BENTHIC DIATOM METABARCODING: DEVELOPING NEW APPROACHES TO RESEARCH AND BIOMONITORING IN AQUATIC ECOSYSTEMS

Javier Pérez Burillo

Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems

Javier Pérez Burillo

Doctoral thesis
2022

IRTA
Institut de Recerca i Tecnologia Agroalimentàries

UNIVERSITAT ROVIRA i VIRGILI

Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems

Doctoral Thesis 2022
Javier Pérez Burillo

IRTA
Institut de Recerca i Tecnologia Agroalimentàries

jperezburillo@gmail.com

UNIVERSITAT ROVIRA i VIRGILI

Doctoral thesis

Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems

Javier Pérez Burillo

Supervised by:

Dr. Francisco Javier Sigró Rodríguez

Dr. Rosa Trobajo Pujadas

Dr. David George Mann

IRTA – Departament de Geografia (URV)

Tarragona 2022

UNIVERSITAT ROVIRA i VIRGILI

FAIG CONSTAR que aquest treball, titulat **Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems**, que presenta Javier Pérez Burillo per a l'obtenció del títol de Doctor, ha estat realitzat sota la meva direcció al Departament de Geografia d'aquesta universitat.

HAGO CONSTAR que el presente trabajo, titulado "**Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems**", que presenta Javier Pérez Burillo para la obtención del título de Doctor, ha sido realizado bajo mi dirección en el Departamento de Geografía de esta universidad.

I STATE that the present study, entitled "**Benthic diatom metabarcoding: developing new approaches to research and biomonitoring in aquatic ecosystems**", presented by Javier Pérez Burillo for the award of the degree of Doctor, has been carried out under my supervision at the Department of Geography of this university.

La Ràpita, 21st of June 2022

El/s director/s de la tesi doctoral
El/los director/es de la tesis doctoral
Doctoral Thesis Supervisor/s

Rosa Trobajo Pujadas - DNI 40311601F (TCAT)
Signat digitalment per Rosa Trobajo Pujadas - DNI 40311601F (TCAT)
Data: 2022.06.21 11:38:59 +02'00'

Dr. Rosa Trobajo Pujadas

David G Mann
Digitally signed by David G_ Mann
DN: CN=David G_ Mann, C=UK, OU=Science Dept, O=Royal Botanic Garden Edinburgh, E=dmann@rbge.org.uk
Reason: I am the author of this document
Location:
Date: 2022-06-21 11:46:32

Dr. David George Mann

SIGRO RODRIGUEZ, FRANCISCO JAVIER (AUTENTICACIÓN)
Firmado digitalmente por SIGRO RODRIGUEZ, FRANCISCO JAVIER (AUTENTICACIÓN)
Fecha: 2022.06.22 11:56:27 +02'00'

Dr. Francisco Javier Sigró Rodríguez

A mi familia y amigos

Agradecimientos

En primer lugar, quiero dar las gracias a la Universitat Rovira i Virgili y al IRTA por ofrecerme la oportunidad de realizar este doctorado mediante el programa de ayudas a la investigación Martí i Franquès. Igualmente, estoy grandemente agradecido a mis supervisores, Xavier Sigró, Rosa Trobajo y David Mann, por haber confiado en mí al darme esta gran oportunidad. En especial, quiero agradecer a Rosa y David por haberme mostrado el apasionante mundo de las diatomeas, por el trabajo y dedicación que han puesto en mi formación, y por todos los conocimientos que me han transmitido.

Agradezco también a Frédéric Rimet, Agnès Bouchez y Louis Jacas por haberme permitido realizar una corta estancia en el centro CARRTEL (INRAE). Junto a ellos, también quiero mencionar a Andrzej Witkowski, François Keck, Greta Valoti, Manoel Leira, Patricia Prado y Valentin Vasselon por su importante contribución en algunos de nuestros artículos.

Por otro lado, quiero agradecer a todo el personal IRTA que me ha ayudado de un modo u otro durante este tiempo, desde la gente de administración y personal de mantenimiento hasta el personal técnico e investigador. Entre ellos, quiero agradecer especialmente a Mari Pau, Ricardo, Vanessa, Edgar, Karl, David Carmona, Ivan, David Vallès, Xus y Yolanda por haber sido siempre tan atentos y generosos. También agradecer a todo el grupo AMIC (Mònica, Jorge, Xavi, Nuria, Margarita, Patricia, Maite y Carles) por haberme enseñado los numerosos aspectos que rodean el mundo científico durante las reuniones de grupo y por los consejos tan importantes que me han dado durante mi doctorado. Igualmente, quiero agradecer a Nil Álvarez, Riki, Maria Belenguer, Joana, Sandra Ramos, Sandra Gimeno, Albertito, Gemma, Lourdes, Mounira, Maria Rey y Greta por haber sido no solo mis compañeros de trabajo, sino también mis amigos durante estos 3 años.

A mis amigos de toda la vida (Guillermo, Castro, Samu, Jorge de la F, Victor, Emilio, Dani Macias, John, Juanfran, Aleandro, Andy, Felipe y Dani Molero) por haber estado siempre ahí a pesar de la distancia. Finalmente, quiero agradecer a mi familia, especialmente a mis padres y hermanos por haberme apoyado en todas las decisiones que he tomado y por haberme inculcado los valores del trabajo, la humildad y el respeto, sin duda, ellos son el artífice de esta tesis.

Table of content

Summary

Diatoms have been one of the most studied groups of protists, partly because they are rich in morphological characters, relative to many flagellate and amoeboid groups, with a siliceous cell wall varying greatly in size, shape and patterning. They are also abundant in aquatic systems and important in biogeochemical cycles. They have applications in biotechnology and stratigraphy and, particularly relevant to this thesis, they are excellent biological indicators. However, estimates suggests that only a small proportion of the total number of extant species have been described so far and many aspects of their ecology remain unknown. One problem in ecological studies and biomonitoring is that they require identification of hundreds of individuals at the species level, which is time-consuming task requiring expert knowledge and considerable microscopical skills. Furthermore there is increasing evidence of cryptic or pseudocryptic species, which differ in few or no discernible morphological characteristics; consequently their geographical distributions and ecological preferences will remain unclear until identification is practical.

DNA metabarcoding (high-throughput sequencing [HTS] of a particular short marker) has recently emerged as an alternative to species identifications based on light microscopic examination (LM). This technology is transforming the way protist diversity can be studied, as thousands of DNA strands can be sequenced in parallel at once, allowing entire communities to be characterised from environmental samples in a relatively simple procedure, to complement the less extensive but richer data provided by microscopy. A key question, however, is the extent to which metabarcoding data faithfully reflect the natural communities from which they are derived. The answer to this question depends on a multitude of factors, including the genetic marker selected, the communities being studied, the molecular processing of the samples and the bioinformatics pipeline used, among others. All of these steps can introduce biases.

The main objective of this thesis was to evaluate the potential and difficulties of using of DNA metabarcoding for the characterisation of some benthic diatom communities in freshwater and coastal environments. There is a special focus on the applicability of the method for Water Framework Directive (WFD) bioassessment of Mediterranean rivers, but we also address ecological and biogeographical questions. The work is organized into 5 chapters, 3 of them represent independent papers that have been published, 1 is under revision and 1 in preparation; these are: 1) Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. 2) Evaluation of two short and similar *rbcL* markers for diatom metabarcoding of

environmental samples: effects on biomonitoring assessment and species resolution. 3) Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters. 4) DNA metabarcoding reveals differences in distribution patterns and ecological preferences among genetic variants within some key freshwater diatom species. 5) Phylogeographical patterns in freshwater diatoms revealed by DNA metabarcoding of a short *rbcL* marker.

Our results for WFD biomonitoring of Mediterranean rivers in Catalonia (NE Spain) builds on previous research, especially in France and the UK, and indicates that *rbcL* metabarcoding of benthic diatoms constitutes an efficient and reliable alternative to LM. One reason for this is that WFD ecological assessments in Catalonia are driven by a relatively small number of common species, for which *rbcL* data are available; i.e. the DNA reference library (essential for converting DNA data into relative abundances of named taxa) is adequate in this region, in contrast to some other regions, e.g. Fennoscandia. However, metabarcoding cannot be considered as an alternative to LM in coastal environments, because of the low coverage of marine benthic diatom species in the reference library. On the other hand, metabarcoding has especial advantages in coastal environments because of its ability to capture information on very delicate or small diatoms and diatoms that exist as endosymbionts in foraminifera and dinoflagellates. In addition, we found that a useful by-product of the *rbcL* metabarcoding protocol is records of other Ochrophyta and Chlorophyta that are co-amplified with diatoms; these can include rarely recorded groups and species.

There are biases in metabarcoding assessments of diatom communities, *relative* to microscopical ones, which need to be taken into account when interpreting results. One bias is caused by interspecific variation in *rbcL* copy number per cell and this seems to have been a factor explaining some discrepancies in the relative abundance of species between LM and DNA metabarcoding in both freshwater and coastal systems. In one case, a major difference between outcomes was the result of failure of the LM approach to faithfully record the abundance of a diatom, *Fistulifera saprophila*, that is destroyed by harsh cleaning methods. We also investigated the relative advantages of two *rbcL* markers that have been proposed (263-bp and 331-bp target regions), finding that the choice has few implications for WFD biomonitoring programmes, but some implications for biodiversity analyses, because of the higher resolution of the 331-bp marker; this allows identification at the species level of certain genetic variants that cannot be separated by the 263-bp marker. In addition, use of the longer marker seems to be more efficient for classifying when using a naïve Bayesian classifier.

Summary

Combining metabarcoding and environmental data for Catalan and French rivers, we created a dataset to investigate whether different Amplicon Sequence Variants (ASVs) within species have the same or different ecological preferences, showing that in *Achnanthidium minutissimum* and *Fistulifera saprophila*, but not in e.g. *Nitzschia inconspicua*, ASVs could be separated into different ecological groupings, some of which deviated from the generally accepted characterizations of the species.

We reanalysed available N American, European and Asian HTS outputs to assemble a large combined dataset of 263-bp Amplicon Sequence Variants (truncating any 331-bp ASVs to 263 bp). We used this to address ecological and biogeographical questions, finding high intraspecific heterogeneity in many cases. Four common phylogeographic patterns were distinguished, which correlate to some extent with biological characteristics: centric diatoms (which are predominantly oogamous and have multiple chloroplast per cell) tend to show lower intraspecific diversity than pennates. 263-bp *rbcL* variants from many species were widely distributed in Europe, North America, the Indian Ocean and Asia, supporting rapid dispersion of diatoms relative to *rbcL* divergence. We concur with other recent assessments of the huge opportunities offered by DNA metabarcoding, even in its current state, which will increase further with technical adjustments and roll-out of long-read technologies.

General Introduction

From barcoding to metabarcoding

The use of short DNA sequences for addressing the taxonomy of microscopic organisms was introduced several decades ago for species difficult to be distinguished on the basis of morphological characteristics (e.g. Arnot et al., 1993; Nanney, 1982; Pace, 1997; Woese & Fox, 1977). Later, Heber et al. (2003) advanced and standardized the use of the mitochondrial gene cytochrome c oxidase I (COI) as a DNA barcode for zoological taxa identification. DNA barcoding thus emerged as a technique capable of providing better taxonomic resolution than that achievable via morphology-based analyses. An ideal DNA barcode is variable enough to enable unambiguous species identification and it must be flanked by regions conservated enough for allowing primers matching. In addition, the DNA barcode should be phylogenetically informative to allow assigning undescribed species or species without a reference barcode to their corresponding taxonomic group (Valentini et al., 2009).

In 2004, the field of molecular ecology experienced a revolution with the arrival of the first high-throughput sequencing (HTS) platform (454 Roche GS FLX System), since HTS technologies, also referred to initially as next-generation sequencing (NGS), enable the sequencing of thousands of DNA strands in parallel. Thus, the large amount of data generated makes it possible to address more and different ecological questions than ever before. The highest transformation has come when coupling HTS with barcoding of a particular marker, which is denominated DNA metabarcoding. This has broken the previous limitation of DNA barcoding based on Sanger sequencing technology which only allowed sequencing a single gene from a single individual in each run. Hence, DNA metabarcoding extends the range of DNA barcoding, from species identification to the characterization of whole communities from environmental samples and importantly, this is done in a relatively simple procedure.

Since 2004, different HTS platforms have been developed and commercialized based on different chemistries, which has led to variations among platforms in read-length, error rate, economical cost, and run time (Morey et al., 2013). As sequencing cost has reduced, the affordability of HTS platforms has enhanced with Illumina's technologies representing the highest reduction in cost and highest gain in throughput sequencing (Reuter et al., 2015). This has transformed the way we study the diversity and ecology of species.

DNA metabarcoding in protists

The protists are a paraphyletic group of eukaryotes that represents most of the microeukaryotic diversity already known. It integrates a vast variety of morphological forms (e.g. amoeboid morphologies, flagellates, ciliates or coccoid forms), and a wide range of nutrition modes (phototrophic, heterotrophic and parasitic) and size ranges (from microorganisms to macroscopic species). In addition, they are distributed in all or most of the earth's environments where they play important roles in biogeochemical cycles (e.g. Caron et al., 2012; Falkowski et al., 1998; Geisen et al., 2018). Despite their enormous ecological relevance, protists have traditionally been understudied compared to eukaryotic macroscopic organisms. In addition, it is particularly difficult to extract reliable data about their diversity and distribution since protists species are prone to be undersampled because of the prevalence of dormant stages, waiting for better conditions to become active, which leads to capturing only a fraction of the total diversity at each sampling time (Foissner et al., 2007). Additionally, the taxonomic classification of protist species is not simple and requires highly skilled personnel.

The emergence of DNA metabarcoding has circumvented many of these problems. Its application in protists has allowed larger sampling efforts at a reduced cost while also achieving high sensitivity and taxonomic resolution (Santoferrara et al., 2020). Moreover, undersampling is less of an issue as both dormant and active stages can be captured. Although some obligate parasites or endosymbionts might still be undersampled because the host organism might correspond to the larger size fraction often removed during sampling (Burki et al., 2021); Alternatively, the host organism might have been lost after an aggressive sample treatment designed for specific groups with resistant cells (e.g. diatoms). On the other hand, and importantly, the bioinformatics classification of sequences derived by metabarcoding is commonly performed by automated approaches. This action itself does not require taxonomic expertise although it is true that the reliability of the classification depends on a reference library curated by taxonomists. All these reasons explain why since the birth of HTS technology, the number of studies applying DNA metabarcoding in protists has importantly increased (Pawlowski et al., 2016) providing new insights into their diversity, biogeographical distribution and functional diversity (e.g. Singer et al., 2021; Malviya et al., 2016).

Nevertheless, some common biological and technical inconveniences can bias the effectiveness of DNA metabarcoding and consequently, the conclusions that are drawn. Particularly, false positives (i.e. the detecting of taxa that are not present in a sample), false negatives (i.e. discarding data that correctly record taxa present in a sample),

artefacts (e.g. non-real sequences mainly derived from sequencing or PCR errors) and defective estimates of relative abundance (mainly related to differences in gene copy number), are errors that can be generated as a consequence of a wide range of factors acting at any of the different steps involved DNA metabarcoding (i.e. from sampling design to bioinformatics analyses) (Santoferrara, 2019). In this regard, the choice of the DNA-barcode constitutes an important factor to be considered. The 18S rDNA gene has been widely used in protist DNA metabarcoding studies because, among others, it is easily amplified and its phylogenetic signal has been proved adequate across many different groups (e.g. de Vargas 1999; Hillis et al., 1991; Stothard et al., 1998) though, other barcodes have been tested and preferred for certain groups of protists and related groups (e.g. Hamsher et al., 2011; Nassonova et al., 2010; Saunders & Kucera, 2010). The effectiveness of a barcode for taxonomic identification, and thereby the dimension achievable by metabarcoding, not only depends on its phylogenetic signal but greatly relies on the degree to which a barcode is covered by a reference library. Among the protists, Baciillariophyta (i.e. diatoms) is probably the group that best represents the existence of both several DNA barcodes with sufficient phylogenetic signal and a curated reference library covering a significant proportion of the total number of described species.

Diatoms: General characteristics and ecological interest

Since C. A. Agardh coined the name “Diatomeae” in 1824 (Mann et al., 2016), diatoms have been one of the most studied organisms among the protists. Diatoms are unicellular diploid cells widely distributed in freshwater and marine systems where they represent a major component of the benthic and planktonic communities. The majority of diatom species are restricted to aquatic habitats but terrestrial specimens are also known. Most of the described diatom species are autotrophs, though some are heterotrophs (Lewin, 1953) and others have been found as endosymbionts in, for example, some foraminifera (e.g. Lee, 2011; Pillet et al., 2011) and dinoflagellate species (e.g. Yamada et al., 2020; You et al., 2015). Diatoms comprise thousands of different species, with estimates of extant species ranging from 100.000 to 200.000 (Mann & Vanormelingen 2013). This immense taxonomic diversity is reflected in a wide variety of morphological forms, size ranges, ecological requirements and reproductive behaviours. Morphologically, diatoms are recognized by examination of their frustule, which constitutes the cell wall. Diatom frustules are characterized by a siliceous nature which makes them strong and resistant structures. According to the symmetry of the frustule and the presence or not of the raphe

system (i.e. the organelle used for locomotion over surfaces), diatoms species have traditionally been divided into two major groups and three different classes: Centric (radial symmetry) and pennate (elongate forms with bilateral symmetry) constitute the two major groups and Coscinodiscophyceae (centric diatoms), Fragillariophyceae (pennate diatoms lacking raphe system or araphid) and Bacillariophyceae (pennate diatoms with raphe system or raphid) are the three differentiated classes (Round et al., 1990). However, it is now evident that the Coscinodiscophyceae and Fragillariophyceae are paraphyletic groups (Theriot et al., 2010).

One of the reasons that explain the great interest in diatoms nowadays is the fact that they have an enormous ecological relevance due to their important role in biogeochemical cycles and carbon fixation. Thus, they are responsible for 20 – 25% of the global carbon dioxide fixation (Mann 1999; Smetacek 1999). More particularly, in some coastal ecosystems, the benthic diatom community can contribute up to 80% of the total benthic production (Cox et al., 2020). In addition to their importance as primary producers, diatoms are a key component of trophic webs because they represent an important part of the diet of a wide range of grazers (e.g. from small protists to molluscs and annelids, Hamels et al., 2004; Lebreton et al., 2011). Nevertheless, despite their great ecological importance, many diatom species remain undescribed and this is especially evident in some ecologically relevant environments such as shallow coastal systems (Mann et al., 2016; Trobajo et al., 2004).

In addition, benthic diatoms have traditionally been used as biological indicators in river biomonitoring programmes around the world because of their rapid and specific response to environmental changes and nutrient conditions, their great diversity and ubiquitous distribution, and the well known ecological preferences of many species (e.g. Bere & Tundisi, 2010; Dalu & Froneman, 2016; Kelly et al., 2008; Rimet, 2012). Another important factor that makes benthic diatoms excellent indicators in aquatic systems is the fact that diatom communities can integrate temporal variability in nutrient conditions over time, reflecting more accurately the health of systems than occasional nutrient measurements (e.g. Lavoie et al., 2008; Smucker & Vis, 2011; Snell et al., 2014).

Diatom barcoding as the basis of current DNA metabarcoding

Identification of diatoms at the genus or species level is a time-consuming task that requires expert knowledge because of the wide range of morphologies existing in the group and the few, or even absent, discernible features existing between some species (e.g. *Nitzschia inconspicua* and *N. soratensis*; Trobajo et al., 2013). In addition, taxonomic boundaries are still not well defined for a large number of species and species complexes, hampering their identification by conventional methods (Mann et al., 2016). Moreover, taxonomic identification for some life stages, such as resting spores, sometimes is not possible (e.g. Kuwata & Takahashi, 1999). All of these explain why, even among trained personnel, disparities in taxonomic classification are common (Kahlert et al., 2009). These difficulties associated with morphological-based identifications represent a limitation for any study or work where diatom species-level identification is required and this is an issue considering the well-known key role of diatoms in aquatic ecosystems and their importance as bioindicators. Consequently, during the past decades, there has been a growing interest in the search for genetic-based methods that could facilitate diatom taxonomic identification.

Thus, efforts have been made to find DNA barcodes with sufficient discriminatory power to, on the one hand, distinguish among species more easily and, on the other hand, to define species boundaries that help in species limitation. In this context, several barcodes have been tested in diatoms species and, among these, the fast-evolving markers ITS-1, ITS-2 (subunits 1 and 2 from the internal transcribed spacer) and COI (cytochrome oxidase subunit 1) have been investigated, each one with particular benefits and inconveniences. More specifically, the higher nucleotide divergence of the ITS region has proved that both ITS-1 and ITS-2 markers are suitable for discrimination among closely related species and some semi-cryptic species (e.g. Amato et al., 2007; Behnke et al., 2004; Vanormelingen et al., 2008). However, intragenomic variation in the ITS region has been observed for some diatom species (Behnke et al., 2004) causing low-quality or unreadable sequences during direct sequencing when different copies are well-represented (e.g. Trobajo et al., 2009). In addition, the alignment of ITS sequences is not very simple and this is important for practical reasons if the objective is species identification or species discovery (Mann et al., 2010). All of these have undermined the importance of ITS for diatom DNA barcoding. COI has also shown sufficient genetic variability for successfully differentiating among species from some groups (Evans et al., 2007) and, it shows two main advantages over ITS: 1) No intraindividual variation has been reported in this marker and 2) its alignment is straightforward since it is a protein-encoding gene. However, the amplification success of COI is reduced for some species

leading to missing data for some lineages that are easily identified by other markers (Trobajo et al., 2010). Thus, COI can be appropriate for certain diatom groups but inefficient for others.

The hypervariable V4 region of the 18S rRNA and the *rbcL* gene (coding for the large subunit of the ribulose-1,5-bisphosphate carboxylase oxygenase) are two markers that show a lower nucleotide divergence than COI and ITS (Evans et al., 2007; Guo et al., 2015). Nevertheless, they are the most widely used markers to date because they contain sufficient variability for discriminating among species, though with exceptions, and both are relatively easily amplified (e.g. Zimmerman et al., 2010). One of the drawbacks of *rbcL* is that it is not valid for diatom species that lack functional plastids, but there are very few of these cases (e.g. *Nitzshia alba*). By contrast, the main advantage of the *rbcL* over the 18S V4 is the fact that *rbcL* sequences are more easily to be aligned since it is a protein-encoding gene which, a part of practical reasons, facilitates the detection of sequencing artefacts by amino acids examination (Mann et al., 2010).

The efforts put over the past years into studying, through DNA data, the ecology, taxonomy and phylogenetic relationships of diatoms have not only increased the knowledge of these aspects but also have made it possible to generate a curated reference library of barcodes (Diat.barcode, formerly called R-Syst::diatom; Rimet et al., 2016, 2019). In addition, the Thonon Culture Collection (TCC) and the UK barcoding project (funded by the UK Environment Agency) have provided another important source of data for filling the reference library (Rimet et al., 2016, 2019). This has promoted the use of DNA metabarcoding of environmental samples for different purposes; from biodiversity and biomonitoring assessments to the study of species biogeography and ecological preferences

Diatom DNA metabarcoding for WFD biomonitoring

Diatom indices used in routinely biomonitoring programmes require species-level identifications and due to the difficulties associated with classical light-microscope examinations, DNA metabarcoding has been considered an alternative method for use in biomonitoring programmes. Prior studies using 454 sequencing platforms already demonstrated the high potential of diatom metabarcoding data for biomonitoring purposes (Kermarrec et al., 2013, 2014). At this stage, the method was mainly limited by the reference library. In this context, very important was, as mentioned in the previous section, the UK diatom metabarcoding project (Kelly et al., 2018, 2020) and the Thonon

Culture Collection (TCC). Other initiatives such as the EU COST Action DNAqua-Net (CA15219) have also contributed to making progress toward the application of genetic tools in biodiversity assessment and biomonitoring programmes of European aquatic systems (Leese et al., 2016). As a result of all these initiatives, Diat.barcode currently covers most of the frequently monitored benthic diatom species in European rivers (Weigand et al., 2019 – note that the coverage of the current version of Diat.barcode is higher as the cited work was based on a previous version, except for those in the far north).

DNA metabarcoding of short markers can already be considered as a realistic alternative for WFD biomonitoring in some countries, as has recently been demonstrated in a number of studies (e.g. Kelly et al., 2020; Mortágua et al., 2019; Rivera et al., 2020). However, before our work (Pérez-Burillo et al., 2020) no exercise had yet been done to test the applicability of DNA metabarcoding for a large set of Mediterranean rivers. Mediterranean rivers are characterized by a highly variable flow regime with heavy rainfall and flooding in winter and drying periods in summer (Pardo & Alvarez, 2006). This natural variability partly explains that Mediterranean ecosystems are one of the most important hot spots of biodiversity in the world (Blondel et al., 2010; Tierno de Figueroa et al., 2012). Nevertheless, such variability is being enhanced further by human activities leading to higher flow intermittency and more frequent periods of drought. This alters the structure and functioning of Mediterranean rivers (loss of tridimensional connectivity and increase of lentic habitats) which ultimately impact the biota occurring in these environments (Bonada et al., 2006; Falasco et al., 2016; Sabater, 2008). Therefore, it is important to have reliable and cost-efficient tools, such as DNA metabarcoding, for monitoring diatom communities that can inform about the different stressors endangering freshwater ecosystems in Mediterranean areas.

There is an interest in assessing the applicability of DNA metabarcoding covering an area that harbours a variety of indicator species since it is crucial to understand to which extent the different biases affecting the method might be leading to an unreal representation of species meaningful for the WFD. The biases affecting the applicability of DNA metabarcoding for biomonitoring programmes include, among others, the incompleteness of the reference library, the bioinformatic strategy used, the gene copy number per cell, and the DNA barcode selected (e.g. Bailet et al., 2019, 2020; Rivera et al., 2020; Vasselon et al., 2017). Concerning the DNA barcode, two similar short regions of the *rbcL* gene are the most common markers used in diatom metabarcoding. Though both have been used successfully for generating biomonitoring assessments (e.g. Kelly et al., 2018, 2020; Mortágua et al., 2019; Rivera et al., 2020), there is no information

about the implications for biodiversity analysis and WFD ecological status assessments of choosing one or other marker. All these previous factors should be specifically addressed for Mediterranean rivers before the implementation of DNA metabarcoding for routine biomonitoring assessments.

The hydrogeographic area of Catalonia (NE Spain) constitutes a suitable subject for testing the applicability of DNA metabarcoding in Mediterranean rivers. This is because this region covers numerous rivers under a Mediterranean climate regime along with a wide variety of geomorphological and physical characteristics that lead to classifying these rivers into 10 different types (ACA, 2010). Moreover, it has been observed that the marked heterogeneity in the physicochemical conditions and pollution levels of these rivers coincide with different benthic diatom communities and thus, different indicator species (Tornés et al., 2007) for which the effectiveness of DNA metabarcoding can be tested.

Assessment of benthic diatom biodiversity in coastal ecosystems by DNA metabarcoding

Coastal environments are located at the interface between terrestrial and marine areas. They include different habitats (e.g. seagrass beds, sandflat communities, coral and bivalve reefs) that support highly productive biological communities and provide numerous ecosystem services (Cloern et al., 2013; Waltham et al., 2020). Microphybenthos (MPB) is one of the essential components of coastal systems because of their role in biogeochemical cycles and of their contribution to both primary production and delivery of multiple ecosystem services (Hope et al., 2019; MacIntyre et al., 1996; Thornton et al., 2002). MPB communities are distributed throughout the sediment in photic zones and are mainly composed of unicellular eukaryotic algae and cyanobacteria. (MacIntyre et al., 1996). More specifically, benthic diatom communities constitute the dominant group of the coastal MPB in respect of biomass and activity (Cox et al., 2020; Underwood et al., 2022). An important aspect of diatoms is that they produce extracellular polymeric substances (EPS) which, among other things, contribute to the formation of microbial biofilms (thus increasing sediment stability), influence organic carbon flux, protect against desiccation and are an important source of carbon for food webs (Czaczyk & Myszka, 2007; Middelburg et al., 2000; Underwood et al., 2004; Widdows et al. 2000).

Because of their intermediate position between land and sea, coastal environments are especially vulnerable to different pollutants derived from human activities. Excess

nutrient inputs, increases in metal contamination and the presence of persistent organic pollutants (POPs) constitute some of the major threats to coastal ecosystems. Importantly, these stressors affect these systems through triggering shifts in community composition, biodiversity loss and decline of ecosystem services provided (e.g. Carstensen et al., 2011; Di Cesare et al., 2020; Lu et al., 2018; Waycott et al., 2009). To assess the impact of these stressors on MPB, it is necessary to have a broad understanding of the ecology of the MPB species. However, it is particularly difficult to study the species composition of MPB communities, as well as aspects related to the factors governing community assemblages, due to the limitations of conventional morphological methodologies. An additional obstacle has been the limited knowledge of the conditions required for unknown species to grow in laboratory cultures.

The arrival of HTS technologies potentially makes studying these aspects more feasible. However, DNA metabarcoding of environmental samples has been rarely tested in coastal ecosystems to study MPB and the few studies conducted to date indicate that conclusions drawn through metabarcoding data are in agreement with those drawn through morphological approaches (Underwood et al., 2022). In addition, these studies have provided interesting insights into species diversity as well as the factors that shape it (Ardura et al., 2021; Bombin et al., 2021; Jeunen et al., 2018; Plante et al., 2021; Rynearson et al., 2020). Considering the importance of diatoms for coastal MPB, and the advances done in the applicability of metabarcoding for their freshwater counterparts, DNA metabarcoding of coastal benthic diatoms becomes therefore an optimal option to increase the knowledge about microeukaryotic diversity patterns and the functioning of these systems. However, for such a purpose, a thorough examination of this technique must be carried out in these systems in order to know its potentialities and limitations. In addition, the complementarity of metabarcoding with traditional morphological analyses must be addressed since morphological-based analyses are still required for ecological interpretations due to the biases associated with DNA metabarcoding and the fact that most of the current knowledge about coastal benthic diatoms is based on such observations (e.g. Witkowski et al., 2000)

DNA metabarcoding for studying patterns of genetic diversity at inter and intraspecific levels

DNA metabarcoding can be used to study genetic variation within species, if the marker is sufficiently variable, through the analyses of environmental samples. For accurately measuring genetic diversity via metabarcoding PCR and sequencing artefacts must be separated from real sequences. For this aim, there are two different bioinformatics strategies applied in DNA metabarcoding. One consists of clustering together sequencing reads that differ less than a defined similarity threshold. These clusters are referred to as operational taxonomic units (OTUs) and the similarity threshold that has often been applied in diatoms ranges from 60% to approximately 97% (e.g. Kelly et al., 2020; Mortágua et al., 2019; Rivera et al., 2020; Vasselon et al., 2017). Although 97% threshold was initially established for 16S rRNA in bacteria (Stackebrandt & Goebel, 1994), some authors have indicated that this cut-off may also be suitable for short diatom *rbcL* markers (kelly et al., 2020). The second most used strategy is based on sequencing "denoise" methods which try to resolve amplicon sequence variants (ASVs; also referred to ESVs [Exact Sequence Variants], sub-OTUs or zero-OTUs) with single-nucleotide resolution from HTS data. For this, sequencing errors are detected and corrected through denoising algorithms. There have been developed several sequencing denoising algorithms such as DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017) and DnoisE (Antich et al., 2022), which differ in the way sequences are corrected. Overall, it has been attributed some advantages for denoise methods over clustering ones. The most important benefit of sequencing denoising approaches has been the fact that ASVs are comparable across studies because they are biological entities independent from the dataset from which they have been inferred. By contrast, OTUs depend on the particular dataset from which they have been defined, which hinders their traceability and reproducibility across studies (Callahan et al., 2017). Furthermore, some studies have reported a higher sensitivity of ASVs approaches compared to OTUs methods (e.g. Kang et al., 2021; Prodan et al., 2020) and it has also been shown that OTUs approaches often overestimate alpha diversity metrics compared to ASVs methods (e.g. Joos et al., 2020; Nearing et al., 2018).

Therefore, DNA metabarcoding based on sequencing denoise methods offers a great opportunity for studying genetic diversity of species, and how such diversity is distributed and structured through wide geographical areas. The study of phylogeographical patterns for a large number of species via metabarcoding is what has been called recently "metaphylogeography" (Turon et al., 2020). As indicated by the authors that

coined the term, this new discipline would allow, among others, to address questions about biogeography, connectivity and dispersal patterns of species in an effective way. However, numerous factors, such as the impossibility of relating HTS sequencing reads to individuals or the low phylogenetic resolution achievable by short markers, are challenging the applicability of DNA metabarcoding for studying some of these aspects, especially at the population level (Adam et al., 2019; Sigsgaard et al., 2019).

Despite these limitations, metabarcoding can still be considered a complementary tool able to provide valuable insights into the genetic diversity and phylogeography of species as some studies have already shown (De Luca et al., 2021; Elbrecht et al,.2018; Shum & Palumbi, 2020; Ruggiero et al., 2022). More in particular, in freshwater diatoms, the opportunity of exploring patterns of distribution of genetic variants is especially useful for studying species complexes. Thus, many diatom species are complexes of genetic variants (e.g. Evans et al., 2008; Trobajo et al., 2010) that show scarcely discernible or no morphological differences (i.e. cryptic species) and therefore it is difficult or impossible to determine their geographical distributions and ecological preferences using traditional methods based on microscopical identifications meaning that the significance of this intraspecific variation is still not clear. On the one hand, it has been suggested that phylogenetically closely related diatoms species often share a similar environmental sensitivity (Keck et al., 2016, 2018) and therefore, it might be thought that the different lineages forming a complex should be similar in terms of spatial distribution and ecological preferences. However, some other studies have shown that lineages within species complexes can do differ in distribution and preferences (e.g. Pinseel et al., 2017; Poulícková et al., 2008; Rynearson et al., 2006). In the same way and more recently, analysis of metabarcoding data has evidenced that variants within some species complexes differ in their tolerance to agriculture stressors (Tapolczai et al., 2021). The application of DNA metabarcoding in certain key species complexes could facilitate disentangling the meaning of their intraspecific diversity, thus leading to more accurate biomonitoring practices in the future.

References

Agardh, C.A., 1824. Systema Algarum. Lund: Literis Berlingianis.

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

Adams, C., Knapp, M., Gemmell, N.J., Jeunen, G.J., Bunce, M., Lamare, M.D., Taylor, H.R., 2019. Beyond Biodiversity: Can Environmental DNA (eDNA) Cut It as a Population Genetics Tool?. Genes 10, 192. https://doi.org/10.3390/genes10030192

Ardura, A., Rick, J., Martinez, J.L., Zaiko, A., Garcia-Vazquez, E., 2021. Stress resistance for unraveling potential biopollutants. Insights from ballast water community analysis through DNA. Mar. Pollut. Bull. 163, 111935. https://doi.org/10.1016/j.marpolbul.2021.112251.

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

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Xu, Z.Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzales, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems 2, e00191–16. https://doi.org/10.1128/mSystems.00191-16

Antich, A., Palacín, C., Turon, X., Wangensteen, O.S., 2022. DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. PeerJ 10:e12758 https://doi.org/10.7717/peerj.12758

Arnot, D.E., Roper, C., Bayoumi, R.A., 1993. Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. Mol. Biochem. Parasitol. 61, 15-24. https://doi.org/10.1016/0166-6851(93)90154-P.

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

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

Behnke, A., Friedl, T., Chepurnov, V.A., Mann, D.G., 2004. Reproductive compatibility and Rdna sequence analyses in the *Sellaphora Pupula* species complex (Bacillariophyta). J. Phycol. 40, 193-208. https://doi.org/10.1046/j.1529-8817.2004.03037.x

Bere, T., Tundisi, J.G., 2011. The effects of substrate type on diatom-based multivariate water quality assessment in a tropical river (Monjolinho), São Carlos, SP, Brazil. Water Air Soil Pollut. 216, 391-409. https://doi.org/10.1007/s11270-010-0540-8,

Blondel, J., Aronson, J., Bodiou, J.Y., Boeuf, G., 2010. The Mediterranean region: biological diversity in space and time. Oxford University, Oxford.

Bombin, S., Wysor, B., Lopez-Bautista, J.M., 2021. Assessment of littoral algal diversity from the northern Gulf of Mexico using environmental DNA metabarcoding. J. Phycol. 57, 269-278. https://doi.org/10.1111/jpy.13087.

Bonada, N., Rieradevall, M., Prat, N., Resh, V.H., 2006. Benthic macroinvertebrate assemblages and macrohabitat connectivity in Mediterranean-climate streams of northern California. J. North Am. Benthol. Soc. 25, 32-43.

Burki, F., Sandin, M.M., Jamy, M., 2021. Diversity and ecology of protists revealed by metabarcoding. Curr. Biol. 31, R1267-R1280. https://doi.org/10.1016/j.cub.2021.07.066.

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

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11**,** 2639–2643 (2017). https://doi.org/10.1038/ismej.2017.119

Caron, D.A., Countway, P.D., Jones, A.C., Kim, D.Y., Schnetzer, A., 2012. Marine protistan diversity. Annu. Rev. Mar. Sci. 4, 467-493. https://doi.org/10.1146/annurev-marine-120709-142802.

Carstensen, J., Sánchez-Camacho, M., Duarte, C.M., Krause-Jensen, D., Marbà, N., 2011. Connecting the dots: responses of coastal ecosystems to changing nutrient concentrations. Environ. Sci. Technol. 45, 9122-9132. https://doi.org/10.1021/es202351y.

Cloern, J.E., Foster, S.Q., Kleckner, A.E., 2013. Review: phytoplankton primary production in the world's estuarine-coastal ecosystems. Biogeosci. Discuss. 10, 17725–17783. https://doi.org/10.5194/bg-11-2477-2014.

Cox, T.E., Cebrian, J., Tabor, M., West, L., Krause, J.W., 2020. Do diatoms dominate benthic production in shallow systems? A case study from a mixed seagrass bed. Limnol. Oceanogr. 5, 425-434. https://doi.org/10.1002/lol2.10167.

Czaczyk, K., Myszka, K., 2007. Biosynthesis of extracellular polymeric substances (EPS) and its role in microbial biofilm formation. Pol. J. Environ. Stud. 16, 60-637.

Dalu, T., Froneman, P.W., 2016. Diatom-based water quality monitoring in southern Africa: challenges and future prospects. Water S.A. 42, 551-559. https://hdl.handle.net/10520/EJC197080

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

de Vargas, C., Norris, R., Zaninetti, L., Gibb, S.W., Pawlowski, J., 1999. Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. Proc. Natl. Acad. Sci. 96, 2864-2868. https://doi.org/10.1073/pnas.96.6.2864.

Di Cesare, A., Pjevac, P., Eckert, E., Curkov, N., Šparica, M. M., Corno, G., Orlić, S., 2020. The role of metal contamination in shaping microbial communities in heavily polluted marine sediments. Environ. Pollut. 265, 114823. https://doi.org/10.1016/j.envpol.2020.114823

Elbrecht, V., Vamos, E.E., Steinke, D., Leese, F., 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644 https://doi.org/10.7717/peerj.4644

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

Evans, K.M., Wortley, A.H., Simpson, G.E., Chepurnov, V.A., Mann, D.G., 2008. A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). J. Phycol. 44, 215-231. https://doi.org/10.1111/j.1529-8817.2007.00454.x

Falasco, E., Piano, E., Bona, F., 2016. Diatom flora in Mediterranean streams: flow intermittency threatens endangered species. Biodivers. Conserv. 25, 2965-2986. https://doi.org/10.1007/s10531-016-1213-8.

Falkowski, P.G., Barber, R.T., Smetacek, V., 1998. Biogeochemical controls and feedbacks on ocean primary production. Science 281, 200-206. DOI: 10.1126/science.281.5374.200.

Foissner, W., 2007. Protist diversity and distribution: some basic considerations. In: Foissner, W., Hawksworth, D.L. (Eds.), Protist Diversity and Geographical Distribution. Topics in Biodiversity and Conservation, 8. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-2801-3_1.

Geisen, S., Mitchell, E.A.D., Adl, S., Bonkowski, M., Dunthorn, M., Ekelund, F., Fernández, L.D., Jousset, A., Krashevska, V., Singer, D., Spiegel, F.W., Walochnik, J., Lara, E., 2018. Soil protists: a fertile frontier in soil biology research. FEMS Microbiol. Rev. 42, 293–323. https://doi.org/10.1093/femsre/fuy006.

Guo, L., Sui, Z., Zhang, S., Ren, Y., Liu, Y., 2015. Comparison of potential diatom 'barcode'genes (the 18S rRNA gene and ITS, COI, rbcL) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. Int. J. Syst. Evol. Microbiol. 65, 1369-1380. https://doi.org/10.1099/ijs.0.000076.

Hamels, I., Mussche, H., Sabbe, K., Muylaert, K., Vyverman, W., 2004. Evidence for constant and highly specific active food selection by benthic ciliates in mixed diatoms assemblages. Limnol. Oceanogr. 49, 58-68. https://doi.org/10.4319/lo.2004.49.1.0058.

Hamsher, S.E., Evans, K.M., Mann, D.G., Poulíčková, A., Saunders, G.W., 2011. Barcoding diatoms: exploring alternatives to COI-5P. Protist 162, 405-422. https://doi.org/10.1016/j.protis.2010.09.005.

Hebert, P.D., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. Royal Soc. B 270, 313-321. https://doi.org/10.1098/rspb.2002.2218.

Hillis, D.M., Dixon, M.T., 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. 66, 411-453. https://doi.org/10.1086/417338.

Hope, J.A., Paterson, D.M., Thrush, S.F., 2019. The role of microphytobenthos in soft-sediment ecological networks and their contribution to the delivery of multiple ecosystem services. J. Ecol. 108, 815-830. https://doi.org/10.1111/1365-2745.13322.

Jeunen, G.-J., Knapp, M., Spencer, H.G., Lamare, M.D., Taylor, H.R., Stat, M., Bunce, M., Gemmell, N.J., 2018. Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. Mol. Ecol. Resour. 19, 426–438. https://doi.org/10.1111/1755-0998.12982.

Joos, L., Beirinckx, S., Haegeman, A., Debode, J., Vandecasteele, B., Baeyen, S., Goormachtig, S., Clement, L., De Tender, C., 2020. Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. BMC Genomics 21, 733. https://doi.org/10.1186/s12864-020-07126-4.

Kahlert, M., Albert, R.L., Anttila, E.L., Bengtsson, R., Bigler, C., Eskola, T., Galman, V., Gottschalk, S., Herlitz, E., Jarlman, A., Kasperoviciene, J., Kokocinski, M., Luup, H., Miettinen, J., Paunksnyte, I., Piirsoo, K., Quintana, I., Raunio, J., Sandell, B., Simola, H., Sundberg, I., Vilbaste, S., Weckstrom, J., 2009. Harmonization is more important than experience – results of the first Nordic-Baltic diatom intercalibration exercise 2007 (stream monitoring). J. Appl. Phycol. 21, 471–482. https://doi.org/10.1007/s10811-008-9394-5.

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría-Galindo, P., Vences, M., Wang, J., Schwalba, A., 2021. Diatom Metabarcoding and Microscopic Analyses from Sediment Samples at Lake Nam Co, Tibet: The Effect of Sample-Size and Bioinformatics on the Identified Communities. Ecol. Indic. 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070.

Keck, F., Bouchez, A., Franc, A., Rimet, F., 2016. Linking phylogenetic similarity and pollution sensitivity to develop ecological assessment methods: a test with river diatoms. J. Appl. Ecol. 53, 856-864. https://doi.org/10.1111/1365-2664.12624.

Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18, 1299–1309. https://doi.org/10.1111/1755-0998.12919.

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

Kelly, M.G., Boonham, N., Juggins, S., Kille, P., Mann, D.G., Pass, D., Sapp, M., Sato, S., Glover, R., 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. Science Report SC140024/R, Environment Agency, Bristol.

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

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F., Bouchez, A., 2013. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Mol. Ecol. Resour. 13, 607-619. https://doi.org/10.1111/1755-0998.12105.

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

Kuwata, A., Takahashi, M., 1999. Survival and recovery of resting spores and resting cells of the marine planktonic diatom *Chaetoceros pseudocurvisetus* under fluctuating nitrate condition. Mar. Biol. 134, 471–478. https://doi.org/10.1007/s002270050563.

Lavoie, I., Campeau, S., Darchambeau, F., Cabana, G., Dillon, P.J., 2008. Are diatoms good integrators of temporal variability in stream water quality?. Freshw. Biol. 53, 827-841. https://doi.org/10.1111/j.1365-2427.2007.01935.x

Lebreton, B., Richard, P., Galois, R., Radenac, G., Pfléger, C., Guillou, G., Mornet, F., Blanchard, G. F., 2011. Trophic importance of diatoms in an intertidal *Zostera noltii* seagrass bed: Evidence from stable isotope and fatty acid analyses. Estuar. Coast. Shelf Sci. 92, 140-153. https://doi.org/10.1016/j.ecss.2010.12.027.

Lee, J.J., 2011. Diatoms as Endosymbionts. In: Seckbach, J., Kociolek, P. (Eds) The Diatom World. Cellular Origin, Life in Extreme Habitats and Astrobiology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1327-7_20

Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., Mergen, P., Pawlowski, J., Piggott, J., Rimet, F., Steinke, D., Taberlet, P., Weigand, A., Abarenkov, K., Beja, P., Bervoets, L., Björnsdóttir, S., Boets, P., Boggero, A., Bones, A., Borja, Á., Bruce, K., Bursić, V., Carlsson, J., Čiampor, F., Čiamporová-Zatovičová, Z., Coissac, E., Costa, F., Costache, M., Creer, S., Csabai, Z., Deiner, K., DelValls, Á., Drakare, S., Duarte, S., Eleršek, T., Fazi, S., Fišer, C., Flot, J., Fonseca, V., Fontaneto, D., Grabowski, M., Graf, W., Guðbrandsson, J., Hellström, M., Hershkovitz, Y., Hollingsworth, P., Japoshvili, B., Jones, J., Kahlert, M., Kalamujic, Stroil, B., Kasapidis, P., Kelly, M., Kelly-Quinn, M., Keskin, E., Kõljalg, U., Ljubešić, Z., Maček, I., Mächler, E., Mahon, A., Marečková, M., Mejdandzic, M., Mircheva, G., Montagna, M., Moritz, C., Mulk, V., Naumoski, A., Navodaru, I., Padisák, J., Pálsson, S., Panksep, K., Penev, L., Petrusek, A., Pfannkuchen, M., Primmer, C., Rinkevich, B., Rotter, A., Schmidt-Kloiber, A., Segurado, P., Speksnijder, A., Stoev, P., Strand, M., Šulčius, S., Sundberg, P., Traugott, M., Tsigenopoulos, C., Turon, X., Valentini, A., van der Hoorn, B., Várbíró, G., Vasquez Hadjilyra, M., Viguri, J., Vitonytė, I., Vogler, A., Vrålstad, T., Wägele, W., Wenne, R., Winding, A., Woodward, G., Zegura, B., Zimmermann, J., 2016 DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Research Ideas and Outcomes 2: e11321. https://doi.org/10.3897/rio.2.e11321

Lewin, J.C., 1953. Heterotrophy in diatoms. Microbiology 9, 305-313. https://doi.org/10.1099/00221287-9-2-305

Lu, Y., Yuan, J., Lu, X., Su, C., Zhang, Y., Wang, C., Cao, X., Li, Q., Su, J., Ittekkot, V., Angus Garbutt, R., Bush, S., Fletcher, S., Wagey, T., Kachur, A., Sweijd, N., 2018. Major threats of pollution and climate change to global coastal ecosystem and enhanced management for sustainability. Environ. Pollut. 239, 670–680. https://doi.org/10.1016/j.envpol.2018.04.016.

MacIntyre, H.L., Geider, R.J., Miller, D.C., 1996. Microphytobenthos: the ecological role of the "secret garden" of unvegetated, shallow-water marine habitats. I. Distribution, abundance and primary production. Estuaries, 19, 186-201. https://doi.org/10.2307/1352224.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A., Bowler, C., 2016. Insights into global diatom distribution and diversity in the world's ocean. Proc. Natl. Acad. Sci. U. S. A. 113, E1516-E1525. https://doi.org/10.1073/pnas.1509523113.

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

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

Mann, D.G., Vanormelingen, P., 2013. An inordinate fondness? The number, distributions, and origins of diatom species. J. Eukaryot. Microbiol. 60, 414–420. https://doi.org/10.1111/jeu.12047.

Mann, D.G., Crawford, R.M., Round, F.E., 2016. Bacillariophyta. In: Archibald, J.M., Simpson, A.G.B., Slamovits, C.H., Margulis, L., Melkonian, M., Chapman, D.J., Corliss, J.O. (Eds.), Handbook of the Protists. Springer, Cham, New York, pp. 1–62. https://doi.org/10.1007/978-3-319-326696_29-1.

Middelburg, J.J., Barranguet, C., Boschker, H.T., Herman, P.M., Moens, T., Heip, C.H., 2000. The fate of intertidal microphytobenthos carbon: An in situ 13C-labeling study. Limnol. Oceanogr. 45, 1224-1234. https://doi.org/10.4319/lo.2000.45.6.1224.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J.M., Couce, M.L., Cocho, J.A., 2013. A glimpse into past, present, and future DNA sequencing. Mol. Genet. Metab. 110, 3-24. https://doi.org/10.1016/j.ymgme.2013.04.024.

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

Nanney, D.L., 1982. Genes and phenes in Tetrahymena. Bioscience, 32, 783-788. https://doi.org/10.2307/1308971

Nassonova, E., Smirnov, A., Fahrni, J., & Pawlowski, J., 2010. Barcoding amoebae: comparison of SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae. Protist 161, 102-115. https://doi.org/10.1016/j.protis.2009.07.003.

Nearing, J.T., Douglas, G.M., Comeau, A.M., Langille, M.G., 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ 6:e5364 https://doi.org/10.7717/peerj.5364.

Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734-740. DOI: 10.1126/science.276.5313.734

Pardo, I., Álvarez, M., 2006. Comparison of resource and consumer dynamics in Atlantic and Mediterranean streams. Limnetica, 25, 271-286. https://doi.org/10.23818/limn.25.19

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., Esling, P., 2016. Protist metabarcoding and environmental biomonitoring: time for change. Eur. J. Protistol. 55, 12-25. https://doi.org/10.1016/j.ejop.2016.02.003.

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

Pillet, L., de Vargas, C., Pawlowski, J., 2011. Molecular identification of sequestered diatom chloroplasts and kleptoplastidy in foraminifera. Protist 162, 394-404. https://doi.org/10.1016/j.protis.2010.10.001

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. Eur. J. Phycol. 52, 264–280. https://doi.org/10.1080/09670262.2017.1283540.

Plante, C.J., Hill-Spanik, K., Cook, M., Graham, C., 2021. Environmental and spatial influences on biogeography and community structure of saltmarsh benthic diatoms. Estuar. Coasts. 44, 147–161. https://doi.org/10.1007/s12237-020-00779-0.

Poulícková, A., Špacková, 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/s10750008-9521-y.

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A.H., Nieuwdorp, M., Levin, E., 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*, e0227434. https://doi.org/10.1371/journal.pone.0227434.

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-throughput sequencing technologies. Mol. Cell. 58, 586-597. https://doi.org/10.1016/j.molcel.2015.05.004.

Rimet, F., 2012. Recent views on river pollution and diatoms. Hydrobiologia 683, 1-24. https://doi.org/10.1007/s10750-011-0949-0

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

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

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

Round, F.E., Crawford, R.M., Mann, D.G., 1990. The diatoms. Biology and morphology of the genera. Cambridge University Press, Cambridge.

Ruggiero, M.V., Kooistra, W.H., Piredda, R., Sarno, D., Zampicinini, G., Zingone, A., Montresor, M., 2022. Temporal changes of genetic structure and diversity in a marine diatom genus discovered via metabarcoding. Environ. DNA. https://doi.org/10.1002/edn3.288

Rynearson, T.A., Newton, J.A., Armbrust, E.V., 2006. Spring bloom development, genetic variation, and population succession in the planktonic diatom *Ditylum brightwellii*. Limnol. Oceanogr. 51, 1249-1261. https://doi.org/10.4319/lo.2006.51.3.1249.

Rynearson, T.A., Flickinger, S.A., Fontaine, D.N., 2020. Metabarcoding reveals temporal patterns of community composition and realized thermal niches of Thalassiosira spp. (Bacillariophyceae)

from the Narragansett Bay long-term plankton time series. Biology 9, 19. https://doi.org/10.3390/biology9010019.

Sabater, S., 2008. Alterations of the global water cycle and their effects on river structure, function and services. Freshw. Rev. 1, 75-88. https://doi.org/10.1608/FRJ-1.1.5

Santoferrara, L.F., 2019. Current practice in plankton metabarcoding: optimization and error management. J. Plankton Res. 41, 571-582. https://doi.org/10.1093/plankt/fbz041.

Santoferrara, L., Burki, F., Filker, S., Logares, R., Dunthorn, M., McManus, G.B., 2020. Perspectives from ten years of protist studies by high-throughput metabarcoding. J. Eukaryot. Microbiol. 67, 612-622. https://doi.org/10.1111/jeu.12813.

Saunders, G.W., Kucera, H., 2010. An evaluation of *rbcL*, *tufA*, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. Cryptogam. Algol. 31, 487.

Shum, P., Palumbi, S.R., 2021. Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. Mol. Ecol. 30, 3355-3373. https://doi.org/10.1111/mec.15851

Singer, D., Seppey, C.V.W., Lentendu, G., Dunthorn, M., Bass, D., Belbahri, L., Blandenier, Q., Debroas, D., de Groot, G.A., de Vargas, C., Domaizon, I., Duckert, C., Izaguirre, I., Koenig, I., Mataloni, G., Schiaffino, M.R., Mitchell, E.A.D., Geisen, S., Lara, E., 2021. Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. Environ. Int. 146. https://doi.org/10.1016/j.envint.2020.106262.

Sigsgaard, E.E., Jensen, M.R., Winkelmann, I.E., Møller, P.R., Hansen, M.M., Thomsen, P.F., 2019. Population-level inferences from environmental DNA—Current status and future perspectives. Evol. Appl. 13, 245-262. https://doi.org/10.1111/eva.12882

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

Smucker, N.J., & Vis, M.L., 2011. Diatom biomonitoring of streams: Reliability of reference sites and the response of metrics to environmental variations across temporal scales. Ecol. Indic. 11, 1647-1657. https://doi.org/10.1016/j.ecolind.2011.04.011.

Snell, M.A., Barker, P.A., Surridge, B.W.J., Large, A.R.G., Jonczyk, J., Benskin, C.M.H., Reaney, S., Perks, M.T., Owen, G.J., Cleasby, W., Deasy, C., Burke, S., Haygarth, P.M., 2014. High frequency variability of environmental drivers determining benthic community dynamics in headwater streams. Environ. Sci.: Process. Impacts 16, 1629-1636.

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 44, 846-849. https://doi.org/10.1099/00207713-44-4-846.

Stothard, D.R., Schroeder-Diedrich, J.M., Awwad, M.H., Gast, R.J., Ledee, D.R., Rodriguez-Zaragoza, S., Dean, C.L., Fuerst, A.P., Byers, T.J., 1998. The evolutionary history of the genus Acanthamoeba and the identification of eight new 18S rRNA gene sequence types. J. Eukaryot. Microbiol. 45, 45-54. https://doi.org/10.1111/j.1550-7408.1998.tb05068.x.

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

Tierno de Figueroa, J.M., López-Rodríguez, M.J., Fenoglio, S., Sánchez-Castillo, P., Fochetti, R., 2013. Freshwater biodiversity in the rivers of the Mediterranean Basin. Hydrobiologia, 719, 137-186. https://doi.org/10.1007/s10750-012-1281-z.

Theriot, E.C., Ashworth, M., Ruck, E., Nakov, T., Jansen, R.K., 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecol. Evol. 143, 278-296. https://doi.org/10.5091/plecevo.2010.418.

Thornton, D.C., Dong, L.F., Underwood, G.J., Nedwell, D.B., 2002. Factors affecting microphytobenthic biomass, species composition and production in the Colne Estuary (UK). Aquat. Microb. Ecol. 27, 285-300. https://doi.org/10.3354/ame027285.

Tornés, E., Cambra, J., Gomà, J., Leira, M., Ortiz, R., Sabater, S., 2007. Indicator taxa of benthic diatom communities: a case study in Mediterranean streams. Ann. Limnol. - Int. J. Lim. 43, 1–11. https://doi.org/10.1051/limn/2007023

Trobajo, R., 2004. Factors affecting the periphytic diatom community in Mediterranean coastal wetlands (Empordà wetlands, NE Spain). Archiv für Hydrobiologie 375-399. 10.1127/0003-9136/2004/0160-0375.

Trobajo, R., Clavero, E., Chepurnov, V.A., Sabbe, K., Mann, D.G., Ishihara, S., Cox, E.J., 2009. Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). Phycologia *48*, 443-459. https://doi.org/10.2216/08-69.1.

Trobajo, R., Mann, D.G., Clavero, E., Evans, K.M., Vanormelingen, P., McGregor, R.C., 2010. The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). Eur. J. Phycol. 45, 413-425. https://doi.org/10.1080/09670262.2010.498586.

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

Turon, X., Antich, A., Palacín, C., Præbel, K., Wangensteen, O.S., 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol. Appl. 30, e02036. https://doi.org/10.1002/eap.2036.

Underwood, G.J.C., Boulcott, M., Raines, C.A., Waldron, K., 2004. Environmental effects on exopolymer production by marine benthic diatoms: dynamics, changes in composition, and pathways of production. J. Phycol. 40, 293-304. https://doi.org/10.1111/j.1529-8817.2004.03076.x

Underwood, G.J.C., Dumbrell, A.J., McGenity, T.J., McKew, B.A., Whitby, C., 2022. The Microbiome of Coastal Sediments. In: Stal, L.J., Cretoiu, M.S. (Eds.) The Marine Microbiome. The Microbiomes of Humans, Animals, Plants, and the Environment. Springer, Cham. https://doi.org/10.1007/978-3-030-90383-1_12

Valentini, A., Pompanon, F., Taberlet, P., 2009. DNA barcoding for ecologists. Trends Ecol. Evol. 24, 110-117. https://doi.org/10.1016/j.tree.2008.09.011.

Vanormelingen, P., Chepurnov, V.A., Mann, D.G., Sabbe, K., Vyverman, W., 2008. Genetic divergence and reproductive barriers among morphologically heterogeneous sympatric clones of *Eunotia bilunaris* sensu lato (Bacillariophyta). Protist 159, 73-90. https://doi.org/10.1016/j.protis.2007.08.004

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

Waltham, N.J., Elliott, M., Lee, S.Y., Lovelock, C., Duarte, C.M., Buelow, C., Simenstad, C., Nagelkerken, I., Claassens, L., Wen, C.K., 2020. UN Decade on Ecosystem Restoration 2021–2030—what chance for success in restoring coastal ecosystems? Front. Mar. Sci. 7, 71. https://doi.org/10.3389/fmars.2020.00071.

Waycott, M., Duarte, C.M., Carruthers, T.J.B., Orth, R.J., Dennison, W.C., Olyarnik, S., Calladine, A., Fourqurean, J.W., Heck Jr., K.L., Hughes, A.R., Kendrick, G.A., Kenworthy, W.J., Short, F.T.,

Williams, S.L., 2009. Accelerating loss of seagrasses across the globe threatens coastal ecosystems. Proc. Natl. Acad. Sci. U S A 106, 12377-12381. https://doi.org/10.1073/pnas.0905620106.

Weigand, H., Beermann, A.J., Čiampor, F., Costa, F.O., Csabai, Z., Duarte, S., Geiger, M.F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A.M., Willassen, E., Wyler, S.A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., Dijkstra, K.-D.B., Eisendle, U., Freyhof, J., Gadawski, P., Graf, W., Haegerbaeumer, A., van der Hoorn, B.B., Japoshvili, B., Keresztes, L., Keskin, E., Leese, F., Macher, J.N., Mamos, T., Paz, G., Pešić, V., Pfannkuchen, D.M., Pfannkuchen, M.A., Price, B.W., Rinkevich, B., Teixeira, M.A.L., Várbíró, G., Ekrem, T., 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Sci. Total Environ. 678, 499-524. https://doi.org/10.1016/j.scitotenv.2019.04.247.

Widdows, J., Brinsley, M.D., Salkeld, P.N., Lucas, C.H., 2000. Influence of biota on spatial and temporal variation in sediment erodability and material flux on a tidal flat (Westerschelde, The Netherlands). Mar. Ecol. Prog. Ser. 194, 23-37. https://doi.org/10.3354/meps194023.

Witkowski, A., Lange-Bertalot, H., Metzeltin, D., 2000. Diatom flora of marine coasts. Iconographia diatomologica, Vol. 7. Koeltz Scientific Books, Königstein, Germany.

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. 74, 5088-5090. https://doi.org/10.1073/pnas.74.11.5088

Yamada, N., Sakai, H., Onuma, R., Kroth, P.G., Horiguchi, T., 2020. Five non-motile dinotom dinoflagellates of the genus *Dinothrix*. Front. Plant Sci. 1764. https://doi.org/10.3389/fpls.2020.591050.

You, X., Luo, Z., Su, Y., Gu, L., Gu, H., 2015. *Peridiniopsis jiulongensis*, a new freshwater dinoflagellate with a diatom endosymbiont from China. Nova Hedwigia 313-326. 10.1127/nova_hedwigia/2015/0272.

Zimmermann, J., Jahn, R., Gemeinholzer, B., 2011. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org. Divers. Evol. 11, 173. https://doi.org/10.1007/s13127-011-0050-6

Objectives

The main objective of this thesis is to evaluate the use of DNA metabarcoding for the characterisation of benthic diatom communities in freshwater and coastal environments. It focuses particularly on the applicability of benthic diatom DNA metabarcoding for rivers WFD bioassessment but also explores the possibilities for addressing other ecological and biogeographical questions, with the following specific objectives:

1. To determine whether DNA metabarcoding constitutes a reliable tool for WFD biomonitoring of rivers under a Mediterranean climate regime (in NE Spain).

2. To identify the species responsible for causing the largest discrepancies between LM and DNA metabarcoding in WFD biomonitoring of freshwater systems, and to determine the reasons behind such discrepancies.

3. To compare the phylogenetic resolution of two similar and short diatom *rbcL* markers at or below the species level and to assess the effect of choosing between these two markers for WFD biomonitoring programmes and biodiversity-related studies. Furthermore, the potential of the region shared (i.e. 263-bp) by both markers to identify non-diatom benthic microeukaryotes is assessed.

4. To use a global freshwater diatom metabarcoding database (263-bp *rbcL*) to identify common phylogeographic patterns and to evaluate the different causes that could explain the differences between species in their intraspecific diversity for this marker.

5. To study the distribution and ecological preferences of different genetic variants within some particular species complexes of diatoms with ecological relevance and importance for WFD biomonitoring programmes.

6. To examine the current status of DNA metabarcoding for assessing benthic diatom communities in shallow coastal environments, and to identify the advantages and disadvantages of this method compared to morphological analyses.

Methodology

The following section presents a very brief overview of the methodology used in this thesis. Note that the procedures and methods used in sampling and morphological analysis were carried out by collaborators. Thus, this thesis focuses on the analysis of diatom metabarcoding data. For more details, see the Material and Methods section in each respective chapter (1, 2, 3, 4 and 5).

1. Study areas and datasets used

The datasets used in chapters 1, 3 and 4 correspond to benthic samples collected in rivers in Catalonia (chapters 1 and 4; 164 samples), France (Chapter 4; 610 samples) and the UK (chapter 3; 1703 samples) as part of the WFD biomonitoring networks held during 2016 and 2017 (in the case of the French and Catalan networks) and during 2014, 2016 and 2017 (in the case of the UK networks). Chapter 5 is based on a much larger benthic diatom metabarcoding database, created from a combination of the above databases, plus 9 additional diatom metabarcoding databases that were collected from the public online repositories Sequence Read Archive (SRA) and Zenodo. These extra databases cover regions in North America (California, Ohio and Ontario), Europe (Fennoscandia, France and Spain), Asia (Tibet) and the Indian Ocean (Mayotte). For chapter 2, the data used corresponded to 9 biofilms samples taken from Alfacs and Fangar bays from the Ebro Delta (NE Spain)

2. Diatom morphological data

In chapters 1 and 2, samples were prepared for morphological analyses using light microscopy (LM). In both cases, the treatment used was based on chemical oxidation (using H_2O_2, HNO_3 orH_2SO_4) and the resulting cleaned diatom valves were mounted with Naphrax resin (Brunel microscopes, Chippenham, UK). At least 400 valves were counted per sample in chapter 1 and between 300 to 400 in chapter 2 using in both cases a 100x objective. In chapter 1, freshwater diatom identifications were performed by several consultancies and mainly followed the taxonomic guides of Krammer and Lange-Bertalot (1986a, b, 1991a, b) and Lange-Bertalot et al. (2017). In the case of marine samples of chapter 2, diatoms identification was carried out following the taxonomic guide of Witkowski et al. (2000).

3. Diatom metabarcoding data

3.1. DNA extraction, PCR amplification and high-throughput sequencing (HTS) library preparation.

In chapters 1, 2 and 4, DNA was extracted using the commercial kits GenElute TM-LPA and NucleoSpin Soil kit (MN-Soil). In chapter 3, DNA from benthic samples was extracted using DNeasy Blood and Tissue kit. In chapter 5, the 9 databases collected from the online repositories were based on a variety of different kits (see Table 1 in chapter 5).

A short *rbcL* region of 312-bp (263-bp without primers) was the marker used in Chapters 1, 2, 4, 5 and part of the data in Chapter 3. This maker was amplified using the primers Diat_rbcL_708F (forward) and R3 (reverse) (Vasselon et al., 2017). The *rbcL* marker used in chapter 3 was 379-bp long (331-bp without primers) and was amplified by the rbcL-646F and rbcL-998R primers (Kelly et al., 2018, 2020).

In all cases, the extracted DNA was sequenced by the illumina Miseq sequencing platform. The exceptions were 2 databases used in chapter 5 (i.e. databases from Mayotte island and French lakes) where the source DNA was sequenced by the PGM Ion Torrent platform due to the time of the work performed.

3.2 Bioinformatic analyses

For all the chapters except 1 (i.e. 2, 3, 4 and 5), DADA2 (Callahan et al., 2016) pipeline was the main bioinformatics pipeline used for the processing of fastq files. Overall, primers were first removed from the R1 and R2 reads using cutadapt (Martin, 2011). The resulting R1 and R2 reads were truncated to 200-240 and 160-200 nucleotides respectively, based on their quality profile (i.e. discarding sequences with median quality score < 30). Reads with ambiguities or an expected error (maxEE) higher than 2 were discarded. Then, the DADA2 denoising algorithm was applied to determine an error rates model in order to infer amplicon sequence variants (ASVs) and ASVs detected as chimeras were discarded using the function “removeBimeraDenovo”.

Fastq files in chapter 1 were processed using Mothur software (Schloss et al., 2009). Overall, reads that showed some of the following properties were removed: lengths < 250 bp, Phred quality score < 23, > 1 mismatch in the primer sequence and homopolymer > 8 bp. The Uchime algorithm (Edgar et al., 2011) was used for discarding chimeras. Finally, reads were clustered into OTUs using 95% as similarity thresholds.

Taxonomic classification of both ASVs and OTUs was performed using the naïve Bayesian classifier method (Wang et al., 2007) and the reference library Diat.barcode

(Rimet et al., 2019). In some cases (chapters 2, 3 and 5) the Basic Local Alignment Search Tool (BLAST) was used against the Nucleotide database of NCBI GenBank or Diat.barcode to check the taxonomy of ASVs. In addition, maximum likelihood trees based on the GRT-Gamma model were performed to evaluate the taxonomy of the ASVs of the species selected in chapters 2, 4 and 5.

3.3 Ecological indexes, statistical analyses and haplotype networks

In chapters 1 and 3, the ecological status of rivers was determined by applying the benthic diatom index IPS (Indice de Polluosensibilité Spécifique; Cemagref 1982). For each site, the IPS was calculated using species inventory data (species composition and relative abundance), obtained from both LM analyses (chapter 1) and DNA metabarcoding (chapters 1 and 3), and IPSS and IPSV values for each species extracted from OMNIDIA v5.5 software (Lecointe et al., 1993). The WFD ecological status class for each sample was assigned by applying the following boundaries (Afnor, 2007): High ($17 \leq IPS \leq 20$), Good ($13 \leq IPS < 17$), Moderate ($9 \leq IPS < 13$), Poor ($5 \leq IPS < 9$), Bad ($1 \leq IPS < 5$). In chapter 1, the contribution of each species to the IPS values was evaluated by a sensitivity analysis and the correlation in IPS between LM and DNA metabarcoding methods was by Pearson's coefficient (also used in chapter 3 for comparing correlation in IPS between markers).

In chapter 2, the following statistical analyses were used for comparing diatom communities between methods and sites: Shannon–Wiener and Sørensen indexes were used to compare diatom diversity between methods (LM vs DNA metabarcoding) and sampling sites. Non-metric multidimensional scaling (NMDS) was used to discriminate patterns in taxon composition among sample sites and, a permutation multivariate analysis of variance (PERMANOVA) was used to evaluate statistically significant differences in the diatom community. Finally, an analysis of similarity percentages (SIMPER) was performed to identify the taxa that accounted for most of the dissimilarities between the LM and DNA metabarcoding inventories.

In chapter 4, redundancy analyses (RDA) models were performed to analyse separately the relationships between the environmental and spatial data and the ASVs. In addition, the ecological preferences of ASVs were evaluated using Threshold Indicator Taxa Analyses (Baker & King, 2010) and Boosted Regression Trees (Elith, 2008).

Haplotype networks were constructed in chapters 3 and 5 to assess how phylogenetic relationships of genetic variants differed as a function of the marker used (chapter 3) and, to explore the phylogeography of species across the regions studied (chapter 5)

Haplotype networks were based on the TCS algorithm (Clement et al. 2002) and visualised using PopART software (Leigh and Bryant, 2015).

References

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

Baker, M.E., King, R.S., 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods Ecol. Evol. 1, 25-37. https://doi.org/10.1111/j.2041-210X.2009.00007.x.

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

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

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

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27, 2194-2200.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Kelly, M.G., Boonham, N., Juggins, S., Kille, P., Mann, D.G., Pass, D., Sapp, M., Sato, S., Glover, R., 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. Science Report SC140024/R, Environment Agency, Bristol.

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

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

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

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

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

Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., 2017. Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment. English

Edition With Updated Taxonomy and Added Species. Koeltz Botanical Books, Schmitten-Oberreifenberg, pp. 1-942.

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

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

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

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

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

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

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. Appl. Environ. Microbiol. 73, 5261-5267. https://doi.org/10.1128/AEM.00062-07.

Witkowski, A., Lange-Bertalot, H., Metzeltin, D., 2000. Diatom flora of marine coasts. Iconographia diatomologica, Vol. 7. Koeltz Scientific Books, Königstein, Germany.

Scientific publications

This doctoral thesis is a compendium of 3 peer-reviewed publications and 2 additional manuscripts, of which one is currently under review and the other is in preparation. These five works represent the core of the thesis. An additional publication is added as an annex to reflect those other articles in which I have been involved during this thesis but with a smaller contribution. In the following paragraphs all publications are listed together with my personal contribution to each of them. In addition, the category, journal impact factor (JIF), and the quartile (Q) of each journal where articles have been published are specified.

Chapter 1: Pérez-Burillo, J., Trobajo, R., Vasselon, V., Rimet, F., Bouchez, A., Mann, D.G., 2020. Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. Sci. Total Environ. 727, 138445. https://doi.org/10.1016/j.scitotenv.2020.138445. (Category: ENVIRONMENTAL SCIENCES; JIF 2020: 7.963; Q1).

Personal contribution: Formal analysis, Investigation, Data curation, Writing (original draft, review & editing) and Visualization.

Chapter 2: Pérez-Burillo, J., Valoti, G., Witkowski, A., Prado, P., Mann, D. G., Trobajo, R., 2022. Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters. Mar. Pollut. Bull. 174, 113183. https://doi.org/10.1016/j.marpolbul.2021.113183. (Categroy: ENVIRONMENTAL SCIENCES; JIF 2020: 5,553; Q1)

Personal contribution: Formal analysis, Investigation, Data curation, Methodology, Writing (original draft, review & editing) and Visualization.

Chapter 3: Pérez-Burillo, J., Trobajo, R., Mann, D. G.. Evaluation of two short and similar *rbcL* markers for diatom metabarcoding of environmental samples: effects on biomonitoring assessment and species resolution. *Chemosphere* (Under Review) (Category: ENVIRONMENTAL SCIENCES; JIF 2020: 7.086; Q1)

Personal contribution: Formal analysis, Investigation, Data curation, Methodology, Writing (original draft, review & editing) and Visualization.

.

Chapter 4: Pérez-Burillo, J., Trobajo, R., Leira, M., Keck, F., Rimet, F., Sigró, J., Mann, D.G., 2021. DNA metabarcoding reveals differences in distribution patterns and ecological preferences among genetic variants within some key freshwater diatom species. Sci. Total Environ. 728, 149029. https://doi.org/10.1016/j.scitotenv.2021.149029. (Category: ENVIRONMENTAL SCIENCES; JIF 2020: 7.963; Q1).

Personal contribution: Formal analysis, Investigation, Data curation, Methodology, Writing (original draft, review & editing) and Visualization.

Chapter 5: Pérez-Burillo, J., Trobajo, R., Mann, D.G.. Phylogeographical patterns in freshwater diatoms revealed by DNA metabarcoding of a short *rbcL* marker. In preparation

Personal contribution: Formal analysis, Investigation, Data curation, Methodology, Writing (original draft, review & editing) and Visualization.

Annex 2: Rambaldo, L., Ávila, H., Casas, M.E., Guivernau, M., Viñas, M., Trobajo, R., Pérez-Burillo, J., Mann, D.G., Fernández, B., Biel, C., Rizzo, L., 2022. Assessment of a novel microalgae-cork based technology for removing antibiotics, pesticides and nitrates from groundwater. Chemosphere, 301, 134777. https://doi.org/10.1016/j.chemosphere.2022.134777. (Category: ENVIRONMENTAL SCIENCES; JIF 2020: 7.086; Q1)

Personal contribution: Writing (reviewing and editing)

Chapter 1

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

Pérez-Burillo, J., Trobajo, R., Vasselon, V., Rimet, F., Bouchez, A., Mann, D.G., 2020.

Sci. Total Environ. 727, 138445.
https://doi.org/10.1016/j.scitotenv.2020.138445.

Science of the Total Environment 727 (2020) 138445

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

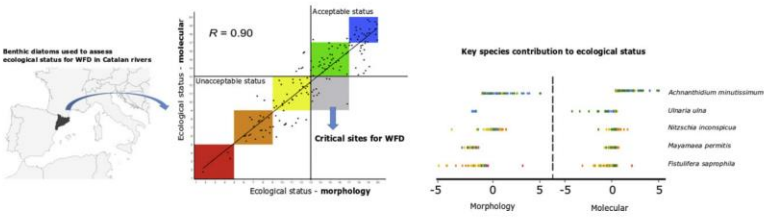
Javier Pérez-Burillo [a,b,*], Rosa Trobajo [a], Valentin Vasselon [c,d], Frédéric Rimet [e,f], Agnès Bouchez [e,f], David G. Mann [a,g]

[a] 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, Catalonia, Spain
[b] Departament de Geografia, Universitat Rovira i Virgili, C/Joanot Martorell 15, E43500 Vila-seca, Catalonia, Spain
[c] Pôle R&D "ECLA", France
[d] AFB, Site INRA UMR CARRTEL, Thonon-les-Bains, France
[e] INRAE, UMR Carrtel, 75 av. de Corzent, FR-74203 Thonon les Bains cedex, France
[f] University Savoie Mont-Blanc, UMR CARRTEL, FR-73370 Le Bourget du Lac, France
[g] Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, Scotland, UK

HIGHLIGHTS

- DNA- and morphology-based diatom assessments of river ecological status are compared.
- Diatom DNA metabarcoding can be a reliable tool for WFD assessment of Catalan rivers.
- Sensitivity analysis shows which species drive ecological status assessments.
- Metabarcoding–morphology ecological status deviations are caused by a few key species.
- Metabarcoding shows some diatoms are seriously underrecorded in light microscopy.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 January 2020
Received in revised form 31 March 2020
Accepted 2 April 2020
Available online 04 April 2020

Editor: Sergi Sabater

Keywords:
Environmental DNA
High-throughput sequencing
*rbc*L
Water Framework Directive
Benthic diatoms
Catalan rivers

ABSTRACT

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

* Corresponding author at: 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, Catalonia, Spain.
E-mail address: Javier.perez@irta.cat (J. Pérez-Burillo).

https://doi.org/10.1016/j.scitotenv.2020.138445

identified by the sensitivity analysis as important for the WFD. Only minor IPS discrepancies were attributed to the incompleteness of the reference library, as most of the abundant and influential species (to the IPS) were well represented there. Finally, we propose that leave-one-out analysis is a good method for identifying priority species for isolation and barcoding.

1. Introduction

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

Several diatom indices have been proposed for ecological status assessment, most of them being derived from the formula of Zelinka and Marvan (1961). One of the most commonly used indices for benthic diatoms is the Indice de Polluosensibilité Spécifique (IPS; Cemagref, 1982) which, like other widely used diatom indices, is calculated on the basis of species' relative frequencies, pollution sensitivity values (IPSS) and pollution tolerance values (IPSV). However, the morphological identifications at species level needed for the calculation of these indices are a time-consuming task and require expert knowledge; furthermore, the taxonomic boundaries are still not well defined in a large number of species and complexes, hampering or even precluding their identification by light microscopy (Mann et al., 2016).

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

For the management of European rivers covered by the WFD, incongruences between methods become especially important when they cause the perceived ecological status of a water body to change class (five classes are recognized: High, Good, Moderate, Poor and Bad). The most important difference occur when morphological analysis (the current methodological standard) assigns "Good" or "High" ecological status to a particular site but the molecular approach assigns instead a "Moderate", "Poor" or "Bad" status. This is because the WFD requires action to be taken to improve those aquatic systems that do not reach at least "Good" ecological status and this often has economic implications. We will therefore focus on these "critical sites" in the current paper (i.e. on those Catalan sites whose status alters from Good/High in LM assessments to Moderate/Poor/Bad with DNA metabarcoding), while accepting that a detailed analysis of movements across other status boundaries may also be of interest and relevance to regulators. In particular, we analyse how different biases may contribute to making the IPS score drop below the critical Good to Moderate threshold. There has previously been some analysis of the extent to which particular diatom species contribute to the final ecological status obtained morphologically (Almeida et al., 2014) and to deviations in IPS values between the molecular and morphological methods (Bailet et al., 2019). In both studies, the analyses were based only on relative abundances of species. However, since the IPS value depends not only on the relative abundances of the species present in a sample, but also on their pollution sensitivity values (IPSS) and tolerance values (IPSV), the contribution of each species to the final IPS score for that sample should take all three parameters into account. This will allow the real impact of each species on the final IPS score to be evaluated and thus identify the main species that lead to IPS discrepancies between methods.

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

2. Material and methods

2.1. Study site

The study area corresponds to the hydrographic area of Catalonia, which is divided into internal and interregional hydrographic basins (Fig. 1). The former comprises a total of eleven main rivers and extends across 16,423 km^2 (52% of the territory of Catalonia). Amongst these eleven, the basins of the rivers Llobregat and Ter are the most extensive and occupy approximately half of the total surface covered by the internal basins. The interregional basins are shared with other Spanish regions and cover the Catalan sections of the rivers Ebro, Garona and Xúquer, with a total extent of 15,567 km^2 (48% of the surface area of Catalonia). For this study, 160 out of the total 164 samples were taken from rivers that belong to the internal basins and the remaining 4 samples were collected from the Lower Ebro river (Fig. 1).

The rivers sampled are influenced predominantly by Mediterranean climatic factors, though some of them are affected by continental or high mountain climates. This climatic diversity, together with the varied geology and the irregular terrain characteristic of Catalonia, has led to Catalan rivers being classified into 10 different types (ACA, 2010). On the other hand, Catalan rivers are affected by various anthropogenic pressures, such as urban and industrial wastewater discharges, urban and industrial land uses, agriculture, and hydromorphological alterations.

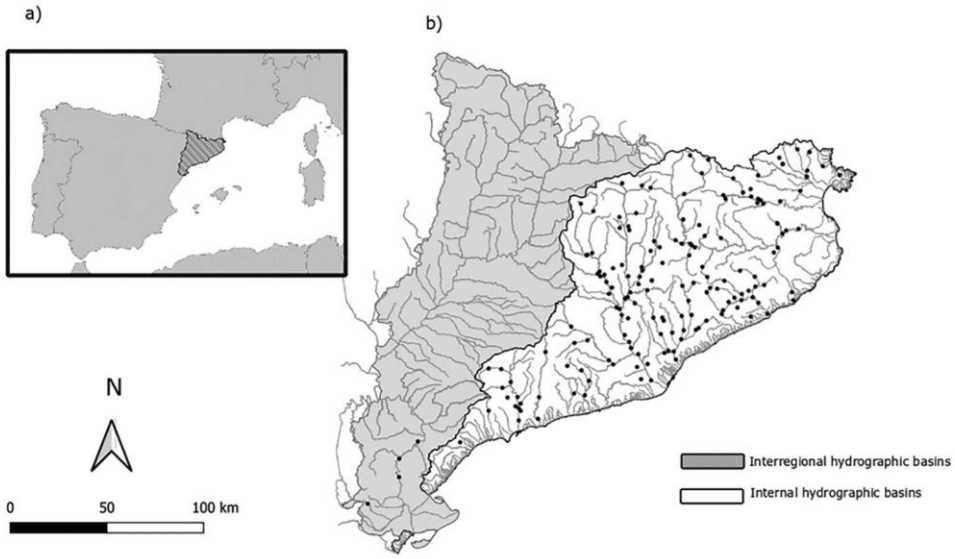


Fig. 1. a) Location of Catalonia (Spain) and b) river sites sampled in the internal (white) and interregional (grey) Catalan hydrographic basins.

2.2. Diatom sampling

All 164 sites were sampled for epilithon between April and July of 2017 following standard procedures (CEN, 2014a). At each site, diatoms were collected from at least 5 stones by brushing their upper surfaces using a toothbrush. The resulting samples were divided into two aliquots, one of which was preserved with formalin or ethanol and used for morphological analyses as part of the statutory monitoring and control program of the Catalan Water Agency (ACA). The second aliquot was preserved by adding >95% 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).

2.3. Morphological analyses

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

2.4. DNA extraction and PCR amplification

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

2.5. High-throughput sequencing

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

2.6. Bioinformatic analysis

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

Singletons were then filtered and samples represented by <3610 reads were removed from the analysis in order to conserve a sufficient sequencing depth to characterize diatom community structure. In order to allow inter-sample comparisons, the remaining samples were then normalized to the same read number using the smallest read abundance amongst them. Diatom molecular inventories were obtained using the taxonomy of OTUs corresponding to the consensus taxonomy of DNA reads with a consensus confidence threshold over 80%.

For brevity, we often use "HTS" to refer to the whole process of deriving ecological status metrics by DNA metabarcoding, when contrasted with the process of obtaining them via light microscopical counts of diatom valves ("LM").

2.7. Ecological status class assignment

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

2.8. HTS correction factor application

Diatom species sometimes differ in the *rbc*L copy number per cell (depending on the number of gene copies per chloroplast and the number of chloroplasts per cell) and Vasselon et al. (2018a) found a strong correlation between *rbc*L copy number per cell and cell biovolume. They therefore suggested that a correction factor (CF) based on cell biovolume should be applied to the proportions of reads before making comparisons with valve counts (morphology). Accordingly, we applied Vasselon et al.'s (2018b) modified CFs (Rivera et al., 2020) to the HTS reads in order to assess their effectiveness in improving the DNA-based ecological status assessments; the CFs were extracted from Diat. barcode v7 (Rimet et al., 2019). The IPS values and the number of critical sites were compared between the LM inventory and both corrected and uncorrected HTS inventories.

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

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

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

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

3. Results

3.1. Light microscopy

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

3.2. Metabarcoding data

A total of 9,941,912 reads were obtained by MiSeq Illumina sequencing of the 164 samples. After quality filtering steps 3,081,893 reads were retained and clustered into 708 OTUs with an average of 78.2 per sample. The maximum and minimum numbers of OTUs per sample were 182 (comprised by a total of 21,654 reads) and 7 (comprised by a total of 14 reads), respectively. To allow inter-sample comparisons, samples were normalized to 3610 reads, representing the minimum number of reads per sample recorded after removal of 2 samples comprising 2033 and 14 reads respectively. The remaining, rarefied data comprised a total of 584,820 reads clustered into 615 OTUs, with an average of 61.1 OTUs per sample, the maximum and minimum being 137 and 10. The OTUs were assigned to a total of 148 taxa, of which 138 were species, with an average of 30.9 species per sample and ranging from 5 to 55 species per sample (Supplementary data). 18.3% of the reads (corresponding to the 51.4% of the total 615 OTUs) were not successfully classified at species level, the percentage of unclassified reads per sample varying from 0.2% to 71.6%. The ten most abundant species were *Achnanthidium minutissimum*, *Fistulifera saprophila*, *Planothidium victorii*, *Mayamaea permitis*, *Cocconeis placentula*, *Melosira varians*, *Craticula subminuscula*, *Gomphonema pumilum* var. *pumilum*, *Ulnaria ulna*, and *Nitzschia inconspicua* (Supplementary Fig. 1).

3.3. Comparison between molecular and morphological inventories

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

3.4. Ecological status comparison between approaches

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

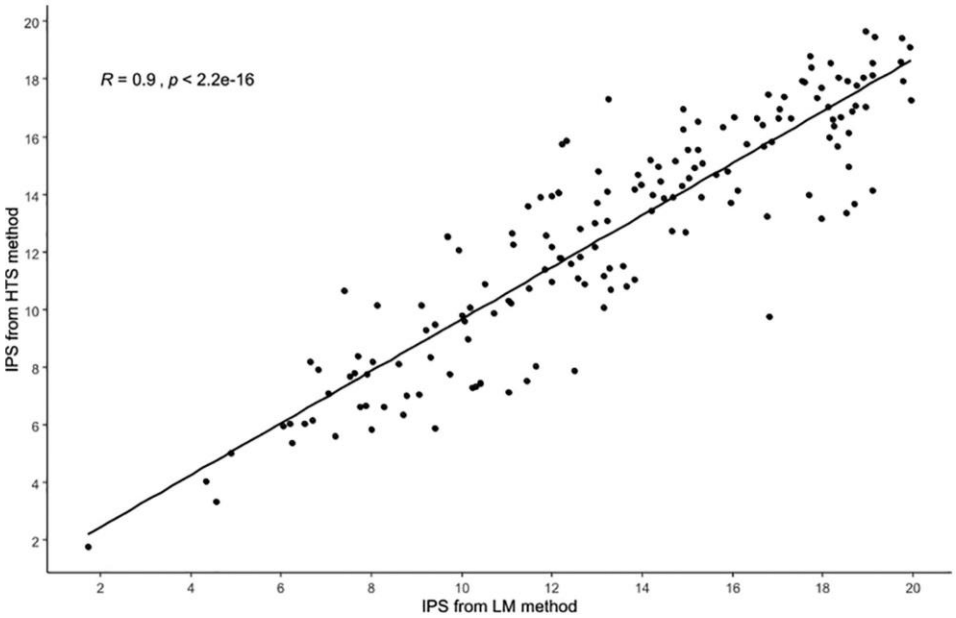


Fig. 2. Correlation of IPS values calculated from LM (x axis) and HTS (y axis) inventories considering the total 162 samples. Pearson's coefficient (R) and p-value are given.

ecological status in both approaches and 49 sites (30.2%) showed 1 class of difference (Table 1).

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

When the biovolume CF was applied to the molecular data, IPS values varied from 19.8 to 2.3 with an average of 12.4. The correlation between IPS values obtained from morphology and from CF corrected HTS was 0.92 (Pearson's R), so slightly higher than without applying the CF. However, the number of sites that shared the same ecological status decreased when the CF was applied (106 sites, representing 65.4% of the samples) and the number of sites that showed 1 and 2 classes of differences increased slightly [51 sites (31.5%) and 5 sites (3.1%) respectively]. Furthermore, and importantly, five new critical sites were obtained when CFs were applied, resulting in a total of 15 critical sites.

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

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

3.5. Species sensitivity analysis

3.5.1. All sites

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

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

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

Overall, the greatest contributions to IPS values were made by the most abundant species. However, lower abundance species (<5%) also made important contributions if their indicator values were very high or very low. Furthermore, and more importantly perhaps, though it is very easily overlooked, the contribution of these species (i.e. low abundance species, with very high or very low IPSS) was influenced by the IPS score of the whole sample. That is, species with very low IPSS values made a relatively greater contribution in samples where the overall IPS score was high (and the reverse was also true). An example is given by the sensitivity analysis results for our samples 76 and 138. In sample 76, *Achnanthidium minutissimum* was recorded (HTS) with a relative abundance of 3.66% and the sensitivity analysis showed a contribution of 0.97 towards the overall IPS (HTS) score of 7.76. In contrast, in sample 138, with an overall

IPS score of 18.05 and in which *A. minutissimum* was recorded (HTS) in almost the same relative abundance (3.77%) as in sample 76, the sensitivity analysis showed a much lower contribution (0.07) of the species to the overall IPS score (Supplementary data).

3.5.2. Critical sites

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

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

Nitzschia inconspicua is also an important contributor to the low IPS values of critical sites but mainly in the LM based assessments. The species was identified by LM in 8 of the 10 critical sites, and was the first-, second- and fourth-ranked species that most negatively affected the IPS in 2, 1 and 2 sites respectively, while with HTS, although it was identified in 9 of the critical sites, it was never amongst the 3 species that most negatively affected the IPS scores (Fig. 4, left). Hence it cannot be crucial for making sites critical. Discrepancies between methods in the contributions to IPS values in the species *Pleurosira laevis* and *Craticula subminuscula* were relevant in determining 2 critical sites. Both were recorded as the first-ranked species that most negatively contributed to IPS in one site by HTS, while they were never ranked amongst the 3 species that most negatively affected the IPS by LM (Fig. 4, left).

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

3.5.3. Critical sites resulting when applying CFs

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

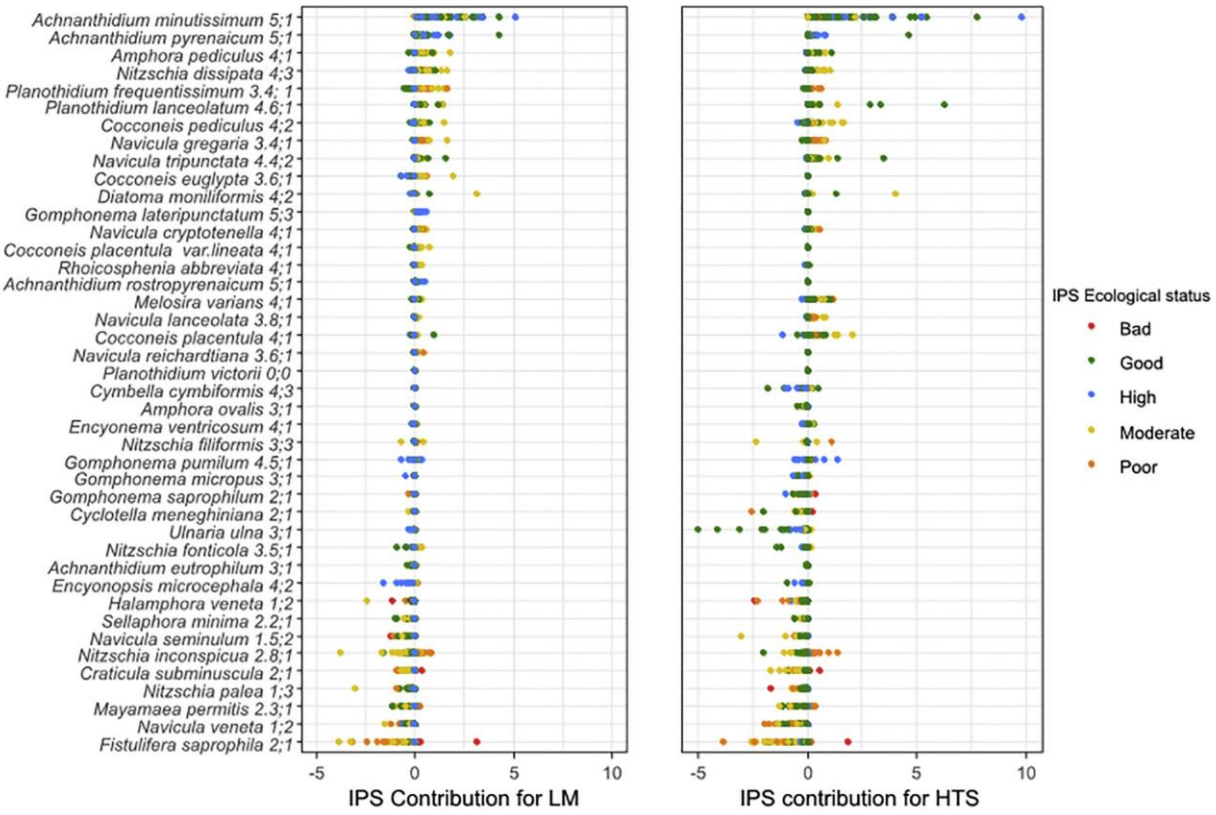


Fig. 3. Species sensitivity analysis (left for LM data, right for the HTS data) calculated by the "leave-one-out" method (see Material and methods) showing the IPS contributions (X axes) of the 35 most abundant species in the LM counts. Species are ordered according to the average of the IPS contributions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

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

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

1) A tendency to be underrepresented by HTS. This was shown in *Planothidium frequentissimum* and *Nitzschia inconspicua*, which were underrepresented in 97% and 90% respectively of the total samples where the species was identified by both methods.
2) The opposite tendency, overrepresentation by HTS, was shown in *Ulnaria ulna*. Of the total of 162 samples analysed, LM recorded the species in only 36 (22%) samples, while it was identified by HTS in 99 (61%). And in those samples where the species was recorded by both methods, it was overrepresented by HTS in 17 samples (61%).
3) Little or no bias overall in the relative abundances between the methods. This is the pattern shown by *Mayamaea permitis* and *Achnanthidium minutissimum*. For example, in the 108 samples where the species was identified by both methods, *M. permitis* was overrepresented by HTS in 50% and underrepresented in 50%. It is worth highlighting that in 9 of the 10 critical sites, *M. permitis* was overrepresented by HTS or not detected at all in LM. In the case of *A. minutissimum* there was a slight tendency to be underestimated by HTS (in 65% of samples where the species was identified by both methods).
4) The pattern shown by *Fistulifera saprophila*. On the one hand, there was a clear bias towards HTS, the species being recorded by this method in 136 samples (84%) out of the total of 162 analysed but in only 76 samples (46%) by LM. On the other hand, in the samples where both methods recorded this species, the pattern seemed to be of underrepresentation by HTS.

4. Discussion

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

Both the strong linear relationship between the ecological status results of both methods (morphology-LM and molecular-HTS), and also the fact that the intercept is close to zero, confirm the high potential of DNA metabarcoding as a new monitoring tool for the WFD assessment of Catalan rivers using benthic diatoms. Recent studies have also demonstrated this same potential for other regions of Europe (rivers in UK, France, Central Portugal and Switzerland: Kelly et al., 2018; Rivera et al., 2020; Mortágua et al., 2019; and Visco et al., 2015) and elsewhere (Mayotte Island rivers: Vasselon et al., 2017b). However, our study is the first to demonstrate the potential for rivers under a Mediterranean climate regime. Interestingly our study found:

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

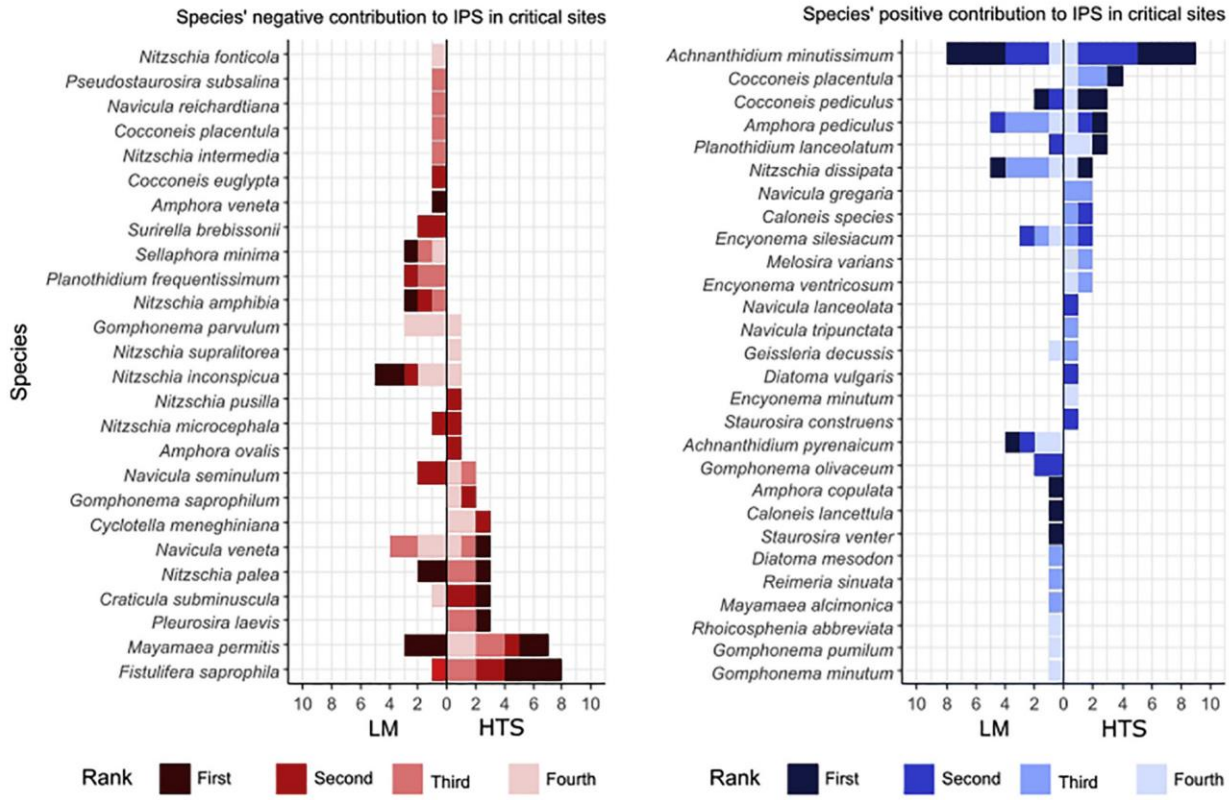


Fig. 4. Relative species contributions to the IPS value in the 10 critical sites (those sites whose status alters from Good/High by LM to Moderate/Poor/Bad by HTS). For each species the number of critical sites in which it was ranked the first, second, third or fourth most important contributor to the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

ii) These shared species accounted for a high percentage of total LM counts (80%) and HTS reads (72%).
iii) A high percentage (48.62%) of all the OTUs were successfully assigned at species level compared with those obtained previously in similar studies; for comparison, these were: 50.7% by Rivera et al. (2020; our calculation from their data); 41% by Rivera et al. (2018b); for lake Bourget); 35.7% by Vasselon et al. (2017b); 32% by Mortágua et al. (2019) and 30% by Keck et al. (2018).
iv) A very high correlation between the IPS values from both methods and also a high % of samples assigned to the same ecological class. To our knowledge, the highest correlation obtained in IPS values between methods is circa 0.83 (Pearson's R; Rivera et al., 2020) while ours is 0.92 after CFs and 0.90 without CFs (Pearson's R). Likewise, in the present work, the proportion of sites that fall into the same ecological status class regardless of the method used is 69.8%, considerably greater than has been obtained in other similar studies (Bailet et al., 2019; Mortágua et al., 2019; Rivera et al., 2020; Vasselon et al., 2017b).

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

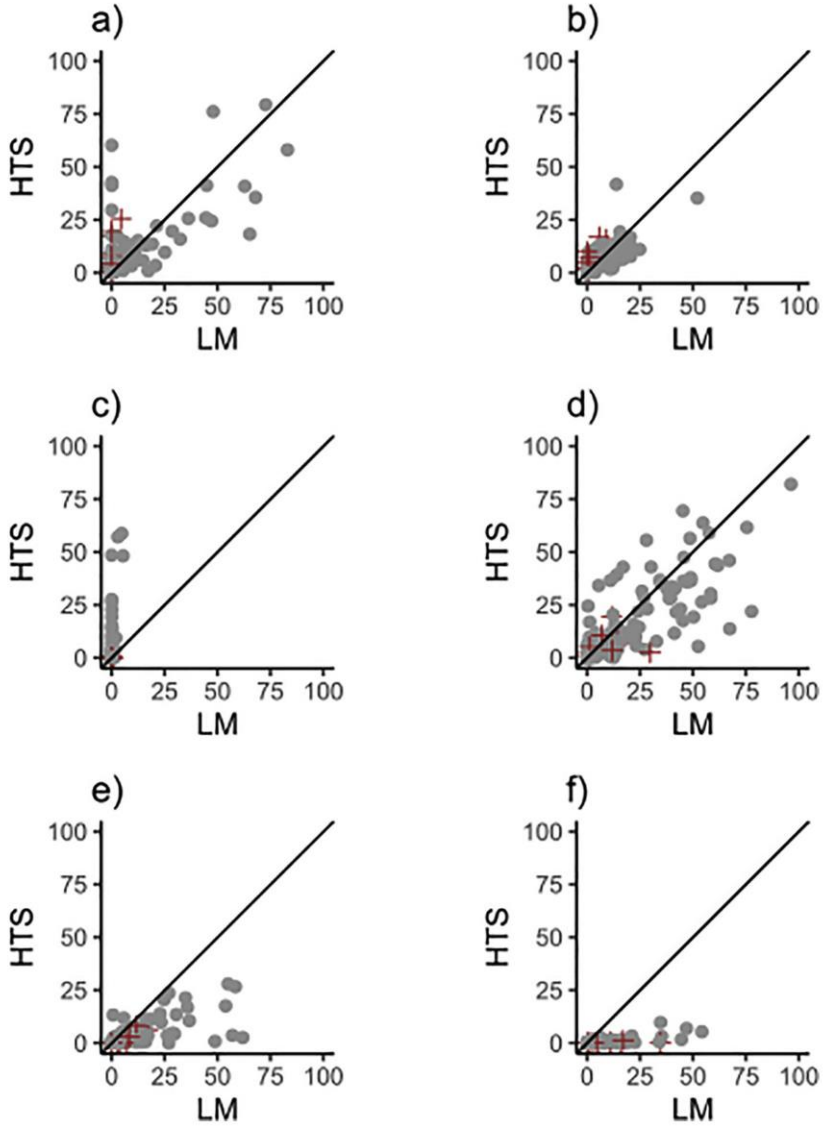


Fig. 5. Relative abundance comparisons between LM valve counts (x axis) and HTS reads (y axis) for methods of selected species. Cross symbol in black correspond to critical sites and circles in grey to non-critical sites. Species represented are the following: a) *Fistulifera saprophila* b) *Mayamaea permitis* c) *Ulnaria ulna* d) *Achnanthidium minutissimum* e) *Nitzschia inconspicua* e) *Planothidium frequentissimum*.

unacceptable Moderate, Poor or Bad status, i.e. the differences responsible for creating "critical" sites.

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

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

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

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

Another case of presumed misidentification by LM, this time partly because of taxonomic and nomenclatural changes, was observed in *Planothidium frequentissimum*, which was overrepresented in LM and indeed, scarcely recorded at all by HTS (Fig. 5; Supplementary Fig. 1). Our results suggest that *P. victorii* was frequently misidentified as *P. frequentissimum* during LM counting, since the relative abundance distribution of *P. frequentissimum* in LM agreed well with the corresponding distribution obtained for *P. victorii* in HTS (Supplementary Fig. 3). In such cases it can be difficult to determine which method (LM or HTS) is likely to be correct. However, in the present case, the sequences of *P. victorii* (and its taxonomic synonym, *P. caputium*) available in the DNA diatom reference database (Diat.barcode v7; Rimet et al., 2019) come from the same clones used to establish the species (Novis et al., 2012; Jahn et al., 2017) and the sequences of *P. frequentissimum* available in the reference library are also likely to have been reliably identified in the taxonomic revision by Jahn et al. (2017). Furthermore, the genetic diversity of these species is apparently well covered (Jahn et al., 2017). Hence, the IPS discrepancies found between the methods should be attributed, not to HTS identification error, but rather to the difficulties in distinguishing between *P. frequentissimum* and *P. victorii* in LM (due to the lack of easily seen morphological differences between them: Jahn et al., 2017), and/or to the difficulties of keeping up-to-date in routine LM counts with all the taxonomic changes being made (guides are often not affordable; the latest taxonomic changes are not always included, etc.). The importance of correctly identifying *P. frequentissimum* by either method lies in the fact that this species is relevant in determining Moderate ecological status because of its intermediate IPS sensitivity value (IPSS = 3.4), which leads to a negative or positive influence on the final IPS value, depending on the other species present (Fig. 3). *P. frequentissium* and *victorii* also illustrate another problem that arises when there are two (or more) taxa that are so similar morphologically that it is impossible to distinguish them during routine LM. This automatically means that we cannot use LM to determine whether they do or do not have the same ecological preferences; in fact, it will be only possible to determine the preferences of such cryptic or pseudocryptic taxa through combining HTS surveys with analyses of accompanying environmental data. Unfortunately, the *Planothidium* example is not unique; there are several small but abundant freshwater species that are similarly difficult or impossible to discriminate under LM, e.g. in *Nitzschia* (e.g. *N. inconspicua* and *N. soratensis*, Trobajo et al., 2013), or in *Amphora* (Levkov 2009). We are currently working on some of these to establish whether the different species/OTUs differ in their ecological preferences.

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

4.3. Pitfalls to be overcome

4.3.1. Gene copy numbers per cell affect the estimates of abundance of important species for WFD

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

As with *Planothidium frequentissimum*, the leave-one-out method revealed that *Nitzschia inconspicua* (IPSS = 2.8) showed a IPS contribution that shifted from positive (in sites classified as having Good or High ecological status) to negative (in sites classified with Bad or Poor ecological status), driving the IPS values towards Moderate ecological status (Fig. 3). The importance for biomonitoring of the relative abundance discrepancies in this species (clear underrepresentation of the taxon by HTS: Fig. 5) is that it will exaggerate the corresponding IPS values either negatively or positively, depending on the starting point. In those sites where *N. inconspicua* is abundant, the ecological status will be wrongly determined by HTS (relative to LM) in two ways: a) in those sites classified by LM as having Good or High ecological status, the IPS will be increased even more (i.e. IPS values overestimated); and, in contrast, b) in those sites classified by LM as having an unacceptable WFD level, the IPS will be lowered making them even worse (i.e. IPS values underestimated). Similar conclusions apply to *Craticula subminuscula*, which showed a similar IPS contribution pattern to *N. inconspicua* (Fig. 3) and was especially relevant for explaining one critical site (Fig. 4).

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

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

The good agreement between LM and HTS methods obtained in this study, in terms of the final IPS score, was likely due in large part to the fact that most of the IPS-determining and abundant benthic diatom species of the Catalan river basin district were represented in the DNA reference database used and could therefore be retrieved when the metabarcoding approach was applied. This reference database, Diat.barcode v7 (Rimet et al., 2019), is becoming widely used in diatom metabarcoding studies (Chonova et al., 2019; Mortágua et al., 2019; Rimet et al., 2018; Rivera et al., 2020) and is continuously curated by experts from different countries. However, it is far from complete and this could potentially be a source of IPS discrepancies between methods, if the missing species are sufficiently abundant and have a strong indicator value. In our case, the taxa amongst the species recovered by LM with a relative abundance >1% that were not identified by HTS, due to the lack of representative barcodes for them in the reference library, were *Cocconeis euglypta*, *Gomphonema lateripunctatum* and *Cocconeis placentula* var. *lineata* (Supplementary Fig. 1). Of these, *C. euglypta* was amongst the 10 species that contributed most to IPS (Fig. 3).

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

The last point we would make in relation to the reference database is that, at least within a limited geographical area and/or a relatively narrow range of water types, the number of "influential" species that must be included to avoid biases in the ecological status assessment may often be quite limited, as we demonstrate here (Fig. 3) and as shown also by Kelly et al. (2018, Fig. 6.10). Hence, although adding any species and genotypes to the reference database will always be useful, before isolating, culturing and Sanger-sequencing new clones it may be worth carrying out an IPS sensitivity analysis of existing LM-based abundance data, to objectively identify priority species for barcoding and hence avoid unnecessary work that may have negligible benefit for WFD biomonitoring.

4.4. Next priority: reference sites

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

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.138445.

CRediT authorship contribution statement

Javier Pérez-Burillo: Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Rosa Trobajo:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Valentin Vasselon:** Software, Writing - review & editing. **Frédéric Rimet:** Writing - review & editing. **Agnès Bouchez:** Writing - review & editing. **David G. Mann:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Supervision.

Declaration of competing interest

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

Acknowledgments

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

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

The authors also acknowledge support from the CERCA Programme/Generalitat de Catalunya. J. Pérez-Burillo acknowledges IRTA-Universitat Rovira i Virgili for his PhD grant (2018PMF-PIPF-22). The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division. This article is based upon work from COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program.

References

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

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

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

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

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

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

CEN, 2014b. CEN_EN 14407: Water Quality - Water Quality Guidance Standard for the Identification, Enumeration and Interpretation of Benthic Diatom Samples From Running Waters. pp. 1–16.

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

Chonova, T., Kurmayer, R., Rimet, F., Labanowski, J., Vasselon, V., Keck, F., Illmer, P., Bouchez, A., 2019. Benthic diatom communities in an alpine river impacted by waste water treatment effluents as revealed using DNA metabarcoding. Front. Microbiol. 10, 1–17. https://doi.org/10.3389/fmicb.2019.00653.

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

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

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

Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18, 1299–1309. https://doi.org/10.1111/1755-0998.12919.

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

Kelly, M., Bennett, C., Coste, M., Delgado, C., Delmas, F., Denys, L., Ector, L., Fauville, C., Ferréol, M., Golub, M., Jarlman, A., Kahlert, M., Lucey, J., Ní Chatháin, B., Pardo, I., Pfister, P., Picinska-Faltynowicz, J., Rosebery, J., Schranz, C., Schaumburg, J., Van Dam, H., Vilbaste, S., 2009. A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. Hydrobiologia 621, 169–182. https://doi.org/10.1007/s10750-008-9641-4.

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

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F., Bouchez, A., 2013. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms. Mol. Ecol. Resour. 13, 607–619. https://doi.org/10.1111/1755-0998.1210.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.M., Humbert, J.F., Bouchez, A., 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshw. Sci. 33, 349–363. https://doi.org/10.1086/675079.

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

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

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

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

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

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

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

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

Mann, D.G., Crawford, R.M., Round, F.E., 2016. Bacillariophyta. In: Archibald, J.M., Simpson, A.G.B., Slamovits, C.H., Margulis, L., Melkonian, M., Chapman, D.J., Corliss, J.O. (Eds.), Handbook of the Protists. Springer, Cham, New York, pp. 1–62. https://doi.org/10.1007/978-3-319-32669-6_29-1.

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

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

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

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

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. Org. Divers. Evol. 18, 51–62. https://doi.org/10.1007/s13127-018-0359-5.

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

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

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

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

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

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

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

Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A., 2017a. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? Freshw. Sci. 36, 162–177. 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). Ecol. Indic. 82, 1–12. https://doi.org/10.1016/j.ecolind.2017.06.024.

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

Vasselon, V., Rimet, F., Bouchez, A., 2018b. Rsyst::diatom_rbcl_align_312bp Database: A Database Adapted to DNA Metabarcoding (Version v7: 23-02-2018). Portail Data Inra, V1 https://doi.org/10.15454/HYRVUH.

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

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. Environ. Sci. Technol. 49, 7597–7605. https://doi.org/10.1021/es506158m.

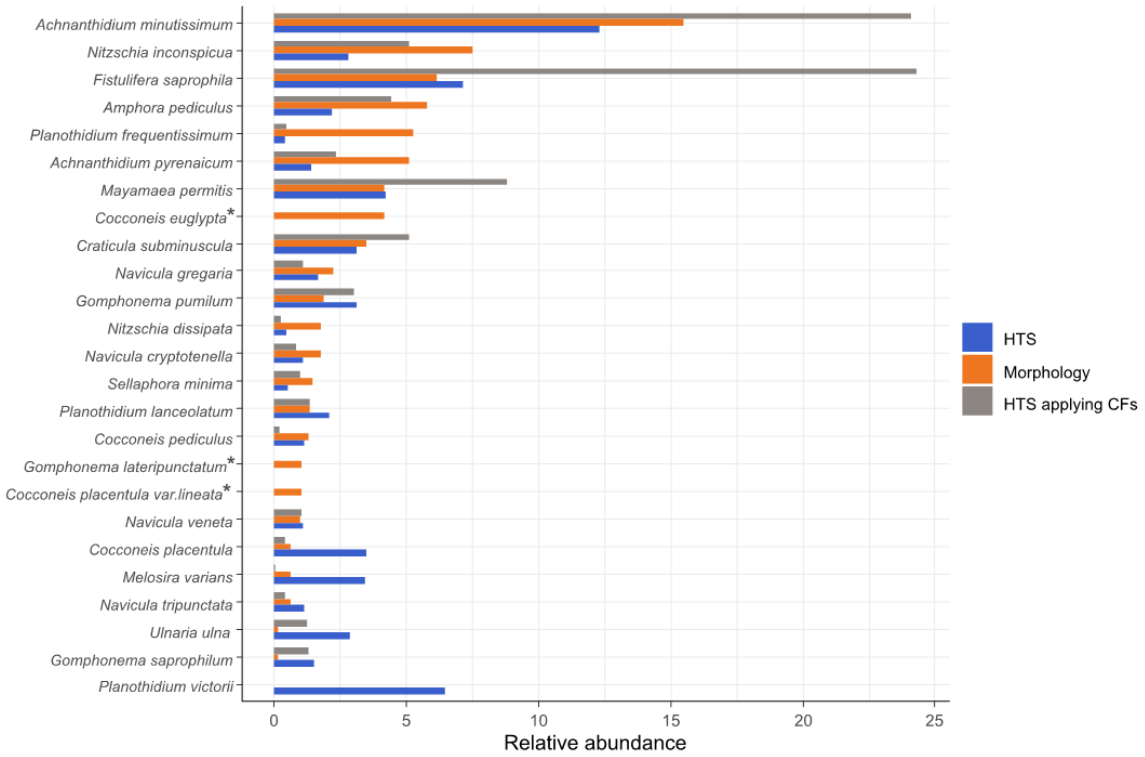
Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267. https://doi.org/10.1128/AEM.00062-07.

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

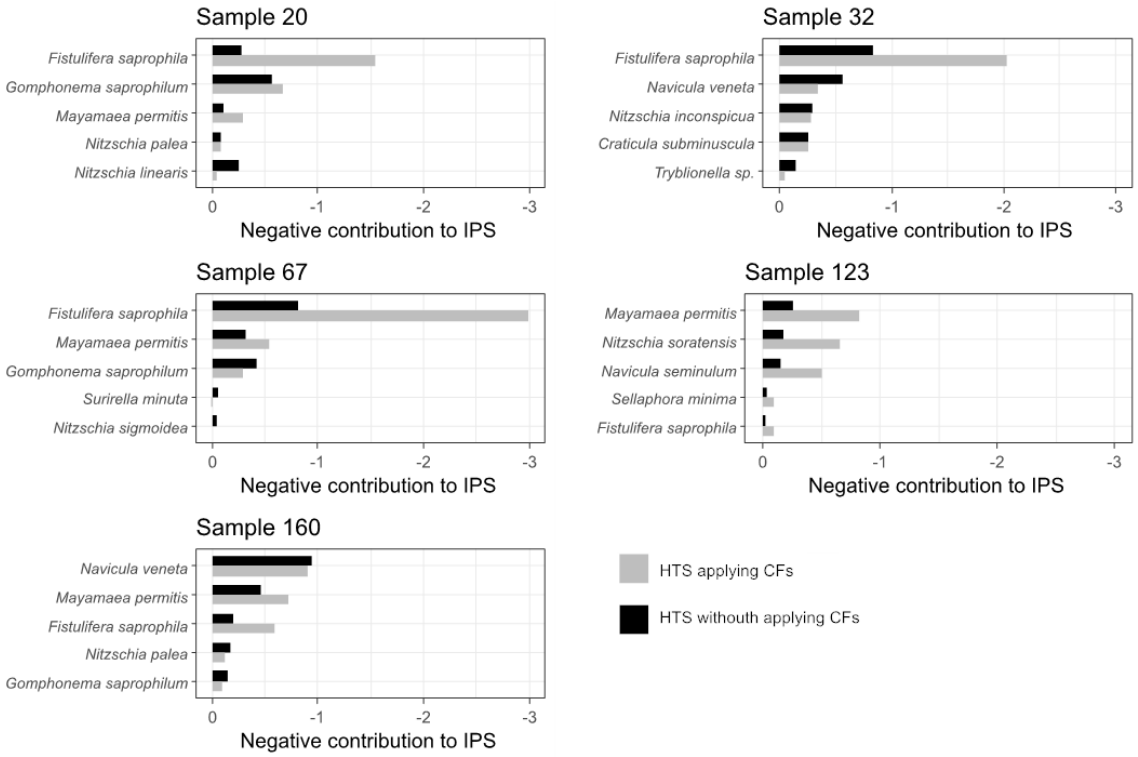
Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., Latala, A., 2013. Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. Diatom Research 28, 431–443. https://doi.org/10.1080/0269249X.2013.833136.

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

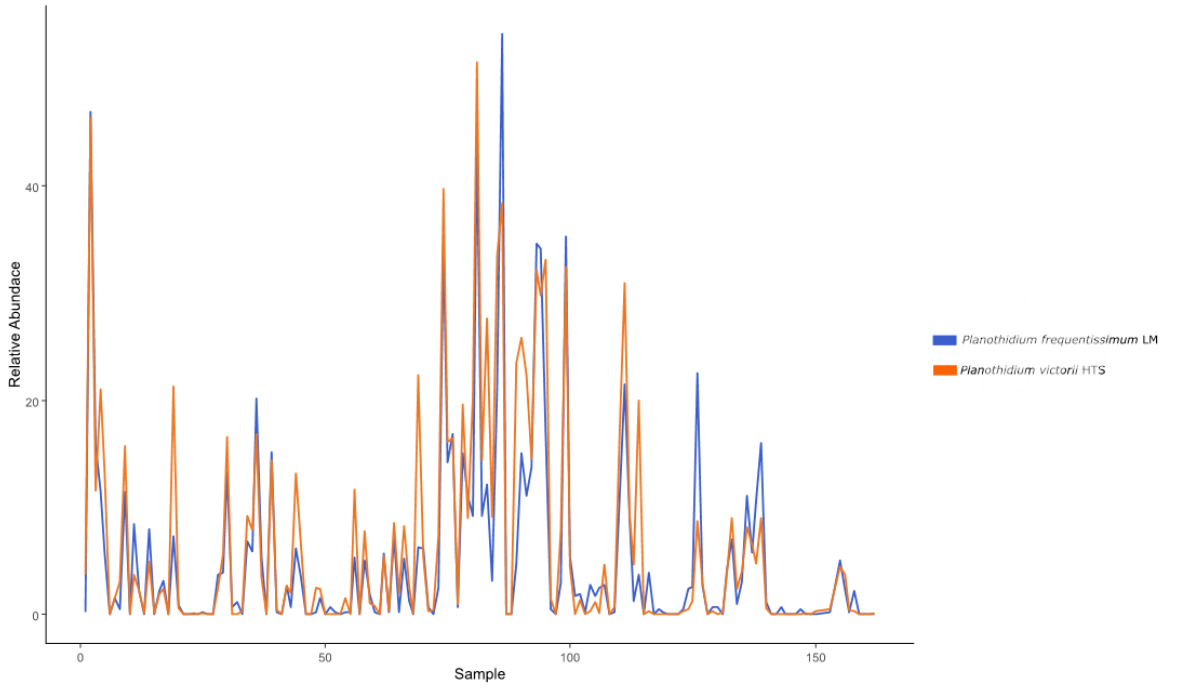
Supplementary material



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



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



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

Chapter 2

Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters

Pérez-Burillo, J., Valoti, G., Witkowski, A., Prado, P., Mann, D. G., Trobajo, R., 2022.

Mar. Pollut. Bull. 174, 113183.
https://doi.org/10.1016/j.marpolbul.2021.113183.

Marine Pollution Bulletin 174 (2022) 113183

Contents lists available at ScienceDirect

Marine Pollution Bulletin

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

Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters

Javier Pérez-Burillo [a,b], Greta Valoti [c], Andrzej Witkowski [d], Patricia Prado [a], David G. Mann [a,e], Rosa Trobajo [a,*]

[a] 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
[b] Departament de Geografia, Universitat Rovira i Virgili, C/ Joanot Martorell 15, E43500 Vila-seca, Tarragona, Spain
[c] Università Politecnica delle Marche, Piazza Roma, 22, IT60131 Ancona, Italy
[d] Institute of Marine and Environmental Sciences, University of Szczecin, Mickiewicza 16a, 70-383 Szczecin, Poland
[e] Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, Scotland, UK

ARTICLE INFO

Keywords:
Diversity
Environmental DNA
Epibiotic
Microalgae
Stramenopiles
rbcL

ABSTRACT

We investigated the advantages and disadvantages of light microscope (LM)-based identifications and DNA metabarcoding, based on a 312-bp *rbcL* marker, for examining benthic diatom communities from Mediterranean shallow coastal environments. For this, we used biofilm samples collected from different substrata in the Ebro delta bays. We show that 1) Ebro delta bays harbour high-diversity diatom communities [LM identified 249 taxa] and 2) DNA metabarcoding effectively reflects this diversity at genus- but not species level, because of the incompleteness of the DNA reference library. Nevertheless, DNA metabarcoding offers new opportunities for detecting small, delicate and rare diatom species missed by LM and diatoms that lack silica frustules. The primers used, though designed for diatoms, successfully amplified rarely reported members of other stramenopile groups. Combining LM and DNA approaches offers stronger support for ecological studies of benthic microalgal communities in shallow coastal environments than using either approach on its own.

1. Introduction

Coastal ecosystems are ecologically important because they are highly productive areas that harbour a great diversity, which is reflected in many types of ecological communities found in these systems, such as seagrass beds, sandflat communities, coral and bivalve reefs (Cloern et al., 2013). These ecosystems are also important from a socio-economic point of view as they provide numerous ecosystem services and contribute importantly to the global total (Costanza et al., 2014). Benthic diatom communities constitute an important component of these systems because of their large contribution to total production (MacIntyre et al., 1996). A recent study of seagrass beds in shallow systems (Cox et al., 2020) has shown that the contribution of diatoms can be over 80% of benthic production and that without them the seagrass beds can be net heterotrophic. In addition, they have a major role in the stabilization of sediments, thanks to the production of extracellular polymeric substances (EPS), and consequently they regulate nutrient fluxes and other biogeochemical processes (Cahoon, 1999, Sundbäck and Granéli, 1988; Sundbäck et al., 1991; Triska and Oremland, 1981; Trobajo and Sullivan, 2010). They can be found in or attached to different substrata, such as the surface of sediments (epipelon), sand grains (epipsammon), seagrasses, macroalgae, and microalgae (epiphyton), or the surface of animals including the shells of molluscs (epizoon). Each of these community types can be very species-rich (Round, 1971) and it has been shown that within communities, different species can play different roles; for instance, in tidal habitats epipelic species show differences in photophysiology and migration activity (Underwood et al., 2005). Hence, it is crucial to combine system-wide estimates of benthic diatom contributions to primary production with an understanding of the roles and functioning of the species comprising these communities. However, the morphological identifications of diatoms at the species level are difficult and require expertise

* Corresponding author.
E-mail address: rosa.trobajo@irta.cat (R. Trobajo).

https://doi.org/10.1016/j.marpolbul.2021.113183
Received 28 September 2021; Received in revised form 16 November 2021; Accepted 20 November 2021

in taxonomy. This is especially true for shallow coastal environments (Mann et al., 2016; Trobajo et al., 2004), despite their ecological and economic importance.

DNA metabarcoding has proved to be a reliable method for studying species diversity from environmental samples (Deiner et al., 2017) and has emerged as an alternative to light microscope-based identifications (LM) due to its speed, reproducibility, and cost (Kermarrec et al., 2014; Zimmermann et al., 2015). It has been broadly tested for freshwater ecological assessment based on benthic diatoms (e.g. Bailet et al., 2019; Kelly et al., 2020; Mortágua et al., 2019; Pérez-Burillo et al., 2020; Vasselon et al., 2017) and for biodiversity studies (e.g. Stoof-Leichsenring et al., 2020; Rimet et al., 2018). DNA metabarcoding has also been applied in marine environments, especially in phytoplankton studies (e.g. De Luca et al., 2021; Malviya et al., 2016; Piredda et al., 2018), but rarely to the phytobenthos of coastal areas, which are very productive and species rich. Exceptions include studies of US saltmarshes (Plante et al., 2021a, 2021b), intertidal sediments in Korea (An et al., 2020), a eutrophic estuary in South Africa (Nunes et al., 2021), and sea turtle biofilms (Rivera et al., 2018).

In the context of ongoing research into the biodiversity and functioning of Mediterranean shallow coastal habitats (e.g. Benito et al., 2015; Carballeira et al., 2017; Prado, 2018, Prado et al., 2020; Rovira et al., 2009), we set out to study the benthic diatom communities in these poorly known systems through the combined use of DNA metabarcoding, based on a 312-bp *rbcL* marker, and LM-based identifications. Sampling aimed for a selection of the different benthic communities dwelling on sediments, seaweeds, seagrasses and molluscs (i.e. epipelic, epiphytic and epizoic/epilithic communities) in coastal areas of the Ebro delta. In particular, Ebro Delta bays sustain a very important shellfish aquaculture of Japanese oyster and Mediterranean mussel (Ramón et al., 2005), providing important substrata for biofilm development. Besides, the area holds one of the last remaining populations of fan mussel (*Pinna nobilis*) after major mass mortality events throughout the Mediterranean (Prado et al., 2014, 2021). Moreover, beds of the seagrasses *Cymodocea nodosa* and *Zostera noltii* are present in the area. In this paper we evaluate the advantages and disadvantages of each survey approach – morphological and molecular – and we assess whether the *rbcL* marker, which was originally developed for diatom biomonitoring of freshwaters, is equally useful in marine environments, where the diversity of related groups of ochrophyte microalgae and macroalgae is much greater.

2. Material and methods

2.1. Study area and sampling collection

Nine biofilm samples were taken in Alfacs and Fangar bays offshore from the Ebro Delta on the Mediterranean coast of the Iberian Peninsula (Fig. 1). The bays constitute semi-enclosed estuarine water bodies that receive freshwater inputs, rich in nutrients and organic matter, from rice fields that border both bays, which have led to eutrophication (Llebot et al., 2011; Prado, 2018). Alfacs Bay encompasses an area of 50 km^2 with an average depth of ~3 m and a maximum of 6 m. Sampling in this bay was conducted by wading in a semi-sheltered area at ca. 60 cm depth near the southern shore, where the seagrass *Cymodocea nodosa* and/or the macroalga *Caulerpa prolifera* constitute the dominant benthic habitat, and where there is also an important population of *Pinna nobilis* (Prado et al., 2014, 2020, 2021). Fangar Bay is smaller, occupying 12 km^2, and has an average depth of 2 m and a maximum of 4 m. Sampling in this bay was conducted within farms of the introduced Pacific oyster *Crassostrea gigas*, located in the southern area of the Bay. Physico-chemical information at each sampling site is shown in Table 1.

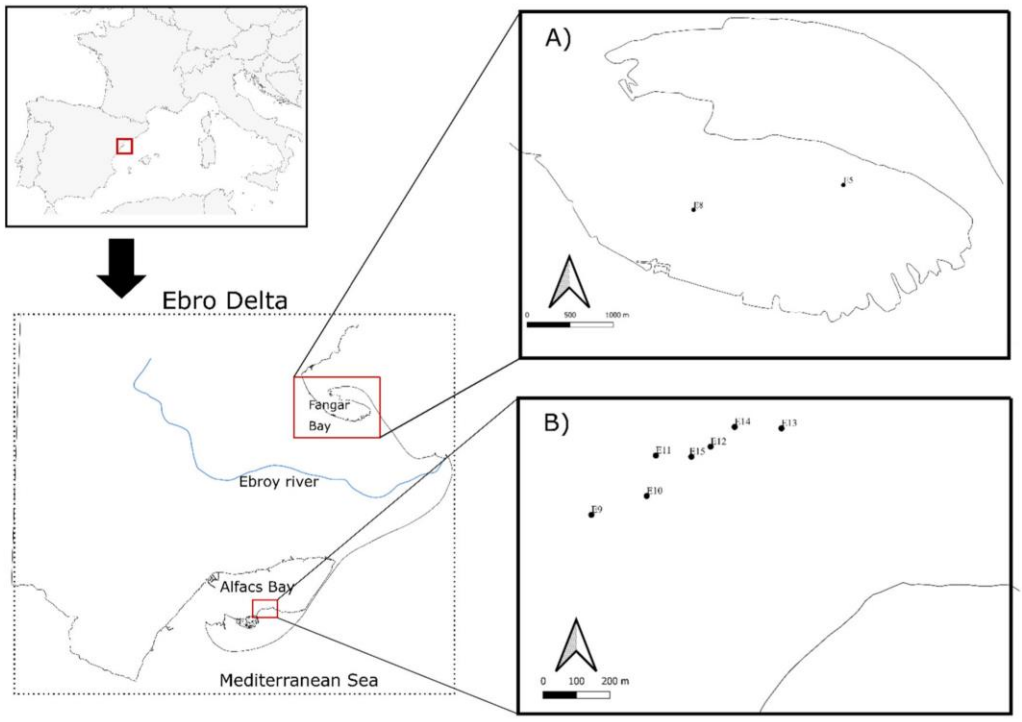


Fig. 1. Location of Ebro Delta (NE of Spain) and samples sites of Fangar (A) and Alfacs (B) bays. Two biofilm samples (E5 and E8) were taken from the surface of *Crassostrea gigas* individuals located in the Fangar Bay (A). Seven biofilm samples were taken from the surface of *Pinna nobilis* (E9, E11 and E12), *Cymodocea nodosa* (E14) and *Caulerpa prolifera* (E15) individuals located in the Alfacs Bay (B). The sediment samples (E10 and E13) were taken from the sediment adjacent to specimens of *P. nobilis* located in the Alfacs Bay (B).

Table 1
Physico-chemical measurements registered in the different sampling sites at the time and date of sampling. Note that there are samples (E9 & E10; E12-E15) with the same physico-chemical data; this is because these samples were collected on the same date from the same small area, being separated from each other by distances in the order of 10s of meters.

Sample	Sampling date	Sampling time	Water temperature (°C)	Salinity (g/L)	pH	Dissolved oxygen (mg/L)	% Dissolved oxygen (mg/L)
E5 - *Crassostrea gigas*	03/03/2020	12:35	11	35.87	7.95	7.49	85.1
E8 - *Crassostrea gigas*	10/03/2020	13:00	12.3	36.99	8.05	6.97	82.2
E9 - *Pinna nobilis* biofilm E10 - *Pinna nobilis* sediment	12/03/2020	12:15	14.9	36.5	8.11	7.79	96.7
E11 - *Pinna nobilis* biofilm	12/03/2020	14:30	16.6	37.27	8.24	8.65	111.6
E12 - *Pinna nobilis* biofilm E13 - *Pinna nobilis* sediment E14 - *Cymodocea nodosa* E15 - *Caulerpa prolifera*	12/03/2020	15:50	15.4	36.24	8.14	7.72	97.6

Samples were collected in March 2020 and comprised seven biofilm samples and two sediment samples. Five of the biofilms were taken from the shell surfaces of *P. nobilis (*three different individuals from Alfacs Bay separated by distances in the order of 10s of metres) and from *Crassostrea gigas* (two individuals from Fangar Bay). The other two biofilm samples were taken from the surfaces of *Caulerpa prolifera* and *Cymodocea nodosa*, respectively, both from the same area of Alfacs Bay as those of *P. nobilis*. Finally, the two sediment samples were collected from the surface sediments (ca. 1–2 cm) immediately adjacent to specimens of *P. nobilis* and transported to the laboratory within small containers. For collecting the biofilm samples, the surface of the organisms was scraped using a different toothbrush for each specimen. Each sample was divided into two aliquots and preserved in ethanol (to a final concentration ≥ 70%). One was used for morphological examinations and the other for DNA metabarcoding analysis.

2.2. Microscopical analysis

Samples for morphological analysis were cleaned using concentrated (37%) hydrogen peroxide (H_2O_2). However, prior to hydrogen peroxide, samples were treated with few millilitres of 10% HCl to remove carbonates. After the reaction with carbonates ceased, samples were washed several times with deionized water. Thereafter samples were boiled with hydrogen peroxide for a few hours to oxidize the organic matter and then washed several times with deionized water at 24 h intervals. Cleaned diatomaceous suspension was dropped onto cover slips and left at room temperature overnight to dry. Permanent slides were mounted with Naphrax (Brunel Microscopes: http://www.brunelmicroscopes.co.uk/), which has a high refractive index. Diatom analysis was performed using a Leica DMLB microscope equipped with 100× PlanApo objective (n.a. 1.4). Approximately 300 to 400 valves were counted in each sample. Problems in identification were resolved with scanning electron microscopy (SEM). For SEM examination, a drop of the cleaned sample was filtered onto Whatman Nuclepore polycarbonate membranes (Fisher Scientific, Schwerte, Germany). Filters were air-dried overnight, mounted onto aluminium stubs and coated with 5 nm of gold. Samples were analysed with an ultra-high field emission Hitachi SU 8020 instrument at West Pomeranian University of Technology in Szczecin.

2.3. DNA extraction, PCR amplification and high-throughput sequencing (HTS) library preparation

A volume of 2 mL of each sample was centrifuged at 4 °C and 11,000 g for 20 min. Ethanol present in the supernatant was removed and the DNA contained in the remaining pellet was extracted using the commercial DNA kit Macheray–Nagel NucleoSpin® Soil extraction kit (MN-Soil). A short *rbcL* region of 312 bp constituted the DNA marker and this was amplified by PCR using an equimolar mix of the modified versions of the Diat_rbcL_708F (forward) and R3 (reverse) primers given by Vasselon et al. (2017). In order to prepare the HTS library using a 2-step PCR strategy, a part of the P5 (TCGTCGGCAGCGTCAG ATGTGTATAAGAGACAG) and P7 (GTCTCGTGGGCTCGGAGATGTGTATA AGAGACA) Illumina adapters was included at the 5′ end of the forward and reverse primers respectively. PCR1 reactions for each DNA sample were performed in triplicate using 1 μL of the extracted DNA in a final volume of 25 μL. The conditions and the reaction mix of the PCR1 were as described in Vasselon et al. (2017). All three PCR1 replicates were pooled and sent to "Plateforme Génome Transcriptome" (PGTB, Bordeaux, France), where the PCR1 products were purified and used as a template for a second round of PCR (PCR2), with Illumina-tailed primers targeting the half of P5 and P7 adapters. The resulting dual-indexed amplicons were pooled for sequencing on an Illumina MiSeq platform using a V2 paired-end sequencing kit (250 bp × 2).

2.4. Bioinformatic analysis

The sequencing facilities performed the demultiplexing of all the samples providing two fastq files per sample, one corresponding to the forward reads (R1) and one to the reverse reads (R2). Primers from all the demultiplexed MiSeq reads were removed by cutadapt (Martin, 2011) and the resulting R1 and R2 reads were processed together using the R package *DADA2* (Callahan et al., 2016). R1 reads were truncated to 225 bases and R2 to 180 bases based on their quality profile (median quality score < 30). Reads with ambiguities or an expected error (maxEE) > 2 were discarded. The DADA2 denoising algorithm was then applied to determine an error rates model to infer Amplicon sequence variants (ASVs). Chimeric ASVs were detected and discarded using the "removeBimeraDenovo" function. The taxonomic affiliations of the ASVs was determined using the database "A ready-to-use database for DADA2: Diat.barcode_rbcL_312bp_DADA2" (Chonova et al., 2020), which is derived from the curated diatom reference library Diat.barcode v9 (Rimet et al., 2019, available at https://www6.inra.fr/carrtel-collection_eng/Barcoding-database and at https://data.inrae.fr/file.xhtml?persistentId=doi:10.15454/TOMBYZ/IEGUXB&version=10.0); the naïve Bayesian classifier method (Wang et al., 2007) was used, with 85% set as the minimum confidence threshold. The taxonomy of unclassified ASVs was checked using the Basic Local Alignment Search Tool (BLAST) against the Nucleotide database of NCBI GenBank, with standard settings (Camacho et al., 2009). Taxonomy was assigned keeping taxa with a percentage of identity higher than 97%. To allow inter-sample comparisons, all samples were resampled to the minimum number of reads recorded in any single sample (5427 reads) using the R package *phyloseq* (McMurdie and Holmes, 2013).

2.5. Data analyses

For assessing the effectiveness of the two methods in identifying

taxa, the percentages of reads or cells identified to species and genus were determined. Furthermore, the percentages of species and genera recorded molecularly that were also identified by the morphological approach and vice versa were calculated. For other statistical analyses, a rarefied molecular inventory was used. To compare diatom diversity between methods and sampling sites, the Shannon–Wiener index was calculated (based on natural logarithms), using the relative abundances of taxa from the corresponding morphological and molecular inventories. The Sørensen index, based on presence/absence data, was also calculated to evaluate the similarities in diatom communities between samples. To visualize patterns in taxon composition (in LM and DNA metabarcoding inventories) among samples, non-metric multidimensional scaling (NMDS) was used, based on Bray–Curtis dissimilarity matrices on ASV, species and genus relative abundance. The correlation between the distance matrices generated by both methods, using diatom species relative abundances, was evaluated by computing a Mantel test (with 999 permutations). Statistically significant differences in diatom community composition at the ASV-, species- and genus level regarding the type of substratum (i.e. biofilm samples taken from *P. nobilis*, *Crassostrea gigas*, *Caulerpa prolifera* and *Cymodocea nodosa* and samples collected from the sediment adjacent to *P. nobilis*) were evaluated through a permutation multivariate analysis of variance (PERMANOVA). To identify the taxa that accounted for most of the dissimilarities between the LM and DNA metabarcoding inventories, an analysis of similarity percentages (SIMPER) was performed on both species and genus relative abundance. The R package *vegan* (Oksanen et al., 2020) was used for performing all these analyses.

2.6. Phylogenetic analyses of non-diatom ASVs

Although the primers used here were designed specifically for freshwater diatom biomonitoring (Vasselon et al., 2017), they do nevertheless sometimes amplify *rbcL* from other groups of algae. For example, the ASV with most reads in the 2017 Catalan rivers dataset used by Pérez-Burillo et al. (2020, 2021) was an unknown green alga related to *Nautococcus* and *Oophila* (Chlorococcaceae), which was present in 116 of 164 samples analysed; Ochrophyta classes (sensu Adl et al., 2019) were also represented, including Xanthophyceae (e.g., *Vaucheria*) and Eustigmatophyceae (e.g., *Neomonodus*). In marine habitats the diversity of ochrophytes and red algae is much greater than in freshwaters and different green algal groups are present. Indeed, preliminary blastn analysis of our reads that were not assigned to any diatom taxon by the Bayesian classifier indicated that some ASVs belonged to different classes or phyla of algae. The majority (both in terms of ASVs and reads) were ochrophytes and we therefore performed phylogenetic analyses of the non-diatom ASVs together with GenBank sequences of selected ochrophytes to elucidate their affiliations and phylogeny. To do this, we assembled the sets of *rbcL* sequences used by Graf et al. (2020) and Wetherbee et al. (2021) and added representatives of other ochrophyte classes (particularly Chrysophyceae and Synurophyceae) to provide a wide coverage of the group. We also added further Phaeophyceae that blastn analysis indicated were close to some ASVs. The sequences were aligned by eye in Mega X (Kumar et al., 2018) after initial use of Muscle (Edgar, 2004), truncated to remove ragged ends and regions poorly represented among the taxa analysed, and exported for phylogenetic analysis to RAxML (Stamatakis, 2014), as implemented in raxmlGUI v. 2.0 (Edler et al., 2021). A Maximum Likelihood (ML) tree was constructed with the alignment partitioned by codon position, using a GRT-Gamma model; 1000 replicates were made for the bootstrap analysis. The tree was visualized, midpoint-rooted, and prepared for publication using iTOL (https://itol.embl.de) (Letunic and Bork, 2021).

The affiliations of the few non-ochrophyte ASVs (Chlorophyta and Rhodophyta) were assessed by blast of NCBI GenBank.

2.7. Trait classification

Alongside analyses of diatom communities based on species composition we also classified the different diatom taxa identified (either microscopically or molecularly) according to their ecological guilds and growth-forms. For this we largely followed Passy (2007) and Rimet and Bouchez (2012) but we split the original euplanktonic group defined by Passy (2007) into planktonic and tychoplanktonic groups. Thus, the resulting growth-forms were: high-profile, low-profile, motile, planktonic and tychoplanktonic. For some taxa, Passy and Rimet & Bouchez provided no information (their focus was on freshwater communities) and for these the growth-form was inferred on the basis of information in Round et al. (1990) and expert knowledge.

3. Results

3.1. Morphological inventory

A total of 249 diatom taxa (including varieties, forms, and species) were identified, the number per sample ranging from 40 to 75, with an average of 58.9. The most abundant diatom taxa were *Navicula* sp. 4, *Amphora helenensis*, *Amphora* cf. *helenensis*, *Cocconeis scutellum* var. *posidoniae*, *Navicula normaloides*, *Nanofrustulum shiloi*, *Cyclotella choctawhatcheeana*, *Navicula normalis*, *Cocconeis scutellum* and *Berkeleya fennica* (Supplementary Table 1). The 249 taxa recorded represented 73 different genera and 128 different species. The number of species identified per sample ranged from 25 to 47, with an average of 36.4. A total of 122 taxa (49%) could not be identified at species level but only at genus level.

Low profile and motile growth forms, mainly represented by species from *Amphora*, *Cocconeis*, *Navicula* and *Nitzschia*, were the predominant groups in all the samples, followed by the high-profile group (Fig. 2a), in which *Berkeleya* was the most abundant genus (Supplementary Table 1). The planktonic and tychoplanktonic groups were less represented and not identified in all the samples. They were more abundant in *P. nobilis* samples (Fig. 2a) and were represented mainly by *Cyclotella*.

3.2. Molecular inventory

MiSeq Illumina sequencing produced a total of 176,248 raw DNA reads from the nine samples. After processing the reads through the DADA2 pipeline, 139,815 reads remained, belonging to 682 ASVs, with an average of 145.1 ASV per sample (Supplementary Table 2). The maximum number of ASVs per sample was 214 (in sample E12 – *Pinna nobilis* biofilm) and the minimum 72 (in sample E15 – *Caulerpa prolifera*). 127 ASVs were classified at diatom species or genus level using the Bayesian classifier, on the basis of Diat.barcode v9 (i.e., bootstrap values at species level $\geq$ 85%). The taxonomic positions of 46 further ASVs that were not allocated to species by the Bayesian classifier (i.e., their bootstrap values at species level were $<$ 85%) were resolved by blastn on GenBank and allocated to species using a percentage of identity $>$97% as threshold. Finally, an additional 181 ASVs that did not fulfil either of the two previous criteria were classified at genus level by using a combination of expert knowledge and examination of the most similar sequences in GenBank.

Altogether, the three approaches described above allowed a total of 354 of the 682 ASVs to be classified to species or genera of diatoms, with 69 species and 73 genera identified. After rarefaction was applied, these numbers were very slightly reduced (the total number of 354 diatom ASVs was reduced to 338 ASVs, accounting for 51.2% of the total of rarefied reads, comprising 69 fully identified species and 71 genera) (Supplementary Table 2). The number of species per sample ranged from 21 to 54, with an average of 37.3, and the ten most abundant diatom taxa in the inventory were: *Thalassiosira profunda*, *Achnanthes longipes*, *Berkeleya fennica*, *Nanofrustulum shiloi*, *Navicula* sp., *Cyclotella* sp., *Haslea howeana*, *Seminavis* cf. *robusta*, *Craspedostauros constricta* and *Licmophora*

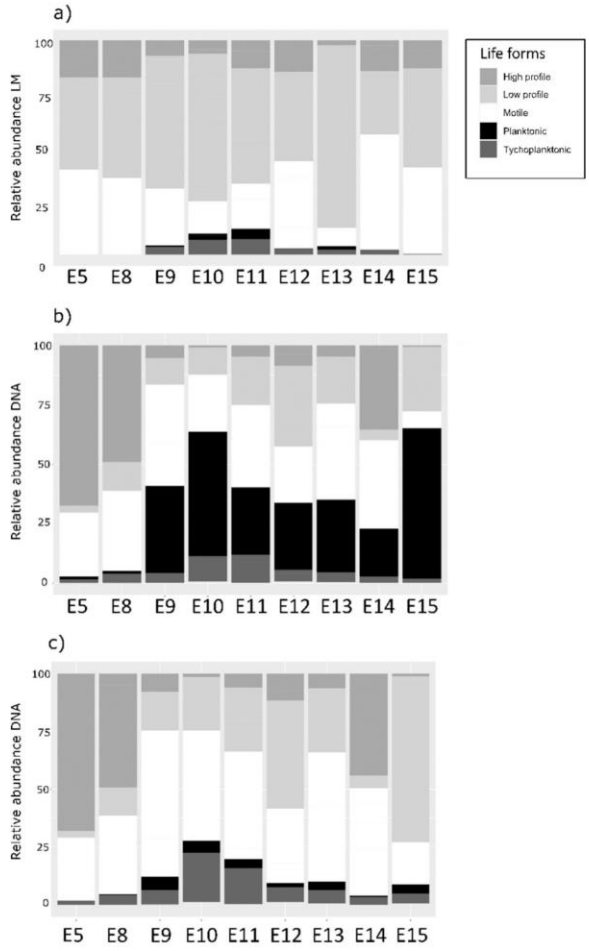


Fig. 2. Relative abundance comparison of diatom growth forms between LM (a), DNA metabarcoding inventories (b) and DNA metabarcoding without considering the planktonic species *Thalassiosira profunda* (c).

paradoxa.

Among the ASVs were several genera and species that were missed or poorly represented in the morphological dataset. One important factor was the lower detection limit of metabarcoding: even in the least productive sample (*Caulerpa* epiphytes) >5000 reads were obtained, offering the possibility to detect rare species undetectable among the c. 400 specimens per sample identified morphologically. It was noticeable too that some ASVs represented species that have very delicate or small cells. Several of these are rarely evident in any cleaned material, such as *Cylindrotheca* and some species of Cymatosirales (comprising *Arcocellulus*, *Extubocellulus*, *Papiliocellulus* and *Minutocellus* in our material). *Cylindrotheca* species are very lightly silicified and often destroyed by oxidative cleaning (Round et al., 1990). Only one sample was recorded to contain *Cylindrotheca* by LM analysis but eleven ASVs were assigned to *Cylindrotheca* by the classifier, one or more occurring in each of the nine samples.

Processing with DADA2 does not remove all artifactual sequences and examination of the sequences of rare diatom ASVs revealed some that could not represent functional genes since they contained one or more stop codons. The most abundant of these was ASV0569, with a total of six reads and occurring in just one of the nine samples. However, rare ASVs were not necessarily artifactual. ASV0645, with just three reads, was an exact match to GenBank accession DQ813818 of *Pseudo-nitzschia delicatissima* (see also Section 3.4 below).

Motile and planktonic growth forms predominated in most of the samples in the molecular analyses and were primarily represented by *Nitzschia* and *Navicula* (motile) and *Thalassiosira* (planktonic) (Fig. 2b and Supplementary Table 2). The exceptions were the samples taken from *Crassostrea gigas* shells, where high profile forms were dominant, and *Cymodocea nodosa* (Fig. 2b), which had approximately equal proportions of high profile and motile forms. The high-profile group was mainly represented by *Achnanthes* and *Berkeleya* species. Conversely to LM, planktonic and tychoplanktonic forms were recorded in all the samples (Fig. 2a and b), while the low-profile group was much less represented; *Nanofrustulum* and *Amphora* genera were the most important representatives for the low-profile group.

The most striking feature of the molecular data was the abundance in most samples (except *Crassostrea*) of *Thalassiosira profunda*, a species for which only three specimens were identified by LM (Supplementary Tables 1 and 2). Because of the systematic bias introduced by this species, we recalculated the relative abundances of the growth forms excluding *T. profunda*. The resulting graphs (Fig. 2c) showed closer agreement with the morphological data.

3.3. Comparative analyses of samples from different substrata

Taken together, the two approaches identified a total of 102 different genera, of which 43 were identified in both inventories (43.4%), and each of both methods recorded exclusively 28 different genera. At species level, both approaches identified a total of 176 different species, of which 19 (10.9%) were identified in both inventories; 107 and 50 were exclusively detected in the morphological and molecular inventories respectively (Supplementary Table 3).

The Shannon diversity index calculated on taxa relative abundances differed between inventories. For almost all the samples, the index values were higher in the LM inventory (Table 2) and the averages obtained for the LM and DNA metabarcoding inventory were 3.29 and 2.31 respectively. Both approaches agreed that the highest diversity was in a sample from a shell of *P. nobilis* (LM = 3.74, DNA metabarcoding = 3.04; Table 2) but disagreed for the lowest diversity; in the LM inventory this was in the sample from *Cymodocea nodosa* (2.61) but in the DNA one it was in the sample from *Caulerpa prolifera* (1.59) (Table 2). A Mantel test indicated that DNA metabarcoding and LM distance matrices calculated on diatom species relative abundances were not significantly correlated (Mantel $r = 0.31$; p value $= 0.077$).

The NMDS and Sørensen similarity index based on DNA metabarcoding data showed a tendency for community composition to be

Table 2
Comparison of taxa richness and Shannon diversity index values obtained for the LM and DNA metabarcoding methods.

Sample	Microscopy		DNA metabarcoding	
	Taxa richness	Shannon index	Taxa richness	Shannon index
E5—*Crassostrea gigas*	69	3.50	37	2.15
E8—*C. gigas*	48	2.90	52	2.92
E9—*Pinna nobilis* biofilm	44	2.81	37	2.43
E10—*Pinna nobilis* sediment	75	3.73	47	2.23
E11—*Pinna nobilis* biofilm	71	3.74	62	3.04
E12—*Pinna nobilis* biofilm	67	3.46	55	3.04
E13—*Pinna nobilis* sediment	72	3.50	57	2.88
E14—*Cymodocea nodosa*	40	2.61	35	2.21
E15—*Caulerpa prolifera*	44	3.34	25	1.59

more similar among samples taken from the same host (Fig. 3a; Supplementary Table 4); this tendency was still evident after the *Thalassiosira profunda* ASVs were removed and the NMDS recalculated (Supplementary Fig. 1). However, these tendencies were not as obvious when NMDS and the Sørensen index were calculated using LM data (Fig. 3b; Supplementary Table 4). In particular, the two samples of *Crassostrea gigas* were widely separated from each other in the LM-based analyses but very close and separated from the rest in the DNA metabarcoding-based ones.

PERMANOVA confirmed the previous tendencies observed, with statistically significant differences in the community composition among different substrata for the DNA metabarcoding inventory (PERMANOVA using ASVs: $F_{4,4} = 2.7965$, $p = 0.012$; using species: $F_{4,4} = 3.3896$, $p = 0.01$; and using genera: $F_{4,4} = 3.5155$, $p = 0.007$) and for the LM inventory at species level (PERMANOVA: $F_{4,4} = 1.362$, $p = 0.032$) but not at genus level though differences were close to being statistically significant (PERMANOVA: $F_{4,4} = 1.6881$, $p = 0.056$).

According to the SIMPER analyses, the five genera that contributed most to the discrepancy between the morphological and molecular approaches were *Thalassiosira* (18.54%), *Navicula* (10.79%), *Amphora* (9.78%), *Cocconeis* (5.80%) and *Achnanthes* (5.61%). Below the genus level, the taxon that most influenced the dissimilarity was *T. profunda*, which was identified only by DNA metabarcoding. This species appeared in all samples analysed and was responsible for 14.37% of the discrepancy between the two inventories (Table 3). The second most important taxon was *Navicula* sp.4, contributing 4.88% of the dissimilarity. It was identified only by LM, and it appeared in most of the samples (Table 3).

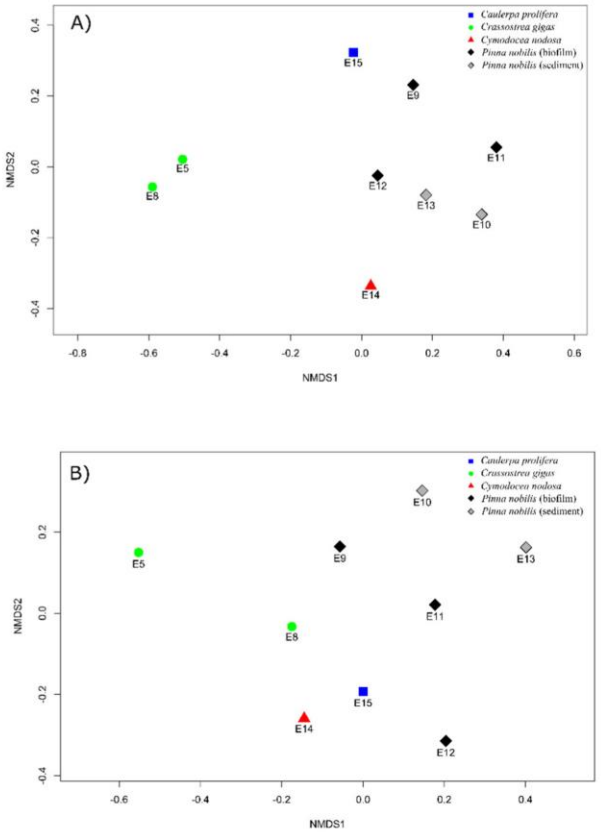


Fig. 3. Non-metric multidimensional scaling of the Bray-Curtis dissimilarity calculated on diatom ASVs relative abundance (A; stress = 0.09) and relative abundance of diatom taxa identified by LM (B; stress = 0.13).

Next was *Amphora helenensis*, which was identified by both methods, but it was much more abundant in the LM inventory. The opposite case was exemplified by *A. longipes*, a large-celled species with many chloroplasts that was much more abundant in the DNA metabarcoding inventory; it was the fourth species most influencing the dissimilarities between the two inventories (4.15%) (Table 3). A total of 83 species identified by LM lacked representative sequences in the reference library and they accounted for 16.32% of the total dissimilarities between inventories. Noteworthy among these were *C. scutellum* var. *posidoniae*, *C. scutellum*, *Navicula normaloides* and *N. normalis*, which together accounted for 6.10% of the total discrepancies (Table 3).

3.4. Diversity and phylogenetic analyses of non-diatom ASVs

Blastn and phylogenetic analyses allowed us to classify many of the non-diatom DNA reads to a class of algae and in some cases to a genus or species. In total, 41 non-diatom ASVs were analysed and allocated, a few of them with considerable hesitation, to an alga class or division. Ten of them were assigned to the Chlorophyta (and were easily recognized in the ASV alignment because all had an extra amino acid relative to the ochrophyte and red algal sequences), mostly with low similarity to any named taxon, except for *Umbraulva*, *Ulvella* and marine *Ulothrix* (the kleptoplasts of *Strombidium* sequenced for GenBank AY257112 are presumably from this genus: Supplementary Table 5); one sequence was apparently related closely to an uncultured *Picochlorum* (Supplementary Table 5). The three red algal ASVs were placed more definitively, as species (or relatives) of the genera *Grania* and *Acrochaetium*, which both grow as branching filaments, and the crustose *Pneophyllum*.

Most of the remaining non-diatom ASVs could be assigned with varying degrees of confidence to one of 10 classes of Ochrophyta (sensu Adl et al., 2019) (Fig. 4, Supplementary Table 5): Chrysophyceae (1 ASV, with low confidence), Synchromophyceae (2 ASVs), Pinguiophyceae (1 ASV), Eustigmatophyceae (1 ASV), Dictyochophyceae (2 ASVs), Pelagophyceae (7 ASVs), Raphidophyceae (2 ASVs), Xanthophyceae (1 ASV), Chrysomeridophyceae (2 ASVs) and Phaeophyceae (10 ASVs). The Phaeophyceae ASVs were mostly allied to species with simple or branched filaments, either in the Ectocarpales (*Hincksia*, *Myrionema*, *Streblonema*, *Elachista*, *Nemacystus*) or the Sphacelariales (*Sphacelaria*). Eight ochrophyte classes were unrepresented in the dataset: the predominantly freshwater Synurophyceae, the picoplanktonic Bolidophyceae, and the Olisthodiscophyceae, Aureanophyceae, Phaeothamniophyceae, Phaeosacciophyceae, Chrysoparadoxophyceae and Schizocladiophyceae.

The only non-diatom ASV that could almost certainly be discounted as an artifact was the very rare ASV0657, which had a low similarity to *Tetraselmis* (c. 80%: Supplementary Table 5) and was represented by just three reads; this sequence contained two stop codons. However, three even rarer non-diatom ASVs, each represented by two reads (ASV0677–0679, belonging to the Rhodophyta and Synchromophyceae), were clearly not artifactual, judging by blastn assignment or phylogenetic analysis (Fig. 4, Supplementary Table 5).

None of the non-diatom ASVs were abundant, the only one exceeding 1% of reads in any sample being an ectocarpalean brown alga (ASV0078, related to *Nemacystus decipiens*) in one of the *Pinna* biofilms (E12 – *Pinna nobilis* biofilm). The most widespread was ASV0183, an unclassified eustigmatophyte that was found in all five *Pinna* biofilm and sediment samples but nowhere else (Supplementary Table 5). Another possibly significant association was between the brown alga *Streblonema maculans* and *Crassostrea*. *Pinna* biofilms were a rich source of non-diatom ASVs, especially in sample E12 (*Pinna nobilis* biofilm).

Table 3
SIMPER analyses showing taxa contribution to the total dissimilarities between DNA metabarcoding and LM methods. Only the first thirty taxa with the greatest contribution to dissimilarities are shown. We also indicate the taxa for which there are or are not representative sequences in the reference library Diat.barcode v9.

Taxon	Relative abundance DNA metabarcoding	Relative abundance LM	Contribution to dissimilarities (%)	Cumulative contribution to dissimilarities (%)	Availability of a representative sequence in Diat.barcode v9
Thalassiosira profunda	27.36	0	14.37	14.37	Yes
Navicula sp.4	0	9.3	4.88	19.25	No
Amphora helenensis	1.74	9.08	4.27	23.52	Yes
Achnanthes longipes	7.64	0.45	4.15	27.67	Yes
Berkeleya fennica	7.17	2.02	3.51	31.18	Yes
Nanofrustulum shiloi	5.17	2.61	2.97	34.15	Yes
Navicula sp.	4.47	0	2.35	36.50	Yes
Amphora cf *helenensis*	0	4.17	2.19	38.69	Yes
Nitzschia spathulata	3.57	0	1.87	40.56	No
Cocconeis scutellum v. posidoniae	0	3.47	1.82	42.38	No
Navicula normaloides	0	2.89	1.52	43.90	Yes
Haslea howeana	2.79	0	1.47	45.37	No
Cyclotella choctawhatcheeana	0.11	2.77	1.43	46.80	Yes
Navicula normalis	0	2.69	1.41	48.21	No
Cocconeis scutellum	0	2.56	1.35	49.56	No
Cyclotella sp.	2.15	0	1.13	50.68	Yes
Seminavis robusta	1.57	0.92	1.11	51.79	Yes
Nitzschia sp.	2.02	0	1.06	52.86	Yes
Pleurosigma sp.	1.94	0	1.02	53.88	Yes
Halamphora sp.2	0	1.81	0.95	54.83	Yes
Navicula perminuta	1.75	0.53	0.92	55.75	No
Serratifera sp.3	0	1.71	0.90	56.65	No
Plagiogramma minus	0	1.71	0.90	57.55	No
Seminavis cf. *robusta*	1.69	0	0.89	58.43	Yes
Navicula subagnita	0	1.68	0.88	59.31	No
Pteroncola marina	0	1.43	0.75	60.06	Yes
Craspedostauros constricta	1.42	0	0.74	60.81	Yes
Psammodictyon sp.	1.39	0	0.73	61.54	No
Mastogloia crucicula	0	1.36	0.72	62.26	Yes
Halamphora sp.	1.36	0	0.71	62.97	No

4. Discussion

4.1. Diatom diversity in shallow coastal environments is very high and DNA metabarcoding is a promising tool for studying it

Our results demonstrate that the shallow coastal ecosystems studied here harbour a very rich diatom community. A total of 126 species were identified by morphology (LM); this is remarkably high when compared with previous studies on coastal environments based on a much larger sampling effort (e.g., 68–328 diatom taxa from 21 to 165 samples; Lobban et al., 2012; Facca and Sfriso, 2007; Kanjer et al., 2019; Virta et al., 2019). To these 126 species identified by LM, an extra 50 were added by metabarcoding. Furthermore, the large number of taxa identified only at generic level or above in both LM and DNA metabarcoding, may indicate that the total number of diatom species in the study area is very much higher. Comparisons with freshwater benthic communities are also instructive. The average diversities of our nine samples exceeded those of periphyton samples from Catalan rivers (for which we had many more samples, Supplementary Table 6), whether the approach taken was metabarcoding or microscopical analysis, emphasizing how rich the diatom communities of the marine benthos can be.

Hence, this first survey of some of the substrata in the Ebro Bays suggests the area is a hot-spot of diatom biodiversity and provides a first step towards understanding how this biodiversity originates and is maintained and the ecological roles that it performs. For instance, diatom biofilms in shallow coastal ecosystems are known to play a major role in sediment stabilization and in providing habitat and food for other organisms (references in Trobajo and Sullivan, 2010; see also Camps-Castellà et al., 2020 for a relevant example from Ebro Bays); moreover, a recent study of benthic diatoms in the Baltic Sea has shown that high diatom diversity supports high ecosystem productivity (Virta et al., 2019). Our study demonstrates that DNA metabarcoding based on the short 312-bp *rbcL* marker also constitutes an efficient method for surveying diatom biodiversity in coastal ecosystems. The effectiveness of the method was reflected in that 1) it recorded the same number of genera as the LM method did, 2) a high proportion of the genera (43.4%) identified by LM were also recorded by DNA metabarcoding and 3) a high number of genera (43) were exclusively identified by DNA metabarcoding. Nevertheless, the LM method showed a greater efficiency for identifying taxa at species level, which was mainly caused by the lack of representative sequences in the DNA reference library for many common species, which consequently could not be retrieved when the molecular approach was applied (further discussion in Section 4.3). Despite this limitation, similarity analyses calculated on the DNA metabarcoding inventory (at ASV, species and genus level) revealed a highly-structured molecular signal, suggesting therefore that our *rbcL*-based metabarcoding was able to discriminate different shallow coastal habitats, as in some other recent studies using other DNA markers (e.g. Bombin et al., 2021; Jeunen et al., 2018; Plante et al., 2021a). In addition, the habitat preferences hinted at by DNA metabarcoding could also indicate a degree of host-specificity among diatom taxa. However, our study was not designed to investigate these aspects in detail but to explore the feasibility of using *rbcL* metabarcoding to study the benthic diatom communities of shallow coastal habitats. Therefore, another study with a greater sampling effort and strategy (e.g., to have matching samples from the two Ebro bays for those hosts present in both; more replication etc) will be needed before any further conclusions can be drawn.

Regarding diatom composition, it was noticeable that a higher proportion of the taxa identified by LM corresponded to the low profile and motile groups, the genera *Amphora*, *Navicula*, *Cocconeis*, *Nitzschia* and *Halamphora* being particularly abundant. These have been recorded as important members of epiphyte communities in the Mediterranean (Car et al., 2019; Mabrouk et al., 2014) or as epizoic on the shells of *P. nobilis*, *C. gigas* and other molluscs (e.g., Andriana et al., 2021; Barille et al., 2017; D'Alelio et al., 2011; Totti et al., 2011). Conversely to LM, DNA metabarcoding better represented the planktonic group. Some planktonic taxa that were only recorded in our benthic samples by DNA metabarcoding have been previously reported as 'epizoic' or 'epiphytic'

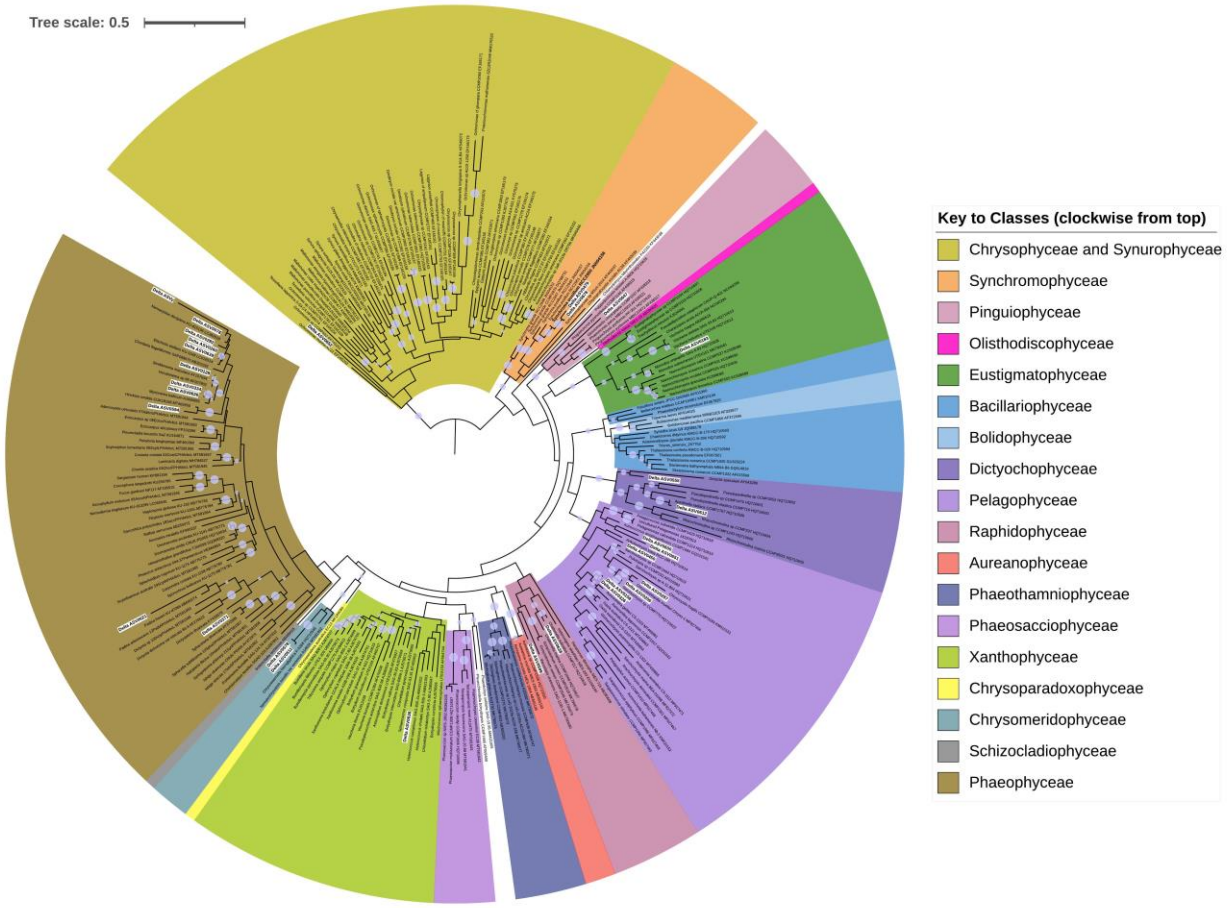


Fig. 4. Maximum likelihood phylogenetic tree based on non-diatom ASVs related to different heterekont classes. *RbcL* representative sequences included in the tree were extracted from Graf et al. (2020), Wetherbee et al. (2021) and GenBank database. The tree was built using raxmlGUI on an alignment partitioned by codon position and setting the GRT-Gamma model with 1000 replicates for the bootstrap analyses. The tree was drawn using iTOL. Bootstrap support values from 70 to 100 are represented. ASVs from *rbcL* metabarcoding are highlighted by white boxes.

in other studies. Examples are *Actinoptychus octonarius*, which has been reported elsewhere as occurring on *Pinna nobilis* (Politis 1949, cited by Round, 1971), and *Asterionellopsis*, found on the seagrass *Posidonia oceanica* (Mabrouk et al., 2014). However, although it is possible that these two are genuinely benthic or tychoplanktonic, it is also possible that the cells represent sedimented phytoplankton: the much lower limit of detection of the metabarcoding approach makes it much easier to detect occasional cells or colonies displaced from their normal habitat. Another planktonic species, *Thalassiosira profunda*, is considered in detail below.

4.2. Opportunities of DNA Metabarcoding

4.2.1. Detection and discrimination of tiny or delicate species

One advantage of the metabarcoding approach is that it is more likely to pick up small and/or delicate species. We recorded *Cylindrotheca* and Cymatosirales species in the metabarcoding dataset, but almost none in the morphological inventory. The *Cylindrotheca closterium* complex is commonly reported from coastal marine phytoplankton samples and it is probably mostly tychoplanktonic. However, these samples are often counted without prior cleaning, whereas benthic samples are not generally examined before oxidative cleaning because of the difficulty of observing sufficient frustule detail in intact biofilms or sediment samples. Hence many records of *Cyindrotheca* have probably been lost. One of the Cymatosirales that we recorded with metabarcoding, *Papiliocellulus simplex*, was first described from intertidal sand at two localities in Great Britain (Gardner and Crawford, 1992) and has subsequently been recorded only planktonically from the Liguro-Provencal basin of the Mediterranean (where it was 'extremely rare' at two stations: Percopo et al., 2011) and from several localities around Australia, mostly from metagenomic data [GBIF query 19 July 2021]. A final example of a small species easily missed or misidentified in light microscopy is *Gedaniella panicellus*, which was detected by DNA at frequencies of 0.1–1% in all but one sample but was not recorded at all in LM. This species was recently described by Li et al. (2018), who noted the difficulties of unambiguous identification by LM. Our ASV differed from the reference sequence (MF092953) by 1 bp and our record extends the known range to Europe from S. Africa and China, and from muddy rockpools to epizoic and epiphytic diatom communities.

4.2.2. Rare species

Another benefit of metabarcoding is the possibility of detecting species that are too rare to be found in routine LM cell counts. It is usually unrealistic to count more than a few 100s of valves per sample in LM, but it is common to obtain 1000s or 10,000s of reads with metabarcoding. However, rare ASVs need to be treated with some caution,

because amplification and sequencing can generate errors. Indeed, a few of our ASVs seem to be artifacts, despite the error modelling and correction incorporated in DADA2, since they contained stop codons. It is much more difficult to detect errors (e.g., in the third codon position) that do not affect the amino acid coded for. Errors can be minimized by imposing an arbitrary criterion – like a minimum number of reads – to try to avoid including artifactual sequences. However, our data illustrate (e.g., the rare ASVs identified as *Pseudo-nitzschia delicatissima*, *Acrochaetium*, *Pneophyllum* and *Synchroma*: see Sections 3.2, 3.4) what would be lost by imposing such a limit. The rarest ASVs can be genuine.

4.2.3. Primers designed for diatoms successfully amplify some non-diatom species

The primers we used were originally designed for use in freshwaters, for biomonitoring and biodiversity studies of diatoms (Vasselon et al., 2017). Our study is one of the first to apply the same primers in marine environments and reveals that the 'diatom-specific' primers do in fact amplify a wide variety of other marine microalgae, and even some green and red algae. The phylogenetic analyses revealed ASVs belonging to 10 different non-diatom classes in the Ochrophyta. Some of these, not surprisingly, were brown algae (various Ectocarpales and Sphacelariales), which probably formed part of the macroscopic structure of the attached communities (though perhaps surprisingly, some also occurred in the sediment samples), but others were microalgae, including some that are seldom recorded. Rather than a weakness of the HTS protocols, as suggested by Grant et al. (2021), we would argue that this 'contamination' of the diatom data is not only tolerable (since the proportion of non-diatom reads was low – c. 1% of the total; for comparison, 25% of the Ion Torrent 18S rDNA reads obtained by Plante et al., 2021b, were non-diatoms) but a valuable bonus, especially because many of the microalgae recorded are probably small and morphologically simple (judging by the nearest relatives that can be identified in GenBank) and will therefore be easily overlooked using microscopy or culturing. Especially interesting was the discovery of ASVs related to the amoeboid alga *Synchroma pusillum* (recently described by Schmidt et al., 2012) and the coccoid *Pinguiococcus pyrenoidosus* (which is difficult to identify in LM due to its small cell diameter of 3–8 μm: Andersen et al., 2002), and lineages of Pelagophyceae and Chrysomeridophyceae. The phylogenetic analyses also revealed one ASV closely related to the raphidophyte species *Chattonella subsalsa*. This species has been reported, among other species of *Chattonella*, to produce red tides and fish kills (Lewitus et al., 2008); the reads came from one of the sediment samples but could perhaps represent stray cells from the bay phytoplankton. Thus, these analyses illustrate the potential of DNA metabarcoding, even when based on primers designed for diatoms, for identifying at least some of the other microeukaryote taxa also present in the community, including ecologically or economically relevant taxa.

4.3. Discrepancies between the LM and metabarcoding results

4.3.1. The case of Thalassiosira profunda

The greatest dissimilarities between the results obtained by the two methods, morphological and metabarcoding, were attributable to one particularly small, delicate species, the centric *Thalassiosira profunda*. This was by far the most abundant species recorded by DNA metabarcoding but only three specimens were identified by LM (2 and 1 respectively in the *C. prolifera* and *C. nodosa* samples). These were found after an additional exhaustive examination of the samples was performed, beyond the normal 300–400 count and prompted by the metabarcoding data, to be sure that this species had not been overlooked in LM. *Thalassiosira profunda* is an extremely small species (valve diameter 1.25–5.5 μm), generally regarded as planktonic, which is very widely distributed (Percopo et al., 2011; Li et al., 2013; Park et al., 2016; Guiry, 2021). The almost complete absence of this species from the morphological counts, which can alternatively be described as gross overrepresentation in the metabarcoding dataset, requires special explanation, because such overrepresentation is generally associated with large-celled species, such as *Ulnaria ulna*, *Pinnularia viridiformis* or *Navicula lanceolata* (Vasselon et al., 2018; Kelly et al., 2020), because they have a larger number of copies of *rbcL* per cell.

Several hypotheses might explain why *T. profunda* was abundant in the DNA reads but extremely rare from the LM inventory. None of them can be discounted entirely; all of them have consequences for planning and interpreting morphological and metabarcoding studies of marine benthic diatoms.

1. *T. profunda* could be detected almost exclusively by metabarcoding because diatoms with tiny valves are easily overlooked and difficult to identify in LM (e.g., see Belcher and Swale, 1986). In the present case, such an explanation can be discounted, given the abundance implied by the molecular data (especially taking into account the likely low *rbcL* copy number per cell) and given that all slides were examined in detail using a × 100 objective. Furthermore, our re-examination of the slide preparations after analysing the metabarcoding data confirmed the almost complete absence of *T. profunda*, while no other *Thalassiosira* species were recorded as abundant in LM. However, the greater number of very small-celled and delicate diatoms (e.g. of Thalassiosirales or Cymatosirales) in many marine habitats, relative to freshwaters, means that there is greater potential for discrepancies between molecular and morphological datasets in marine studies.
2. Valves of delicate species like *T. profunda* can be destroyed during preparation for LM, as has been reported for the weakly-silicified cells of the freshwater *Fistulifera saprophila* (Kelly et al., 2020; Pérez-Burillo et al., 2020; Zgrundo et al., 2013). Small size also predisposes them to be lost, since centrifugation and sedimentation during washing steps will be less effective (e.g., Andrews, 1972). The solution is clearly to examine material before it is cleaned or retain aliquots for examination later. Unfortunately, we did not do this, but the detection of a few intact valves of *T. profunda* in the *C. prolifera* and *C. nodosa* samples, following an exhaustive search for the species, undermines destruction as a reason for 'overrepresentation' in the molecular dataset.
3. The molecular signal captured for *T. profunda* may not be contemporary with the morphologically characterized benthic communities but come from an earlier bloom. Some planktonic species form resting stages following a bloom (McQuoid and Hobson, 1996; Inoue and Taniguchi, 1999; Sugie and Kuma, 2008), leading to the deposition of a large numbers of resting spores in the sediment. These might be detectable using DNA but more difficult by LM due to morphological differences from the vegetative cells (cf. Kuwata and Takahashi, 1999). However, this strategy is not known to occur in *T. profunda*. In any case, diatom resting spores and resting cells are usually more robust than vegetative cells (Krawczyk et al., 2012) and should have been found in our material if present. Alternatively, an earlier bloom of *T. profunda* could perhaps have left a molecular trace even though the frustules had redissolved in situ. A moderate abundance of DNA reads of *Thalassiosira* and other planktonic species, including *T. profunda*, was recently reported in saltmarsh sediments in S Carolina, USA, by Plante et al. (2021a), who suggested this could reflect deposition of faecal pellets or recent phytoplankton blooms. However, their study did not include accompanying cell counts from LM.
4. Finally, it is possible that intact *T. profunda* cells were present in the samples but lacked frustules, so that they were undetectable in material prepared for microscopy. As far as we know there is no confirmed report of *free-living* diatoms lacking a silica cell wall, apart from some morphotypes of *Phaeodactylum* (Round et al., 1990), but this does not mean that none exist. But wall-less diatoms certainly do occur as endosymbionts, for example in some foraminifera (Lee, 2011) and dinoflagellates (Yamada et al., 2020), while other foraminifera and dinoflagellates ingest diatoms and jettison the frustules,

retaining their chloroplasts (as 'kleptoplasts') for days or months afterwards (e.g., Pillet et al., 2011; Yamada et al., 2019), and hence also, perhaps, their *rbcL*. In freshwaters, some Thalassiosirales are known to be endosymbionts of dinoflagellates (e.g., Takano et al., 2007; You et al., 2015), while in marine environments chloroplasts of Thalassiosirales are retained by some foraminifera, e.g., *Elphidium* (Pillet et al., 2011) and *Nonionellina* species (Jauffrais et al., 2019), at least one of which (*Elphidium*) occurs in the Ebro Bays (Benito et al., 2016).

The possibility that *T. profunda* could have been living endosymbiotically or as kleptoplasts in the communities we sampled receives further specific support from the study of Schmidt et al. (2018), who looked at the endosymbionts of the benthic foraminifera *Pararotalia calcarioformata* growing in the East Mediterranean. Among the endosymbiotic diatoms they extracted and cultured from *P. calcarioformata* was one species identified and illustrated as *Minidiscus* sp. (Schmidt et al., 2018), which we would identify instead, from the specimen illustrated (op. cit., fig 4.12), as *T. profunda*. In any case, it is clear that several Thalassiosirales do occur without frustules in marine environments, providing a possible reconciliation of our molecular and morphological data. The potential of kleptoplasts to confuse metabarcoding results extends beyond the photic zone, since functioning diatom kleptoplasts (again from Thalassiosirales, though apparently of *Skeletonema* not *Thalassiosira*) have been recorded, with intact *rbcL*, from depths of more than 500 m in the foraminifer *Nonionella stella* (Gomaa et al., 2021).

4.3.2. The effects of taxonomic resolution, reference library and gene copy number

Another important reason for the discrepancies observed between methods was the impossibility of identifying some taxa at species level. Marine littoral diatoms have been much less studied than their freshwater counterparts in rivers and lakes, so that several species in our samples – some of them abundant – remained unidentified at species level in LM, despite the great taxonomic effort and resources applied (including thorough LM identifications supported by SEM and TEM). The number of taxa identified as sp. or confer (cf) or affinis (aff) illustrates the incompleteness of the taxonomy underlying the morphological approach. This was particularly true for *Amphora* and also for *Navicula*, since a total of eight *Navicula* species could not be identified to known species. One of these, *Navicula* sp. 4 was very abundant and contributed greatly to the discrepancies between the LM and metabarcoding outputs. The prevalence of *Navicula* spp. without a species assignation in epibiotic communities has been shown also in other recent studies (e.g., Andriana et al., 2021; Car et al., 2019; Kanjer et al., 2019; Medlin and Juggins, 2018).

Concerning the metabarcoding inventories, the impossibility of reaching species-level identifications of the ASVs was often due to the incompleteness of the reference library since many species identified by LM lack representative DNA sequences, again mainly due to understudied environments. Diat.barcode (Rimet et al., 2019) aims to list and check all available *rbcL* sequences for diatoms, whether marine, coastal, or freshwater, but it depends to a considerable extent on what sequences have been deposited in NCBI GenBank, which reflects historical trends in systematic and other research. Overall, the data available show a strong bias towards freshwater diatoms, which account for around 60% of the *rbcL* entries in Diat.barcode v. 9 (>4500 sequences), despite the greater diversity of marine diatoms. Perhaps not surprisingly, therefore, many of our ASVs were not assigned to a species, or even a genus, by the Bayesian classifier. The contrast with freshwater biomonitoring analyses is illustrated in Supplementary Table 7 which shows (for the marine samples and two campaigns of metabarcoding in Catalan rivers) the proportions of the ASVs that find exact matches with reference sequences or matches at 95–99% similarity levels: the marine analysis lags well behind.

The incompleteness of the reference database explained c. 16% of dissimilarities between methods and some of the species missed are known for being important components in the epizoic and epiphytic diatom communities, so they could be considered priorities for future barcoding. Among them are several *Navicula* species, including *N. normaloides*, *N. normalis* and *N. subagnita*; these taxa were identified in all or most of the samples, indicating their importance in the study area. *Navicula normaloides* and *N. subagnita* have also been recorded as epiphytic on leaves from *Posidonia oceanica* and *Caulerpa taxifolia* in the Adriatic Sea (Kanjer et al., 2019; Car et al., 2019). It is important to emphasize, however, that in many cases the lack of representative sequences only partially prevents interpretation of metabarcoding data, though it reduces the resolution achieved. For example, despite the lack of representative sequences for the particular *Navicula* species known (from LM) to occur in our samples, the coverage of the genus in the reference dataset was sufficient for the classifier to assign many ASVs at genus level. These assignments could then be checked by individual blastn searches and can be examined further in future by phylogenetic approaches, as with the non-diatom ASVs.

Other important species underestimated by DNA metabarcoding due to lack of reference sequences were *Cocconeis* species, notably *C. scutellum* and *C. scutellum* var. *posidoniae*. Both are cosmopolitan taxa and important components of the attached diatom communities (e.g., De Stefano et al., 2008; Polifrone et al., 2020; Ryabushko and Ryabushko, 2000; Witak et al., 2020). Overall, the genus *Cocconeis* was very poorly represented by DNA metabarcoding despite the very high diversity of *Cocconeis* species revealed by LM. Furthermore, and contrary to what we found with *Navicula* species, only a small proportion of reads and ASVs unclassified at the species level by the Bayesian classifier could be convincingly related to *Cocconeis* even at genus level. This can be explained by the fact that the reference database contains few *Cocconeis* (and almost all of them are freshwater species whose relationship to the marine species remains unknown), making it impossible for the classifier to assign ASVs to *Cocconeis* or related genera at any level. A few ASVs were tentatively identified as possible *Cocconeis* or Cocconeidaceae species on the basis of the spread of hits from blastn interrogation of GenBank, but overall it seems that the reference library is currently the main limitation to study *Cocconeis* diversity by DNA metabarcoding. Due to the importance of these diatoms in marine attached communities, further efforts should be made to increase their representation in the DNA reference library. A further and more worrying possibility is that some *Cocconeis* taxa may carry mutations in critical parts of one or both primer regions, but this too cannot be known without long reference sequences of the marine species. In contrast to *Cocconeis*, genera like *Pseudo-nitzschia*, *Haslea* and *Achnanthes* are well represented in the DNA reference database.

Finally, some discrepancies between methods can probably be attributed to variation among species in the *rbcL* copy number per cell, as noted previously by Vasselon et al. (2018), Kelly et al. (2020) and Pérez-Burillo et al. (2020). This variation depends on the number of gene copies per chloroplast and the number of chloroplasts per cell. A correlation between the *rbcL* copy number and cell biovolume has been reported, leading to much higher relative abundances for high biovolume species, e.g., *Ulnaria ulna*, large *Pinnularia* or *Pleurosira laevis*, in metabarcoding outputs (Vasselon et al., 2018). This very likely explains the higher abundances obtained by the DNA method for *Achnanthes longipes* and *Pleurosigma*. These taxa are characterized by high biovolume and either high numbers of chloroplasts per cell (*A. longipes*) or highly complex, large chloroplasts (*Pleurosigma*).

5. Concluding remarks

As mentioned earlier, diatoms can contribute well over 50% of primary benthic production in marine habitats where the only visually obvious photosynthetic organisms are seagrasses (Cox et al., 2020), while on apparently bare sediments lacking macrophytes, diatoms

generally dominate, except in summer when cyanobacteria are often important (e.g., Scholz and Liebezeit, 2012b). Furthermore, individual diatom species, including species that coexist, can exhibit different responses to macronutrients (e.g., Underwood and Provot, 2000), different vertical migration patterns (Underwood et al., 2005), and different seasonality (e.g., Scholz and Liebezeit, 2012a). Hence, to understand marine benthic communities, it is important to identify and quantify species, and hence to have resources that facilitate consistent accurate identification.

The main aim of this paper was to use a small dataset to examine the advantages and disadvantages of metabarcoding and morphological approaches to study the benthic diatom communities of shallow coastal environments. Our results show that both approaches are more difficult to implement than in freshwater environments and in both cases the cause is essentially the same: marine microphytobenthic communities have been greatly understudied, despite their ecological and economic importance. As a result, the traditional morphology-based taxonomy has yet to advance to the level achieved for freshwaters, while the lack of reference sequences limits the resolution achievable with metabarcoding, though this did not prevent the molecular method from separating the samples according to the type of substratum. There are also special features of the marine benthos – such as the presence of a wider range of related microalgal groups – that offer extra opportunities for studying non-diatom diversity but also mean that the reference database needs to be more inclusive than in freshwaters for efficient identification of ASVs. Clearly, then, both approaches, morphological and metabarcoding, are in some senses incomplete for marine benthic diatom communities, but together they offer a strong foundation for ecological and biogeographical studies. We suggest that the way forward, for the moment, is to develop metabarcoding and morphological approaches in parallel and exploit their particular strengths and complementarity: for example, the far greater resolution and sensitivity of metabarcoding (and the albeit limited capacity to detect non-diatom components), combined with the insights into life-form, cell surface area: volume relationships and functional group membership that are inherent in the morphological approach and can never be fully realized with metabarcoding, even when the reference database is complete and allows all ASVs to be allocated to known species.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.marpolbul.2021.113183.

CRediT authorship contribution statement

Javier Pérez-Burillo: Formal analysis, Investigation, Data curation, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Greta Valoti:** Formal analysis, Investigation, Data curation, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Andrzej Witkowski:** Formal analysis, Writing – review & editing. **Patricia Prado:** Writing – review & editing. **David G. Mann:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Rosa Trobajo:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

We acknowledge the Erasmus+ programme of the European Union which supported the traineeship of Greta Valoti (Università Politecnica delle Marche, Italy) in IRTA. J. Pérez-Burillo acknowledges IRTA and the Universitat Rovira i Virgili for his Martí Franqués PhD grant (2018PMF-PIPF-22). The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division. This article was facilitated by COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program. We also acknowledge support from the CERCA Programme/Generalitat de Catalunya. P Prado was contracted under the INIA-CCAA cooperative research program for postdoctoral incorporation from the Spanish National Institute for Agricultural and Food Research and Technology (INIA). The authors would like to thank the Biodiversity Foundation from the Spanish Ministry for Ecological Transition for providing additional support for fieldwork sampling in the context of the Recupera Pinna project. Prof. Rafał J. Wróbel (West Pomeranian University of Technology) is acknowledged for his help with SEM examination of the diatomaceous samples. G. Valoti acknowledges Prof. Cecilia Totti (Università Politecnica delle Marche) for her support with respect to GV's IRTA visit. The authors would like to thank Vanessa Castan, IRTA technician, who collected the individuals of *Crassotrea gigas*. The authors state no conflicts of interest.

References

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D.J.G., Lara, E., Le Gall, L., Lynn, D.H., Mann, D.G., Massana, R., Mitchell, E.A.D., Morrow, C., Park, J.S., Pawlowski, J.W., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Torruella, G., Youssef, N., Zlatogursky, V., Zhang, Q., 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. J. Euk. Microbiol. 66, 4–119. https://doi.org/10.1111/jeu.12691.

An, S.M., Choi, D.H., Noh, J.H., 2020. High-throughput sequencing analysis reveals dynamic seasonal succession of diatom assemblages in a temperate tidal flat. Estuar. Coast. Shelf Sci. 237, 106686. https://doi.org/10.1016/j.ecss.2020.106686.

Andersen, R.A., Potter, D., Bailey, J.C., 2002. *Pinguiococcus pyrenoidosus* gen. et sp. nov. (Pinguiophyceae), a new marine coccoid alga. Phycol. Res. 50, 57–65. https://doi.org/10.1111/j.1440-1835.2002.tb00136.x.

Andrews, G.W., 1972. Some fallacies of quantitative diatom paleontology. Nova Hedwigia, Beih. 39, 285–295.

Andriana, R., Engel, F.G., Gusmao, J.B., Eriksson, B.K., 2021. Intertidal mussel reefs change the composition and size distribution of diatoms in the biofilm. Mar. Biol. 168, 24. https://doi.org/10.1007/s00227-020-03819-2.

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

Barille, L., Le Bris, A., Méléder, V., Launeau, P., Robin, M., Louvrou, I., Ribeiro, L., 2017. Photosynthetic epibionts and endobionts of Pacific oyster shells from oyster reefs in rocky versus mudflat shores. PLoS ONE 12, e0185187. https://doi.org/10.1371/journal.pone.0185187.

Belcher, J.H., Swale, E.M.F., 1986. Notes on some small *Thalassiosira* species (Bacillariophyceae) from the plankton of the lower Thames and other British estuaries (identified by transmission electron microscopy). Br. Phycol. J. 21, 139–145. https://doi.org/10.1080/00071618600650161.

Benito, X., Trobajo, R., Ibáñez, C., 2015. Benthic diatoms in a Mediterranean delta: ecological indicators and a conductivity transfer function for paleoenvironmental studies. J. Paleolimnol. 54, 171–188. https://doi.org/10.1007/s10933-015-9845-3.

Benito, X., Trobajo, R., Cearreta, A., Ibáñez, C., 2016. Benthic foraminifera as indicators of habitat in a Mediterranean delta: implications for ecological and palaeoenvironmental studies. Estuar. Coast. Shelf Sci. 180, 97–113. https://doi.org/10.1016/j.ecss.2016.06.001.

Bombin, S., Wysor, B., Lopez-Bautista, J.M., 2021. Assessment of littoral algal diversity from the northern Gulf of Mexico using environmental DNA metabarcoding. J. Phycol. 57, 269–278. https://doi.org/10.1111/jpy.13087.

Cahoon, L.B., 1999. The role of benthic microalgae in neritic ecosystems. Oceanogr. Mar. Biol. Ann. Rev. 37, 47–86. https://doi.org/10.1201/9781482298550-4.

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., 2009. BLAST plus: architecture and applications. BMC Bioinform. 10, 421. https://doi.org/10.1186/1471-2105-10-421.

Camps-Castellà, J., Romero, J., Prado, P., 2020. Trophic plasticity in the sea urchin *Paracentrotus lividus*, as a function of resource availability and habitat features. Mar. Ecol. Prog. Ser. 637, 71–85. https://doi.org/10.3354/meps13235.

Car, A., Witkowski, A., Dobosz, S., Jasprica, N., Ljubimir, S., Zgłobicka, I., 2019. Epiphytic diatom assemblages on invasive *Caulerpa taxifolia* and autochthonous

Halimeda tuna and *Padina* sp. seaweeds in the Adriatic Sea – summer/autumn aspect. Oceanol Hidrobiol Stud. 48, 209–226. https://doi.org/10.2478/ohs-2019-0019.
Carballeira, R., Trobajo, R., Leira, M., Benito, X., Sato, S., Mann, D.G., 2017. A combined morphological and molecular approach to *Nitzschia varelae* sp. nov., with discussion of symmetry in Bacillariaceae. Eur. J. Phycol. 52, 342–359. https://doi.org/10.1080/09670262.2017.1309575.
Chonova, T., Keck, F., Bouchez, A., Rimet, F., 2020. A ready-to-use database for DADA2: Diat.barcode_rbcL_263bp_DADA2 based on Diat.barcode v9. Portail Data INRAE, V2, 10.15454/QBLSXP.
Cloern, J.E., Foster, S.Q., Kleckner, A.E., 2013. Review: phytoplankton primary production in the world's estuarine-coastal ecosystems. Biogeosci. Discuss. 10, 17725–17783. https://doi.org/10.5194/bg-11-2477-2014.
Cox, T.E., Cebrian, J., Tabor, M., West, L., Krause, J.W., 2020. Do diatoms dominate benthic production in shallow systems? A case study from a mixed seagrass bed. Limnol Oceanogr. 5, 425–434. https://doi.org/10.1002/lol2.10167.
Costanza, R., de Groot, R., Sutton, P., van der Ploeg, S., Anderson, S., Kubiszewski, I., Farber, S., Turner, R., 2014. Changes in the global value of ecosystem services. Glob. Environ. Chang. 26, 152–158. https://doi.org/10.1016/j.gloenvcha.2014.04.002.
D'Alelio, D., Cante, M.T., Russo, G.F., Totti, C., De Stefano, M., 2011. Epizoic diatoms on gastropod shells: when substrate complexity selects for microcommunity complexity. In: Dubinsky, Z., Seckbach, J. (Eds.), All Flesh Is Grass. Plant-animal interrelationships, Springer, Netherlands, pp. 349–364.
Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., Vere, N., Pfrender, M.E., Bernatchez, L., 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol. Ecol. 26, 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 J. https://doi.org/10.1038/s41396-021-00895-0.
De Stefano, M., Romero, O.E., Totti, C., 2008. A comparative study of *Cocconeis scutellum* Ehrenberg and its varieties (Bacillariophyta). Bot. Mar. 51, 506–536. https://doi.org/10.1515/BOT.2008.058.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.
Edler, D., Klein, J., Antonelli, A., Silvestro, D., 2021. raxmlGUI 2.0: a graphical interface and toolkit for phylogenetic analyses using RAxML. Methods Ecol. Evol. 12, 373–377. https://doi.org/10.1111/2041-210X.13512.
Facca, C.A., Sfriso, A., 2007. Epipelic diatom spatial and temporal distribution and relationship with the main environmental parameters in coastal waters. Estuar. Coast. Shelf Sci. 75, 35–49. https://doi.org/10.1016/j.ecss.2007.03.033.
Gomaa, F., Utter, D.R., Powers, C., Beaudoin, D.J., Edgcomb, V.P., Filipsson, H.L., Hansel, C.M., Wankel, S.D., Zhang, Y., Bernhard, J.M., 2021. Multiple integrated metabolic strategies allow foraminiferan protists to thrive in anoxic marine sediments. Sci. Adv. 7, eabf1586 https://doi.org/10.1126/sciadv.abf1586.
Gardner, C., Crawford, R.M., 1992. A description of the diatom *Papiliocellulus simplex* sp. nov. (Cymatosiraceae, Bacillariophyta) using light and electron microscopy. Phycologia 31, 246–252. https://doi.org/10.2216/i0031-8884-31-3-4-246.1.
Graf, L., Yang, E.C., Han, K.Y., Küpper, F.C., Benes, K.M., Oyadomari, J.K., Herbert, R.J. H., Verbruggen, H., Wetherbee, R., Andersen, R.A., Yoon, H.S., 2020. Multigene phylogeny, morphological observation and re-examination of the literature lead to the description of the Phaeosacciophyceae classis nova and four new species of the Heterokontophyta SI clade. Protist 171, 125781. https://doi.org/10.1016/j.protis.2020.125781.
Grant, D.M., Brodnicke, O.B., Evankow, A.M., Ferreira, A.O., Fontes, J.T., Hansen, A.K., Jensen, M.R., Kalaycı, T.E., Leeper, A., Patil, S.K., Prati, S., Reunamo, A., Roberts, A. J., Shigdel, R., Tyukosova, V., Bendiksby, M., Blaalid, R., Costa, F.O., Hollingsworth, P.M., Stur, E., Ekrem, T., 2021. The future of DNA barcoding: reflections from early career researchers. Diversity 2021 (13), 313. https://doi.org/10.3390/d13070313.
Guiry, G.M., 2021. *Thalassiosira profunda*. In: Guiry, M.D., Guiry, G.M. (Eds.), AlgaeBase. World-wide electronic publication, National University of Ireland, Galway searched on 16 July 2021. http://www.algaebase.org.
Inoue, T., Taniguchi, A., 1999. Seasonal distribution of vegetative cells and resting spores of the arcto-boreal diatom Thalassiosira nordenskioeldii cleve in Onagawa Bay, northeastern Japan. In: Mayama, I., Koizumi, I. (Eds.), Proceedings of the 14th International Diatom Symposium. Tokyo, pp. 263–276.
Jauffrais, T., LeKieffre, C., Schweizer, M., Geslin, E., Metzger, E., Bernhard, J.M., Jesus, B., Filipsson, H.L., Mare, O., Meiborn, A., 2019. Kleptoplastidic benthic formanifera from aphotic habitats: insights into assimilation of inorganic C, N and S studied with sub-cellular resolution. Env. Microbiol. 21, 125–141. https://doi.org/10.1111/1462-2920.14433.
Jeunen, G.-J., Knapp, M., Spencer, H.G., Lamare, M.D., Taylor, H.R., Stat, M., Bunce, M., Gemmell, N.J., 2018. Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. Mol. Ecol. Resour. 19, 426–438. https://doi.org/10.1111/1755-0998.12982.
Kanjer, L., Mucko, M., Car, A., Bosak, S., 2019. Epiphytic diatoms on *Posidonia oceanica* (L.) Delile leaves from eastern Adriatic Sea. Nat. Croat. 28, 1–20. https://doi.org/10.20302/NC.2019.28.1.
Kelly, M.G., Juggins, S., Mann, D.G., Sato, S., Glover, R., Boonham, N., Sapp, M., Lewis, E., Hany, U., Kille, P., Jones, T., Walsh, K., 2020. Development of a novel metric for evaluating diatom assemblages in rivers using DNA metabarcoding. Ecol. Indic. 118, 106725 https://doi.org/10.1016/j.ecolind.2020.106725.
Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.M., Humbert, J.F., Bouchez, A., 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshw. Sci. 33, 349–363. https://doi.org/10.1086/675079.
Krawczyk, D.W., Witkowski, A., Wroniecki, M., Waniek, J., Kurzydłowski, K.J., Płociński, T., 2012. Reinterpretation of two diatom species from the West Greenland margin — *Thalassiosira kushirensis* and *Thalassiosira antarctica* var. *borealis* – hydrological consequences. Mar. Micropaleontol. 88–89, 1–14. https://doi.org/10.1016/j.marmicro.2012.02.004.
Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549. https://doi.org/10.1093/molbev/msy096.
Kuwata, A., Takahashi, M., 1999. Survival and recovery of resting spores and resting cells of the marine planktonic diatom *Chaetoceros pseudocurvisetus* under fluctuating nitrate condition. Mar. Biol. 134, 471–478. https://doi.org/10.1007/s002270050563.
Lee, J.J., 2011. Diatoms as endosymbionts. In: Seckbach, J., Kociolek, P. (Eds.), The Diatom World. Cellular Origin, Life in Extreme Habitats and Astrobiology. Springer, Dordrecht, pp. 439–464, 0.1007/978-94-007-1327-7_20.
Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296. https://doi.org/10.1093/nar/gkab301.
Lewitus, A.J., Brock, L.M., Burke, M.K., DeMattio, K.A., Wilde, S.B., 2008. Lagoonal stormwater detention ponds as promoters of harmful algal blooms and eutrophication along the South Carolina coast. Harmful Algae. 8, 60–65. https://doi.org/10.1016/j.hal.2008.08.012.
Li, Ch.-L., Witkowski, A., Ashworth, M.P., Dąbek, P., Sato, S., Zgłobicka, I., Witak, M., Khim, J.S., Kwon, C.J., 2018. The morphology and molecular phylogenetics of some marine diatom taxa within the Fragilariaceae, including twenty undescribed species and their relationship to *Nanofrustulum*, *Opephora* and *Pseudostaurosira*. Phytotaxa. 355, 1–104. https://doi.org/10.11646/phytotaxa.355.1.1.
Li, Y., Zhao, Q., Lü, S., 2013. The genus *Thalassiosira* off the Guangdong coast, South China Sea. Bot. Mar. 56, 83–110, 0.1515/bot-2011-0045.
Llebot, C., Solé, J., Delgado, M., Fernández-Tejedor, M., Campa, J., Estrada, M., 2011. Hydrographical forcing and phytoplankton variability in two semi-enclosed estuarine bays. J. Mar. Syst. 86, 69–86. https://doi.org/10.1016/j.jmarsys.2011.01.004.
Lobban, C.S., Schefter, M., Jordan, R.W., Arai, Y., Sasaki, A., Theriot, E.C., Ashworth, M., Ruck, E., Chiara, P., 2012. Coral-reef diatoms (Bacillariophyta) from Guam: new records and preliminary checklist, with emphasis on epiphytic species from farmer-fish territories. Micronesica 43, 237–479.
Mabrouk, L., Ben Brahim, M., Hamza, A., Mahfoudhi, M., Bradai, M.N., 2014. A comparison of abundance and diversity of epiphytic microalgal assemblages on the leaves of the seagrasses *Posidonia oceanica* (L.) and *Cymodocea nodosa* (Ucria) Asch in Eastern Tunisia. J. Mar. Biol. 2014, 1–10. https://doi.org/10.1155/2014/275305.
MacIntyre, H.L., Geider, R.J., Miller, D.C., 1996. Microphytobenthos: the ecological role of the secret garden of unvegetated, shallow-water marine habitats. I. Distribution, abundance and primary production. Estuaries 12, 186–201. https://doi.org/10.2307/1352224.
Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, J., Ludicone, D., De Vargas, C., Bittner, L., Zingone, A., Bowler, C., 2016. Insights into global diatom distribution and diversity in the world's ocean. Proc. Natl. Acad. Sci. 113, 1516–1525. https://doi.org/10.1073/pnas.1509523113.
Mann, D.G., Crawford, R.M., Round, F.E., 2016. Bacillariophyta. In: Archibald, J.M., Simpson, A.G.B., Slamovits, C.H., Margulis, L., Melkonian, M., Chapman, D.J., Corliss, J.O. (Eds.), Handbook of the Protists. Springer, Cham, New York, pp. 1–62. https://doi.org/10.1007/978-3-319-32669-6_29-1.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, 10–12. https://doi.org/10.14806/ej.17.1.200.
McMurdie, P.J., Holmes, S., 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8, e61217. https://doi.org/10.1371/journal.pone.0061217.
McQuoid, M.R., Hobson, L.A., 1996. Diatom resting stages. J. Phycol. 32, 889–902. https://doi.org/10.1111/j.0022-3646.1996.00889.x.
Medlin, L.K., Juggins, S., 2018. Multivariate analyses document host specificity, differences in the diatom metaphyton vs. epiphyton, and seasonality that structure the epiphytic diatom community. Estuar. Coast. Shelf Sci. 213, 314–330. https://doi.org/10.1016/j.ecss.2018.06.011.
Mortágua, A., Vasselon, V., Oliveira, R., Elias, C., Chardon, C., Bouchez, A., Rimet, F., João Feio, M., Almeida, S.F., 2019. Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers using diatoms. Ecol. Indic. 106, 105470 https://doi.org/10.1016/j.ecolind.2019.105470.
Nunes, M., Lemley, D.A., Matcher, G.F., Adams, J.B., 2021. The influence of estuary eutrophication on the benthic diatom community: a molecular approach. Afr. J. Mar. Sci. 43, 171–186. https://doi.org/10.2989/1814232X.2021.1897039.
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.
Park, J.S., Jung, S.W., Lee, J.H., Yun, S.M., Lee, J.H., 2016. Species diversity of the genus *Thalassiosira* (Thalassiosirales, Bacillariophyta) in South Korea and its biogeographical distribution in the world. Phycologia 55, 403–423. https://doi.org/10.2216/15-66.1.
Passy, S.I., 2007. Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters. Aquat. Bot. 86, 171–178. https://doi.org/10.1016/j.aquabot.2006.09.018.
Percopo, I., Siano, R., Cerino, F., Sarno, D., Zingone, A., 2011. Phytoplankton diversity during the spring bloom in the northwestern Mediterranean Sea. Bot. Mar. 54, 243–267. https://doi.org/10.1515/bot.2011.033.

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

Pérez-Burillo, J., Trobajo, R., Leira, M., Keck, F., Rimet, F., Sigró, J., Mann, D.G., 2021. DNA metabarcoding reveals differences in distribution patterns and ecological preferences among genetic variants within some key freshwater diatom species. Sci. Total Environ. 728, 149029 https://doi.org/10.1016/j.scitotenv.2021.149029.

Pillet, L., de Vargas, C., Pawlowski, J., 2011. Molecular identification of sequestered diatom chloroplasts and kleptoplastidy in foraminifera. Protist 162, 394–404. https://doi.org/10.1016/j.protis.2010.10.001.

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. Sci. Rep. 8, 18059. https://doi.org/10.1038/s41598-018-36345-9.

Plante, C.J., Hill-Spanik, K., Cook, M., Graham, C., 2021. Environmental and spatial influences on biogeography and community structure of saltmarsh benthic diatoms. Estuar. Coasts. 44, 147–161. https://doi.org/10.1007/s12237-020-00779-0.

Plante, C.J., Hill-Spanik, K., Lowry, J., 2021. Controls on diatom biogeography on South Carolina (USA) barrier island beaches. Mar. Ecol. Progr. Ser. 661, 17–33. https://doi.org/10.3354/meps13598.

Polifrone, M., Viera-Rodríguez, M.A., Pennesi, C., Conte, M.T., Del Pino, A.S., Stroobant, M., De Stefano, M., 2020. Epiphytic diatoms on Gelidiales (Rhodophyta) from Gran Canaria (Spain). Eur. J. Phycol. 55, 404–411. https://doi.org/10.1080/09670262.2020.1737967.

Prado, P., Caiola, N., Ibañez, C., 2014. Habitat use by a large population of *Pinna nobilis* in shallow waters. Sci. Mar. 78, 555–565. https://doi.org/10.3989/scimar.04087.03A.

Prado, P., 2018. Seagrass epiphytic assemblages are strong indicators of agricultural discharge but weak indicators of host features. Estuar. Coast. Shelf Sci. 204, 140–148. https://doi.org/10.1016/j.ecss.2018.02.026.

Prado, P., Andree, K.B., Trigos, S., Carrasco, N., Caiola, N., García-March, J.R., Tena, J., Fernández-Tejedor, M., Carella, F., 2020. Breeding, planktonic and settlement factors shape recruitment patterns of one of the last remaining major population of *Pinna nobilis* within Spanish waters. Hydrobiologia 847, 771–786. https://doi.org/10.1007/s10750-019-04137-5.

Prado, P., Grau, A., Catanese, G., Cabanes, P., Carella, F., Fernández-Tejedor, M., Andree, K.B., Añón, T., Hernandis, S., Tena, J., García-March, J.R., 2021. Pinna nobilis in suboptimal environments are more tolerant to disease but more vulnerable to severe weather phenomena. Mar. Environ. Res. 163, 105220 https://doi.org/10.1016/j.marenvres.2020.105220.

Ramón, M., Cano, J., Peña, J.B., Campos, M.J., 2005. Current status and perspectives of mollusc (bivalves and gastropods) culture in the spanish Mediterranean. Bol. Inst. Esp. Oceanogr. 21, 361–373.

Rimet, F., Bouchez, A., 2012. Life-forms, cell-sizes and ecological guilds of diatoms in european rivers. Knowl. Manag. Aquat. Ecosyst. 406, 1–14. https://doi.org/10.1051/kmae/2012018.

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. Org. Divers. Evol. 18, 51–62. https://doi.org/10.1007/s13127-018-0359-5.

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

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.

Round, F.E., 1971. Benthic marine diatoms. Oceanogr. Mar. Biol. Ann. Rev. 9, 83–139.

Round, F.E., Crawford, R.M., Mann, D.G., 1990. The diatoms. Biology and morphology of the genera. Cambridge University Press, Cambridge.

Rovira, L., Trobajo, R., Ibañez, C., 2009. Periphytic diatom community in a Mediterranean salt wedge estuary: the Ebro estuary (NE Iberian Peninsula). Acta Bot. Croat. 68, 285–300.

Ryabushko, L.I., Ryabushko, V.I., 2000. Communities of diatoms on the shells of mollusks of the genus *Mytilus* L. Int. J. Algae. 2, 15–22. https://doi.org/10.1615/InterJAlgae.v2.i2.20.

Schmidt, C., Morard, R., Romero, O., Kucera, M., 2018. Diverse internal symbiont community in the endosymbiotic foraminifera *Pararotalia calcariformata*: implications for symbiont shuffling under thermal stress. Front. Microbiol. 9, 2018. https://doi.org/10.3389/fmicb.2018.02018.

Schmidt, M., Horn, S., Flieger, K., Ehlers, K., Wilhelm, C., Schnetter, R., 2012. *Synchroma pusillum* sp. nov. and other new algal isolates with chloroplast complexes confirm the Synchromophyceae (Ochrophyta) as a widely distributed group of amoeboid algae. Protist 163, 544–559. https://doi.org/10.1016/j.protis.2011.11.009.

Scholz, B., Liebezeit, G., 2012. Microphytobenthic dynamics in a Wadden Sea intertidal flat – part I: seasonal and spatial variation of diatom commnities in relation to macronutrient supply. Eur. J. Phycol. 47, 105–119. https://doi.org/10.1080/09670262.2012.663793.

Scholz, B., Liebezeit, G., 2012. Microphytobenthic dynamics in a Wadden Sea intertidal flat – part II: seasonal and spatial variability of non-diatom community components in relation to abiotic parameters. Eur. J. Phycol. 47, 120–137. https://doi.org/10.1080/09670262.2012.665251.

Stamatakis, A., 2014. Raxml version 8: a tool for phylogenetic analysis andpost-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

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

Sugie, K., Kuma, K., 2008. Resting spore formation in the marine diatom *Thalassiosira nordenskioeldii* under iron- and nitrogen-limited conditions. J. Plankton. Res. 30, 1245–1255. https://doi.org/10.1093/plankt/fbn080.

Sundbäck, K., Granéli, W., 1988. Influence of microphytobenthos on the nutrient flux between sediment and water: a laboratory study. Mar. Ecol. Prog. Ser. 43, 63–69. https://doi.org/10.3354/meps043063.

Sundbäck, K., Enoksson, V., Granéli, W., Pettersson, K., 1991. Influence of sublittoral microphytobenthos on the oxygen and nutrient flux between sediment and water: a laboratory continuous-flow study. Mar. Ecol. Prog. Ser. 74, 263–279. https://doi.org/10.3354/meps074263.

Takano, Y., Hansen, G., Fujita, D., Horiguchi, T., 2007. Serial replacement of diatom endosymbionts in two freshwater dinoflagellates, *Peridiniopsis* spp. (Peridiniales, Dinophyceae). Phycologia 47, 41–53. https://doi.org/10.2216/07-36.1.

Totti, C., Romagnoli, T., De Stefano, M., Di Camillo, C.G., Bavestrello, G., 2011. The diversity of epizoic diatoms: relationships between diatoms and marine invertebrates. In: Seckbach, J., Dubinsky, Z. (Eds.), All Flesh Is Grass. Plant-Animal Interrelationships. Springer, Netherlands, pp. 327–343.

Triska, F.J., Oremland, R.S., 1981. Denitrification associated with periphyton communities. Appl. Environ. Microbiol. 42, 745–748. https://doi.org/10.1128/aem.42.4.745-748.1981.

Trobajo, R., Sullivan, M.J., 2010. Applied diatom studies in estuaries and shallow coastal environments. In: Smol, J.P., Stoermer, E.F. (Eds.), The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press, Cambridge, UK, pp. 309–323.

Trobajo, R., Quintana, X.D., Sabater, S., 2004. Factors affecting the periphytic diatom community in Mediterranean coastal wetlands (Empordà wetlands, NE Spain). Arch. Hydrobiol. 160, 375–399. https://doi.org/10.1127/0003-9136/2004/0160-0375.

Underwood, G.J.C., Provot, L., 2000. Determining the environmental preferences of four epipelic diatom taxa: growth across a range of salinities, nitrate and ammonium conditions. Eur. J. Phycol. 35, 173–182. https://doi.org/10.1080/09670260010001735761.

Underwood, G.J.C., Perkins, R.G., Consalvey, M.C., Hanlon, A.R.M., Oxborough, K., Baker, N.R., Paterson, D.M., 2005. Patterns in microphytobenthic primary productivity: species-specific variation in migratory rhythms and photosynthetic efficiency in mixed-species biofilms. Limnol. Oceanogr. 50, 755–767. https://doi.org/10.4319/lo.2005.50.3.0755.

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

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

Virta, L., Gammal, J., Järnström, M., Bernard, G., Soininen, J., Norkko, J., Norkko, A., 2019. The diversity of benthic diatoms affects ecosystem productivity in heterogeneous coastal environments. Ecology 100, e02765. https://doi.org/10.1002/ecy.2765.

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. Appl. Environ. Microbiol. 73, 5261–5267. https://doi.org/10.1128/AEM.00062-07.

Wetherbee, R., Bringloe, T.T., Costa, J.F., van de Meene, A., Andersen, R.A., Verbruggen, H., 2021. New pelagophytes show a novel mode of algal colony development and reveal a perforated theca that may define the class. J. Phycol. 57, 396–411. https://doi.org/10.1111/jpy.13074.

Witak, M., Pędziński, J., Olwa, S., Hetko, D., 2020. Biodiversity of benthic diatom flora in the coastal zone of Puck Bay (southern Baltic Sea): a case study of the Hel Peninsula. Oceanol. Hydrobiol. Stud. 49, 304–318. https://doi.org/10.1515/ohs-2020-0027.

Yamada, N., Bolton, J.J., Trobajo, R., Mann, D.G., Dąbek, P., Witkowski, A., Onuma, R., Horiguchi, T., Kroth, P.G., 2019. Discovery of a kleptoplastic 'dinotom' dinoflagellate and the unique nuclear dynamics of converting kleptoplastids to permanent plastids. Sci. Rep. 9, 10474. https://doi.org/10.1038/s41598-019-46852-y.

Yamada, N., Sakai, H., Onuma, R., Kroth, P.G., Horiguchi, T., 2020. Five non-motile dinotom dinoflagellates of the genus *Dinothrix*. Front. Plant Sci. 11, 1764.

You, X., Luo, Z., Su, Y., Gu, L., Gu, H., 2015. *Peridiniopsis jiulongensis*, a new freshwater dinoflagellate with a diatom endosymbiont from China. Nova Hedwigia 101, 313–326. https://doi.org/10.1127/nova_hedwigia/2015/0272.

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

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

Supplementary material

Supplementary Table 1. Valve count of the species identified by LM across all the 9 samples and total relative abundance for the whole inventory (%). The growth-form of each taxon is also indicated. Note that *Thalassiosira profunda* is not listed in the table because it was not found during the count of 300-400 valves but it was detected by a more exhaustive examination of the slides performed after analysing the metabarcoding data.

Taxon	Growth-forms	E5 - Crassostrea gigas	E8 - Crassostrea gigas	E9 - Pinna nobilis biofilm	E10 - Pinna nobilis sediment	E11 - Pinna nobilis biofilm	E12 - Pinna nobilis biofilm	E13 - Pinna nobilis sediment	E14 - Cymodocea nodosa	E15 - Caulerpa prolifera	Relative abundance (%)
Berkeleya fennica	High profile	11	19	1	1	2	2	0	22	8	2.18
Berkeleya cf. fragilis	High profile	10	26	0	0	0	0	0	6	0	1.39
Navicula cf. ramosissima	High profile	0	0	0	0	0	30	0	0	4	1.12
Toxarium undulatum	High profile	0	0	5	10	2	2	1	5	0	0.82
Berkeleya rutilans	High profile	4	4	0	0	0	0	0	12	0	0.66
Hyalosynedra laevigata	High profile	0	0	0	1	16	1	0	0	0	0.59
Hyalosynedra sp.1	High profile	0	0	0	0	6	5	2	1	4	0.59
Grammatophora marina	High profile	0	5	0	0	4	0	0	0	8	0.56
Achnanthes longipes	High profile	13	3	0	0	0	0	0	0	0	0.53
Hyalosynedra sub-laevigata	High profile	2	0	0	1	6	0	2	0	4	0.49
Nitzschia angularis	High profile	0	0	14	1	0	0	0	0	0	0.49
Licmophora paradoxa	High profile	1	0	0	1	0	0	0	2	5	0.3
Parlibellus berkeleyi	High profile	0	9	0	0	0	0	0	0	0	0.3
Ardissonea crystallina	High profile	0	0	2	3	2	0	1	0	0	0.26
Melosira nummuloides	High profile	8	0	0	0	0	0	0	0	0	0.26
Navicula ramosissima	High profile	7	0	0	0	0	0	0	0	0	0.23
Licmophora debilis	High profile	5	0	0	0	0	0	0	0	0	0.16
Grammatophora oceanica	High profile	0	0	0	0	0	0	0	0	4	0.13
Licmophora abbreviata	High profile	4	0	0	0	0	0	0	0	0	0.13
Licmophora flabellata	High profile	0	1	0	0	1	1	0	1	0	0.13
Achnanthes cf brevipes	High profile	3	0	0	0	0	0	0	0	0	0.1
Berkeleya scopulorum	High profile	0	0	1	0	0	0	0	1	0	0.07
Hyalosira sp.1	High profile	0	0	0	0	0	0	0	1	1	0.07
Hyalosynedra parietina	High profile	0	0	0	0	0	2	0	0	0	0.07
Neosynedra provincialis	High profile	0	0	0	0	0	0	0	1	1	0.07
Ardissonea sp.1	High profile	0	0	0	0	0	1	0	0	0	0.03
Cyclophora tenuis	High profile	0	0	0	0	0	0	0	1	0	0.03
Divergita toxoneides	High profile	0	0	0	0	0	0	1	0	0	0.03
Licmophora oedibus	High profile	0	0	0	0	0	1	0	0	0	0.03

Nitzschia martiana	High profile	0	0	0	0	0	0	1	0	0	0.03
Nitzschia vidovichi	High profile	0	0	0	0	0	1	0	0	0	0.03
Striatella unipunctata	High profile	0	0	0	0	0	0	0	1	0	0.03
Parlibellus sp.1	High profile?	0	1	0	0	0	0	0	0	0	0.03
Amphora helenensis	Low profile	47	80	86	22	11	12	2	0	20	9.24
Amphora cf helenensis	Low profile	0	62	46	10	0	0	4	0	8	4.29
Cocconeis scutellum v. posidoniae	Low profile	3	1	10	27	3	11	7	25	14	3.33
Nanofrustulum shiloi	Low profile	1	0	0	3	0	13	60	0	3	2.64
Cocconeis scutellum	Low profile	6	2	5	12	6	22	7	0	13	2.41
Serratifera sp.3	Low profile	0	0	0	0	0	0	46	0	7	1.75
Halamphora sp.2	Low profile	0	0	0	0	0	7	0	16	29	1.72
Plagiogramma minus	Low profile	0	0	0	15	8	1	21	0	4	1.62
Pteroncola marina	Low profile	0	1	0	0	4	0	1	41	0	1.55
Amphora aff. helenensis	Low profile	45	0	0	0	0	0	0	0	0	1.48
Mastogloia crucicula	Low profile	0	0	5	5	11	1	14	0	3	1.29
Halamphora cf. luciae	Low profile	2	7	2	8	3	1	10	0	1	1.12
Serratifera sp.2	Low profile	0	0	2	0	25	4	0	0	0	1.02
Cocconeis euglypta	Low profile	0	1	0	0	9	10	1	1	8	0.99
Opephora pacifica	Low profile	0	0	1	12	3	3	6	2	0	0.89
Tabularia investiens	Low profile	0	7	2	0	0	1	0	3	8	0.69
Amphora marina	Low profile	0	0	0	7	0	0	12	0	0	0.63
Halamphora acutiuscula	Low profile	0	1	12	4	0	0	0	0	1	0.59
Serratifera sp. 4	Low profile	0	0	0	17	0	0	0	0	0	0.56
Amphora cf. marina	Low profile	6	3	2	0	4	2	0	0	0	0.56
Cocconeis dirupta	Low profile	15	2	0	0	0	0	0	0	0	0.56
Cocconeis peltoides	Low profile	0	1	0	6	1	0	8	0	0	0.53
Diplonei vacillans	Low profile	0	0	11	0	1	0	1	0	1	0.46
Amphora exilitata	Low profile	0	0	0	4	1	0	8	0	1	0.46
Tabularia cf.parva	Low profile	0	0	0	0	0	4	0	2	8	0.46
Amphora inconspicua	Low profile	0	1	2	3	1	1	2	2	0	0.4
Halamphora semperpaloru m	Low profile	0	0	0	0	0	2	7	2	1	0.4
Halamphora sp.1	Low profile	0	0	0	0	0	12	0	0	0	0.4
Lunella ghalebii	Low profile	0	0	0	1	1	2	8	0	0	0.4
Rhopalodia acuminata	Low profile	1	0	0	0	11	0	0	0	0	0.4
Delphineis livingstonii	Low profile	0	0	1	5	2	0	2	0	1	0.36
Nagumonea sp.1	Low profile	0	0	0	0	0	0	0	11	0	0.36
Cocconeis neothumensis var marina	Low profile	3	0	0	0	0	0	3	0	4	0.33
Gedaniella guenter-grassii	Low profile	0	0	0	0	0	0	10	0	0	0.33
Mastogloia cf. emarginata	Low profile	0	0	0	0	10	0	0	0	0	0.33
Opephora sp.2	Low profile	0	0	0	0	0	0	10	0	0	0.33

Tabularia cf. fasciculata	Low profile	1	0	0	1	0	0	5	3	0	0.33
Plagiogramma nanum	Low profile	0	0	0	1	4	0	4	0	0	0.3
Opephora cf. marina	Low profile	0	0	0	0	2	0	6	0	0	0.26
Amphora sp.2	Low profile	0	0	0	0	7	0	0	0	0	0.23
Mastogloia pusilla	Low profile	0	0	0	1	2	1	3	0	0	0.23
Achnanthes meridionalis	Low profile	6	0	0	0	0	0	0	0	0	0.2
Serratifera sp.1	Low profile	0	0	0	0	6	0	0	0	0	0.2
Achnanthes sp.1	Low profile	4	0	0	0	0	0	0	0	0	0.13
Amphora cf. proteus	Low profile	1	0	0	1	0	0	2	0	0	0.13
Amphora sp.4	Low profile	0	0	0	0	0	0	4	0	0	0.13
Auricula hoffmannii	Low profile	0	0	0	0	0	0	4	0	0	0.13
Cocconeis septentrionalis	Low profile	0	4	0	0	0	0	0	0	0	0.13
Cocconeis sp.2	Low profile	0	0	0	4	0	0	0	0	0	0.13
Halamphora coffeaeformis	Low profile	0	0	0	0	1	2	0	1	0	0.13
Hippodonta sp.2	Low profile	0	0	0	4	0	0	0	0	0	0.13
Mastogloia ovulum	Low profile	0	0	0	0	3	1	0	0	0	0.13
Mastogloia robusta	Low profile	0	0	0	0	2	0	2	0	0	0.13
Serratifera cf. andersonii	Low profile	0	0	0	4	0	0	0	0	0	0.13
Tabularia sp.1	Low profile	0	4	0	0	0	0	0	0	0	0.13
Amphora cf. arenaria	Low profile	3	0	0	0	0	0	0	0	0	0.1
Amphora kolbei	Low profile	0	2	1	0	0	0	0	0	0	0.1
Cocconeis diaphana	Low profile	2	0	0	0	1	0	0	0	0	0.1
Cocconeis distans	Low profile	0	1	0	0	0	1	1	0	0	0.1
Cocconeis pelta	Low profile	0	1	0	0	1	0	1	0	0	0.1
Cocconeis sp.1	Low profile	0	0	3	0	0	0	0	0	0	0.1
Diploneis sp.1	Low profile	1	0	2	0	0	0	0	0	0	0.1
Fallacia cf. clepsidroides	Low profile	0	0	0	2	1	0	0	0	0	0.1
Halamphora sp.3	Low profile	0	0	0	0	0	0	3	0	0	0.1
Madinithidium sp.1	Low profile	0	0	0	0	1	2	0	0	0	0.1
Mastogloia cf. multicostata	Low profile	0	0	0	0	1	2	0	0	0	0.1
Mastogloia erythraea	Low profile	0	0	0	0	0	0	3	0	0	0.1
Opephora horstiana	Low profile	0	0	0	0	0	0	3	0	0	0.1
Opephora sp.1	Low profile	0	0	0	3	0	0	0	0	0	0.1
Stauroforma sp.1	Low profile	0	0	0	3	0	0	0	0	0	0.1
Stauroforma sp.3	Low profile	0	0	0	0	0	0	3	0	0	0.1
Tabularia sp.1	Low profile	3	0	0	0	0	0	0	0	0	0.1
Vikingea sp.1	Low profile	3	0	0	0	0	0	0	0	0	0.1
Amicula specululum	Low profile	0	0	0	0	0	2	0	0	0	0.07
Amphicocconeis sp.1	Low profile	0	0	0	1	1	0	0	0	0	0.07
Amphora cf. inconspicua	Low profile	2	0	0	0	0	0	0	0	0	0.07
Amphora incrassata	Low profile	0	0	0	2	0	0	0	0	0	0.07
Amphora micrometra	Low profile	2	0	0	0	0	0	0	0	0	0.07
Astartiella sp.1	Low profile	0	0	0	1	0	0	1	0	0	0.07

Chamaepinnularia cf. wiktoriae	Low profile	0	0	0	2	0	0	0	0	0	0.07
Cymbellonitzschia sp.1	Low profile	0	0	0	0	2	0	0	0	0	0.07
Fallacia forcipata	Low profile	0	0	1	1	0	0	0	0	0	0.07
Fragilaria cf. bronkei	Low profile	0	0	0	1	0	0	1	0	0	0.07
Halamphora kolbei	Low profile	2	0	0	0	0	0	0	0	0	0.07
Halamphora tenerrima	Low profile	0	0	0	0	0	2	0	0	0	0.07
Hippodonta sp.4	Low profile	0	0	0	0	0	2	0	0	0	0.07
Mastogloia cuneata	Low profile	0	0	0	0	0	0	2	0	0	0.07
Mastogloia ovata	Low profile	0	0	0	0	0	2	0	0	0	0.07
Mastogloia sp.1	Low profile	2	0	0	0	0	0	0	0	0	0.07
Proschkinia browderiana	Low profile	0	0	2	0	0	0	0	0	0	0.07
Stauroforma sp.2	Low profile	0	0	0	0	2	0	0	0	0	0.07
Tabularia sp.2	Low profile	0	0	0	0	2	0	0	0	0	0.07
Tabularia sp.3	Low profile	0	0	0	0	0	2	0	0	0	0.07
Achnanthes pseudogroenlandica	Low profile	1	0	0	0	0	0	0	0	0	0.03
Achnanthes sanctipauli	Low profile	1	0	0	0	0	0	0	0	0	0.03
Amphora caroliniana	Low profile	0	0	0	1	0	0	0	0	0	0.03
Amphora immarginata	Low profile	1	0	0	0	0	0	0	0	0	0.03
Amphora pannucea	Low profile	1	0	0	0	0	0	0	0	0	0.03
Amphora sp.1	Low profile	0	0	0	1	0	0	0	0	0	0.03
Campylodiscus cf fastuosus	Low profile	0	0	0	0	0	1	0	0	0	0.03
Campyloneis sp.1	Low profile	0	0	0	0	0	1	0	0	0	0.03
Cocconeiopsis cf. patrickae	Low profile	0	0	0	1	0	0	0	0	0	0.03
Cocconeis barleyi	Low profile	0	0	0	0	1	0	0	0	0	0.03
Cocconeis costata	Low profile	1	0	0	0	0	0	0	0	0	0.03
Cocconeis krammeri	Low profile	1	0	0	0	0	0	0	0	0	0.03
Cocconeis molesta	Low profile	0	1	0	0	0	0	0	0	0	0.03
Diploneis cf. papula	Low profile	0	0	0	1	0	0	0	0	0	0.03
Diploneis decipiens var. parallela	Low profile	0	0	0	0	1	0	0	0	0	0.03
Diploneis sp.2	Low profile	0	0	0	1	0	0	0	0	0	0.03
Fallacia floriniae	Low profile	0	0	0	1	0	0	0	0	0	0.03
Halamphora spriggerica	Low profile	1	0	0	0	0	0	0	0	0	0.03
Halamphora yundangensis	Low profile	0	0	1	0	0	0	0	0	0	0.03
Hippodonta sp.1	Low profile	0	0	0	0	0	0	0	0	1	0.03
Hippodonta sp.3	Low profile	0	0	0	0	1	0	0	0	0	0.03
Lunella sp.1	Low profile	0	0	0	0	0	1	0	0	0	0.03
Lyrella cf abrupta	Low profile	0	0	0	1	0	0	0	0	0	0.03
Lyrella cf. atlantica	Low profile	0	0	0	0	0	0	1	0	0	0.03
Mastogloia acutiuscula	Low profile	0	0	0	0	1	0	0	0	0	0.03
Mastogloia binotata	Low profile	0	0	0	1	0	0	0	0	0	0.03

Mastogloia biocellata	Low profile	0	0	0	0	1	0	0	0	0	0.03
Mastogloia cf. corsicana	Low profile	0	0	0	1	0	0	0	0	0	0.03
Mastogloia cf. lanceolata	Low profile	0	0	0	0	0	0	1	0	0	0.03
Mastogloia sp.2	Low profile	0	0	0	1	0	0	0	0	0	0.03
Mastogloia sp.3	Low profile	0	0	0	1	0	0	0	0	0	0.03
Prestauroneis sp.1	Low profile	0	1	0	0	0	0	0	0	0	0.03
Prestauroneis sp.2	Low profile	0	0	0	0	1	0	0	0	0	0.03
Pteroncola sp.	Low profile	0	0	0	0	0	0	0	1	0	0.03
Navicula sp.4	Motile	31	35	0	8	19	57	1	129	14	9.7
Navicula normaloides	Motile	17	34	10	1	6	9	1	5	9	3.04
Navicula normalis	Motile	6	0	26	5	1	6	3	6	24	2.54
Navicula subagnita	Motile	0	2	16	7	2	0	1	22	1	1.68
Nitzschia frustulum	Motile	12	6	0	2	1	2	0	1	16	1.32
Nitzschia sp section Dissipatae	Motile	1	10	0	3	2	3	1	9	0	0.96
Seminavis robusta	Motile	1	6	4	0	12	3	1	0	0	0.89
Craspedostauros sp.2	Motile	25	0	0	0	0	0	0	0	0	0.82
Navicula pavillardi	Motile	0	5	1	1	0	5	2	10	1	0.82
Navicula sp.6	Motile	3	19	0	0	0	0	1	0	0	0.76
Navicula sp.3	Motile	0	0	0	0	0	0	0	0	22	0.73
Nitzschia liebetruthii	Motile	0	0	0	0	0	5	3	1	12	0.69
Nitzschia inconspicua	Motile	9	3	0	3	1	1	0	1	2	0.66
Navicula perminuta	Motile	6	6	0	0	0	0	0	0	5	0.56
Nitzschia cf. grossestriata	Motile	1	0	1	0	0	2	0	2	9	0.49
Craspedostauros sp.3	Motile	14	0	0	0	0	0	0	0	0	0.46
Nitzschia dissipata	Motile	0	0	12	0	0	0	0	0	0	0.4
Caloneis formosa var. densestriata	Motile	0	0	0	0	1	10	0	0	0	0.36
Navicula sp.2	Motile	0	0	0	0	0	0	0	10	0	0.33
Craspedostauros sp.1	Motile	9	0	0	0	0	0	0	0	0	0.3
Nitzschia paleacea	Motile	2	0	0	1	3	0	0	0	3	0.3
Seminavis sp.3	Motile	0	1	0	0	8	0	0	0	0	0.3
Navicula sp.1	Motile	0	0	0	0	0	6	0	0	0	0.2
Arcuatasigma sp.1	Motile	0	0	1	1	0	3	0	0	0	0.16
Navicula cancellata	Motile	0	0	0	3	1	0	1	0	0	0.16
Navicula cf. dehissa	Motile	0	1	1	1	1	0	1	0	0	0.16
Nitzschia cf. linkei	Motile	0	0	0	0	0	0	5	0	0	0.16
Gyrosigma coelophilum	Motile	0	0	4	0	0	0	0	0	0	0.13
Navicula gregaria	Motile	1	3	0	0	0	0	0	0	0	0.13
Psammodictyon sp.2	Motile	0	0	0	0	0	4	0	0	0	0.13
Tetramphora sulcata	Motile	0	1	0	1	0	0	0	2	0	0.13
Navicula phylleptosoma	Motile	0	3	0	0	0	0	0	0	0	0.1
Nitzschia cf. hybrida	Motile	3	0	0	0	0	0	0	0	0	0.1
Psammodictyon sp.3	Motile	0	0	0	0	0	0	3	0	0	0.1
Bacillaria sp.	Motile	0	0	0	1	0	0	1	0	0	0.07
Caloneis liber	Motile	0	1	0	0	0	1	0	0	0	0.07

Navicula johnsonii	Motile	0	0	0	2	0	0	0	0	0	0.07
Navicula sp.5	Motile	2	0	0	0	0	0	0	0	0	0.07
Navicula sp.7	Motile	0	2	0	0	0	0	0	0	0	0.07
Navicula sp.8	Motile	0	0	0	0	0	2	0	0	0	0.07
Navicula veneta	Motile	0	0	0	0	2	0	0	0	0	0.07
Nitzschia cf. aurariae	Motile	1	0	0	0	1	0	0	0	0	0.07
Nitzschia cf. palea	Motile	2	0	0	0	0	0	0	0	0	0.07
Nitzschia navicularis	Motile	0	0	0	0	0	2	0	0	0	0.07
Nitzschia sp.1	Motile	0	0	0	0	0	2	0	0	0	0.07
Okedenia cf inflexa	Motile	0	0	2	0	0	0	0	0	0	0.07
Plagiotropis pusilla	Motile	0	0	0	1	0	0	1	0	0	0.07
Pleurosigma cf. aestuarii	Motile	0	0	0	0	1	0	0	0	1	0.07
Psammodictyon coarctata	Motile	0	1	0	0	0	0	1	0	0	0.07
Seminavis cf. insignis	Motile	2	0	0	0	0	0	0	0	0	0.07
Trachyneis aspera	Motile	0	0	0	0	1	1	0	0	0	0.07
Arcuatasigma sp.2	Motile	0	0	0	0	0	0	1	0	0	0.03
Craspedostauros sp.4	Motile	1	0	0	0	0	0	0	0	0	0.03
Donkinia sp.1	Motile	0	0	0	0	0	0	0	1	0	0.03
Entomoneis decussata	Motile	0	0	0	1	0	0	0	0	0	0.03
Entomoneis sp.1	Motile	0	0	1	0	0	0	0	0	0	0.03
Entomoneis sp.2	Motile	0	0	1	0	0	0	0	0	0	0.03
Haslea sp.1	Motile	0	0	0	0	0	1	0	0	0	0.03
Navicula bipustulata	Motile	1	0	0	0	0	0	0	0	0	0.03
Navicula cf. oblonga	Motile	0	0	0	0	0	0	1	0	0	0.03
Navicula cf. pavillardii	Motile	1	0	0	0	0	0	0	0	0	0.03
Navicula phyllepta	Motile	0	1	0	0	0	0	0	0	0	0.03
Navicula salinarum	Motile	0	0	1	0	0	0	0	0	0	0.03
Nitzschia angularis v. minor	Motile	0	0	0	0	0	1	0	0	0	0.03
Nitzschia cf. composita	Motile	1	0	0	0	0	0	0	0	0	0.03
Nitzschia composita	Motile	0	1	0	0	0	0	0	0	0	0.03
Nitzschia insignis	Motile	0	0	0	1	0	0	0	0	0	0.03
Nitzschia microcephala	Motile	1	0	0	0	0	0	0	0	0	0.03
Nitzschia sigma	Motile	0	0	1	0	0	0	0	0	0	0.03
Nitzschia sp.2	Motile	0	0	1	0	0	0	0	0	0	0.03
Plagiotropis cf. lepidoptera	Motile	0	0	0	0	0	0	0	0	1	0.03
Psammodictyon cf. areolatum	Motile	1	0	0	0	0	0	0	0	0	0.03
Psammodictyon sp.1	Motile	0	0	0	1	0	0	0	0	0	0.03
Psammodictyon sp.4	Motile	0	0	0	0	0	0	1	0	0	0.03
Psammodictyon sp.5	Motile	0	0	0	0	0	0	0	1	0	0.03
Seminavis sp.1	Motile	1	0	0	0	0	0	0	0	0	0.03
Seminavis sp.2	Motile	0	0	1	0	0	0	0	0	0	0.03
Tetramphora ostrearia	Motile	0	0	0	1	0	0	0	0	0	0.03
Tetramphora sp.1	Motile	0	0	0	0	0	1	0	0	0	0.03
Tryblionella apiculata	Motile	1	0	0	0	0	0	0	0	0	0.03

Thalassionema bacillaris	Planktonic	0	0	2	2	10	0	3	0	0	0.56
Chaetoceros sp.1	Planktonic	0	0	0	5	4	0	0	0	0	0.3
Neofragilaria sp. nov.	planktonic	0	0	0	0	0	0	3	0	0	0.1
Thalassionema nitzschioides var. lanceolata	Planktonic	0	0	0	1	0	0	0	0	0	0.03
Cyclotella choctawhatcheeana	Tychoplanktonic	0	0	12	21	21	8	6	9	1	2.57
Nitzschia linkei	Tychoplanktonic	0	0	0	0	1	1	0	0	0	0.07
Cyclotella sp.1	Tychoplanktonic	0	0	0	0	0	0	1	0	0	0.03
Cylindrotheca sp.1	Tychoplanktonic	0	0	0	0	0	1	0	0	0	0.03
Nitzschia socialis	Tychoplanktonic	0	0	0	0	0	0	1	0	0	0.03

Supplementary Table 2. List of diatom ASVs identified in this study together with their corresponding taxonomy affiliation, growth-form, reads distribution among the 9 samples analysed and their relative abundance over all the inventory. Taxonomy of ASVs was determined using Diat.barcode v9 and setting 85% as the minimum confidence threshold in DADA2. When ASVs were not classified on the basis of previous criteria, the taxonomic affiliation was assigned in the cases where ASVs shared ≥ 97% of similarity with sequences from the database of NCBI GenBank. Note that some ASVs could not be classified by any of the previous criteria but they could at genus level through an evaluation of the most similar sequences in GenBank

ASV id	Taxonomy based on Diat.barcode v9	Taxonomy based on GenBank	Growth-forms	E5-Crassostrea gigas	E8-Crassostrea gigas	E9-Pinna nobilis biofilm	E10-Pinna nobilis sediment	E11-Pinna nobilis biofilm	E12-Pinna nobilis biofilm	E13-Pinna nobilis sediment	E14-Cymodocea nodosa	E15-Caulerpa prolifera	Relative abundance (%)
ASV0001	*Thalassiosira profunda*	*N/A*	Planktonic	27	19	566	1591	671	669	823	516	2041	27.69
ASV0002	*Berkeleya fennica*	*N/A*	High profile	364	157	77	6	19	73	36	841	2	6.30
ASV0003	*Achnanthes longipes*	*N/A*	High profile	1180	821	0	0	0	0	0	0	0	8.00
ASV0006	*Nanofrustulum shiloi*	*N/A*	Low profile	2	5	91	58	78	274	291	25	472	5.18
ASV0010	*Unclassified*	*Navicula sp.*	Motile	0	0	119	265	16	61	311	24	0	3.18
ASV0014	*Haslea howeana*	*N/A*	Motile	0	0	37	0	50	11	14	398	0	2.04
ASV0019	*Cyclotella sp.*	*N/A*	TychoPlanktonic	0	0	28	223	206	6	19	11	49	2.17
ASV0022	*Unclassified*	*Seminavis cf. robusta*	Motile	0	0	0	0	394	0	0	0	0	1.58
ASV0026	*Licmophora paradoxa*	*N/A*	High profile	97	96	0	0	4	43	51	58	13	1.45
ASV0030	*Navicula perminuta*	*N/A*	Motile	46	0	0	0	0	72	7	175	9	1.24
ASV0031	*Seminavis robusta*	*N/A*	Motile	77	261	0	0	0	0	0	0	0	1.35
ASV0034	*Psammodictyon sp.*	*N/A*	Motile	15	16	11	31	27	44	109	7	63	1.29
ASV0035	*Craspedostauros constricta*	*N/A*	Motile	348	18	0	0	0	0	0	0	0	1.46
ASV0037	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	0	0	38	187	0	0	0.90
ASV0039	*Seminavis sp.*	*N/A*	Motile	0	0	0	93	7	36	33	59	0	0.91

ASV0040	*Dimeregramma sp.*	*N/A*	Low profile	0	0	6	47	86	10	26	4	70	1.00
ASV0041	*Thalassiosira angulata*	*N/A*	Planktonic	0	0	42	40	43	17	47	11	50	1.00
ASV0045	*Unclassified*	*Halamphora maritima*	Low profile	0	0	0	0	0	209	1	0	0	0.84
ASV0046	*Nitzschia spathulata*	*N/A*	Motile	0	0	162	0	29	8	0	0	0	0.80
ASV0047	*Unclassified*	*Bacillaria sp.*	TychoPlanktonic	30	97	38	0	26	24	0	0	0	0.86
ASV0049	*Gedaniella panicellus*	*N/A*	Low profile	2	0	27	22	28	33	45	4	67	0.91
ASV0051	*Nitzschia traheaformis*	*N/A*	Motile	60	22	6	0	16	59	21	0	0	0.74
ASV0052	*Serratifera andersonii*	*N/A*	Low profile	0	0	6	24	48	37	30	2	18	0.66
ASV0053	*Amphora helenensis*	*N/A*	Low profile	48	25	12	0	11	33	0	7	64	0.80
ASV0062	*Haslea howeana*	*N/A*	Motile	63	75	0	0	0	0	0	0	0	0.55
ASV0063	*Pleurosigma sp.*	*N/A*	Motile	0	0	0	84	40	0	9	0	0	0.53
ASV0066	*Unclassified*	*Amphora helenensis*	Low profile	17	50	8	0	9	16	0	0	46	0.58
ASV0067	*Nanofrustulum sp.*	*N/A*	Low profile	0	0	12	17	25	18	21	0	44	0.55
ASV0069	*Navicula avium*	*N/A*	Motile	0	128	0	0	0	0	0	0	0	0.51
ASV0071	*Unclassified*	*Nitzschia cf. dubiiformis*	Motile	0	120	0	0	0	0	0	0	0	0.48
ASV0073	*Nitzschia ovalis*	*N/A*	Motile	0	0	6	23	3	20	33	4	26	0.46
ASV0074	*Pleurosigma sp.*	*N/A*	Motile	0	0	71	5	0	11	7	2	30	0.50
ASV0075	*Unclassified*	*Nitzschia fontifuga*	Motile	0	0	0	0	19	0	82	0	0	0.40
ASV0077	*Nitzschia spathulata*	*N/A*	Motile	0	5	118	0	0	0	0	0	0	0.49
ASV0081	*Amphora helenensis*	*N/A*	Low profile	0	101	0	0	0	0	0	0	0	0.40
ASV0082	*Achnanthes brevipes*	*N/A*	High profile	111	5	0	0	0	0	0	0	0	0.46
ASV0083	*Melosira nummuloides*	*N/A*	High profile	69	59	0	0	0	0	0	0	0	0.51

ASV0084	*Seminavis robusta*	*N/A*	Motile	0	112	0	0	0	0	0	0	0	0.45
ASV0090	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	5	9	0	69	0	0.33
ASV0092	*Entomoneis paludosa*	*N/A*	Motile	17	8	9	0	0	42	4	0	0	0.32
ASV0094	*Unclassified*	*Navicula*	Motile	0	106	0	0	0	0	0	0	0	0.42
ASV0097	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	50	23	6	7	0	0	0.34
ASV0101	*Unclassified*	*Trachyneis sp.*	Motile	0	0	0	16	0	0	13	27	0	0.22
ASV0103	*Tabularia laevis*	*N/A*	Low profile	0	15	0	0	0	11	0	43	0	0.28
ASV0106	*Navicula perminuta*	*N/A*	Motile	0	0	6	0	0	0	0	56	0	0.25
ASV0110	*Unclassified*	*Entomoneis infula*	Motile	0	0	4	33	2	5	20	0	0	0.26
ASV0112	*Striatella unipunctata*	*N/A*	High profile	0	80	0	0	0	0	0	0	0	0.32
ASV0113	*Hyalosynedra lanceolata*	*N/A*	High profile	0	0	0	0	62	0	0	0	0	0.25
ASV0115	*Unclassified*	*Nitzschia sp.*	Motile	29	20	0	1	0	1	2	11	8	0.29
ASV0116	*Striatella unipunctata*	*N/A*	High profile	5	64	0	0	0	9	0	0	0	0.31
ASV0117	*Unclassified*	*Parlibellus berkeleyi*	High profile	2	68	0	0	0	0	0	0	0	0.28
ASV0120	*Berkeleya fennica*	*N/A*	High profile	0	85	0	0	0	0	0	0	0	0.34
ASV0122	*Unclassified*	*Plagiotropis*	Motile	0	0	0	17	0	32	2	2	11	0.26
ASV0127	*Nitzschia adhaerens*	*N/A*	Motile	0	0	0	0	0	0	50	0	0	0.20
ASV0129	*Berkeleya fennica*	*N/A*	High profile	0	61	0	0	0	0	0	0	0	0.24
ASV0132	*Unclassified*	*Haslea cf. howeana*	Motile	0	0	0	9	0	0	10	27	0	0.18
ASV0136	*Achnanthes longipes*	*N/A*	High profile	67	0	0	0	0	0	0	0	0	0.27
ASV0140	*Unclassified*	*Actinoptychus octonarius*	TychoPlanktonic	0	0	0	5	38	0	5	0	0	0.19
ASV0141	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	0	0	32	21	0	0	0.21

ASV0142	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	7	0	0	0	0	0	0	40	0	0.19
ASV0144	*Unclassified*	*Nitzschia*	Motile	0	0	25	0	21	0	0	0	0	0.18
ASV0146	*Dimeregramma sp.*	*N/A*	Low profile	0	0	0	18	21	0	8	0	0	0.19
ASV0155	*Petrodictyon sp.*	*N/A*	Low profile	0	58	0	0	0	0	0	0	0	0.23
ASV0157	*Nitzschia spathulata*	*N/A*	Motile	0	0	28	0	5	3	0	0	0	0.14
ASV0159	*Pleurosigma sp.*	*N/A*	Motile	0	0	0	20	0	16	7	0	0	0.17
ASV0160	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	0	7	0	12	2	16	3	0.16
ASV0161	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	3	37	0	0	3	0	0.17
ASV0163	*Striatella unipunctata*	*N/A*	High profile	0	20	0	0	0	21	8	0	0	0.20
ASV0170	*Striatella unipunctata*	*N/A*	High profile	0	0	0	0	0	17	5	19	0	0.16
ASV0171	*Unclassified*	*Nitzschia inconspicua*	Motile	0	0	0	7	27	0	0	0	0	0.14
ASV0173	*Nitzschia dalmatica*	*N/A*	Motile	0	0	8	4	2	8	16	0	0	0.15
ASV0174	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	2	26	0	0	0.11
ASV0175	*Ditylum intricatum*	*N/A*	Planktonic	0	0	12	22	4	2	1	0	0	0.16
ASV0176	*Unclassified*	*Nitzschia*	Motile	2	0	0	0	0	0	5	28	0	0.14
ASV0178	*Berkeleya fennica*	*N/A*	High profile	0	0	0	0	0	4	24	0	0	0.11
ASV0179	*Unclassified*	*Halamphora*	Low profile	0	0	0	1	0	18	23	0	0	0.17
ASV0182	*Unclassified*	*Opephoroid sp.*	Low profile	0	0	0	5	19	0	0	4	0	0.11
ASV0185	*Licmophora abbreviata*	*N/A*	High profile	29	11	0	0	0	0	0	0	0	0.16
ASV0189	*Unclassified*	*Tetramphora*	Motile	0	0	0	0	0	0	27	0	0	0.11
ASV0192	*Pleurosigma sp.*	*N/A*	Motile	0	0	38	0	0	0	0	0	0	0.15
ASV0194	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	19	28	0	0	0.19

ASV0196	*Unclassified*	*Navicula sp.*	Motile	0	0	0	0	0	0	0	25	0	0.10
ASV0198	*Unclassified*	*Navicula*	Motile	0	0	0	16	10	0	3	5	0	0.14
ASV0199	*Unclassified*	*Trachyneis sp.*	Motile	0	0	0	0	23	0	0	0	0	0.09
ASV0202	*Unclassified*	*Cylindrotheca sp.*	TychoPlanktonic	0	0	0	11	0	10	10	0	0	0.12
ASV0203	*Unclassified*	*Cocconeis*	Low profile	0	0	0	0	0	20	0	3	35	0.23
ASV0207	*Unclassified*	*Planothidium*	Low profile	0	0	0	0	0	20	0	4	0	0.10
ASV0209	*Haslea ostrearia*	*N/A*	Motile	0	0	0	0	0	0	24	0	0	0.10
ASV0210	*Nitzschia traheaformis*	*N/A*	Motile	36	0	0	0	0	0	0	0	0	0.14
ASV0211	*Unclassified*	*Amphora fusca*	Low profile	0	0	0	0	27	2	0	0	0	0.12
ASV0212	*Pleurosigma sp.*	*N/A*	Motile	0	0	0	0	0	0	8	16	0	0.10
ASV0213	*Unclassified*	*Amphora*	Low profile	0	0	0	0	0	20	0	0	0	0.08
ASV0214	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	33	0	0	0	0	0	0	0	0.13
ASV0220	*Grammatophora oceanica*	*N/A*	High profile	0	33	0	0	0	0	0	0	0	0.13
ASV0222	*Skeletonema costatum*	*N/A*	Planktonic	0	0	0	5	0	0	8	0	0	0.05
ASV0223	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	29	0	0	3	0	0	0.13
ASV0225	*Grammatophora oceanica*	*N/A*	High profile	0	0	4	0	27	0	0	2	0	0.13
ASV0226	*Amphora sulcata*	*N/A*	Low profile	0	31	0	0	0	0	0	0	0	0.12
ASV0227	*Unclassified*	*Entomoneis sp.*	Motile	0	0	0	0	0	5	17	0	4	0.10
ASV0228	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	0	0	0	30	0	0	0.12
ASV0231	*Tabularia laevis*	*N/A*	Low profile	2	19	0	0	0	0	0	0	0	0.08
ASV0233	*Odontella mobiliensis*	*N/A*	Planktonic	0	0	0	0	11	7	1	0	0	0.08
ASV0236	*Unclassified*	*Amphora*	Low profile	0	0	0	0	21	0	0	0	0	0.08

ASV0237	*Navicula rhynchocephala var. hankensis*	*N/A*	Motile	20	8	0	0	0	0	0	0	0	0.11
ASV0239	*Unclassified*	*Nitzschia*	Motile	0	0	0	6	19	0	0	0	0	0.10
ASV0240	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	25	0	0	0	0	0	0.10
ASV0241	*Unclassified*	*Amphora*	Low profile	0	0	0	0	0	12	0	0	34	0.18
ASV0242	*Parlibellus hamulifer*	*N/A*	High profile	0	32	0	0	0	0	0	0	0	0.13
ASV0243	*Unclassified*	*Gomphonella*	Undetermined	0	0	0	14	0	0	7	0	0	0.08
ASV0245	*Unclassified*	*Trachyneis sp.*	Motile	11	19	0	0	0	0	0	0	0	0.12
ASV0247	*Unclassified*	*Plagiotropis*	Motile	0	0	0	0	0	22	0	0	0	0.09
ASV0248	*Cyclotella choctawhatcheeana*	*N/A*	TychoPlanktonic	0	0	0	22	9	0	0	0	0	0.12
ASV0249	*Navicula perminuta*	*N/A*	Motile	0	0	0	0	31	0	0	0	0	0.12
ASV0250	*Unclassified*	*Amphora*	Low profile	0	0	0	0	0	5	3	0	6	0.06
ASV0251	*Paralia sulcata*	*N/A*	High profile	0	0	6	0	8	0	0	11	0	0.10
ASV0252	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	24	0	0	0	0.10
ASV0254	*Unclassified*	*Nitzschia*	Motile	2	35	0	0	0	0	0	0	0	0.15
ASV0255	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	0	0	17	0	0.07
ASV0257	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	3	0	0	5	10	0	0	0.07
ASV0259	*Chaetoceros socialis*	*N/A*	Planktonic	3	0	0	0	9	0	4	0	0	0.06
ASV0261	*Unclassified*	*Tetramphora*	Motile	0	22	0	0	0	0	0	0	0	0.09
ASV0262	*Unclassified*	*Entomoneis*	Motile	0	0	5	7	0	0	0	0	30	0.17
ASV0263	*Chaetoceros tenuissimus*	*N/A*	Planktonic	0	10	0	6	0	1	8	0	0	0.10
ASV0264	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	0	0	3	18	0	0	0	0.08

ASV0265	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	0	18	0	0	0.07
ASV0266	*Unclassified*	*Psammodictyon*	Motile	0	0	20	0	0	0	0	0	0	0.08
ASV0267	*Unclassified*	*Seminavis*	Motile	0	0	0	0	17	0	0	0	0	0.07
ASV0268	*Haslea ostrearia*	*N/A*	Motile	0	0	0	0	0	0	16	0	0	0.06
ASV0269	*Unclassified*	*Achnanthidium*	Low profile	0	18	0	0	0	0	0	0	0	0.07
ASV0270	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	17	0	0	0	0	0	0.07
ASV0271	*Unclassified*	*Amphora*	Low profile	0	0	0	1	2	7	5	0	0	0.06
ASV0272	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	16	0	0	0	0	0.06
ASV0273	*Nitzschia spathulata*	*N/A*	Motile	0	0	11	0	8	0	0	0	0	0.08
ASV0274	*Unclassified*	*Nitzschia sp.*	Motile	7	12	0	0	0	0	0	0	0	0.08
ASV0275	*Unclassified*	*Nitzschia*	Motile	0	24	0	0	0	0	0	0	0	0.10
ASV0276	*Unclassified*	*Odontella*	Planktonic	0	0	10	6	0	0	5	0	0	0.08
ASV0278	*Unclassified*	*Nitzschia*	Motile	0	0	0	18	0	0	0	4	0	0.09
ASV0279	*Unclassified*	*Licmophora*	High profile	0	0	0	0	0	5	8	0	0	0.05
ASV0280	*Unclassified*	*Halamphora*	Low profile	0	0	0	13	2	0	4	0	0	0.08
ASV0281	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	2	11	0	0	0.05
ASV0283	*Unclassified*	*Halamphora*	Low profile	0	0	21	0	0	0	0	0	0	0.08
ASV0284	*Unclassified*	*Donkinia*	Motile	0	0	0	0	14	0	0	0	0	0.06
ASV0285	*Unclassified*	*Achnanthidium*	Low profile	5	17	0	0	0	0	0	0	0	0.09
ASV0289	*Unclassified*	*Seminavis*	Motile	0	0	0	0	0	0	0	19	0	0.08
ASV0291	*Unclassified*	*Haslea*	Motile	0	0	0	0	0	6	9	0	0	0.06
ASV0292	*Unclassified*	*Navicula*	Motile	0	0	0	1	0	0	8	3	0	0.05
ASV0293	*Pleurosigma sp.*	*N/A*	Motile	0	0	0	9	0	0	10	0	0	0.08

ASV0294	*Unclassified*	*Planothidium*	Low profile	0	16	0	0	0	0	0	0	0	0.06
ASV0295	*Papiliocellulus simplex*	*N/A*	TychoPlanktonic	3	3	0	0	0	10	0	0	0	0.06
ASV0297	*Unclassified*	*Haslea/Navicula*	Motile	0	0	0	0	14	0	0	0	0	0.06
ASV0298	*Unclassified*	*Halamphora*	Low profile	2	0	0	0	4	12	0	0	0	0.07
ASV0299	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	0	0	0	11	0	0	0.04
ASV0300	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	17	0	0	0	0	0	0.07
ASV0301	*Unclassified*	*Caloneis*	Motile	0	0	3	0	5	0	8	0	0	0.06
ASV0302	*Unclassified*	*Cocconeis*	Low profile	0	0	0	0	0	12	0	0	0	0.05
ASV0305	*Haslea pseudostrearia*	*N/A*	Motile	0	0	0	0	0	0	11	0	0	0.04
ASV0308	*Biddulphia alternans*	*N/A*	High profile	0	0	0	11	0	0	0	0	0	0.04
ASV0309	*Nitzschia volvendirostrata*	*N/A*	Motile	0	0	0	0	19	0	0	0	0	0.08
ASV0310	*Unclassified*	*Seminavis*	Motile	4	15	0	0	0	0	0	0	0	0.08
ASV0311	*Caloneis sp.*	*N/A*	Motile	0	0	0	0	9	0	0	0	0	0.04
ASV0312	*Unclassified*	*Navicula sp.*	Motile	0	0	0	0	0	16	0	0	0	0.06
ASV0313	*Unclassified*	*Navicula*	Motile	0	0	0	11	0	0	0	0	0	0.04
ASV0314	*Unclassified*	*Cocconeis cf. sigillata*	Low profile	0	0	0	0	2	0	5	0	9	0.06
ASV0316	*Amphora helenensis*	*N/A*	Low profile	0	13	0	0	3	0	0	0	0	0.06
ASV0318	*Haslea howeana*	*N/A*	Motile	0	0	0	17	0	0	0	0	0	0.07
ASV0321	*Unclassified*	*Stricosus*	Undetermined	0	0	0	0	0	9	4	0	0	0.05
ASV0323	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	0	0	0	8	0	15	0.09
ASV0327	*Hyalosira delicatula*	*N/A*	High profile	0	0	0	0	0	16	0	0	0	0.06
ASV0328	*Unclassified*	*Haslea*	Motile	0	0	0	0	0	0	11	0	0	0.04

ASV0329	*Unclassified*	*Halamphora*	Low profile	0	0	0	8	4	0	0	0	0	0.05
ASV0330	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	0	0	15	0	0	0	0	0.06
ASV0332	*Unclassified*	*Pleurosigma*	Motile	0	0	0	0	0	10	0	0	0	0.04
ASV0335	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	0	0	0	30	0.12
ASV0339	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	6	0	0	0	0	0.02
ASV0341	*Unclassified*	*Haslea sp.*	Motile	0	8	0	0	0	4	4	0	0	0.06
ASV0342	*Unclassified*	*Bacillaria*	TychoPlanktonic	0	0	0	0	0	1	8	0	0	0.04
ASV0343	*Navicula hippodontafallax*	*N/A*	Motile	4	6	0	0	0	0	0	0	0	0.04
ASV0344	*Pleurosigma sp.*	*N/A*	Motile	0	18	0	0	0	0	0	0	0	0.07
ASV0345	*Extubocellulus spinifer*	*N/A*	TychoPlanktonic	0	0	0	18	0	0	0	0	0	0.07
ASV0347	*Unclassified*	*Halamphora banzuensis*	Low profile	0	0	0	0	13	0	0	0	0	0.05
ASV0349	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	8	0	0	0	0.03
ASV0351	*Unclassified*	*Actinoptychus*	TychoPlanktonic	0	0	0	0	0	0	14	0	0	0.06
ASV0353	*Unclassified*	*Halamphora*	Low profile	0	0	0	10	1	0	4	0	0	0.06
ASV0354	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	5	1	0	0	0	0.02
ASV0355	*Unclassified*	*Nitzschia*	Motile	0	0	0	1	0	0	1	0	0	0.01
ASV0358	*Unclassified*	*Lyrella*	Low profile	0	0	0	13	0	0	0	0	0	0.05
ASV0359	*Nitzschia inconspicua*	*N/A*	Motile	15	0	0	0	0	0	0	0	0	0.06
ASV0360	*Hyalodiscus scoticus*	*N/A*	Low profile	0	15	0	0	0	0	0	0	0	0.06
ASV0362	*Unclassified*	*Halamphora*	Low profile	0	0	0	17	0	0	0	0	0	0.07
ASV0364	*Unclassified*	*Diploneia*	Low profile	0	0	0	0	11	0	0	0	0	0.04
ASV0366	*Unclassified*	*Halamphora*	Low profile	0	0	0	10	0	0	0	0	0	0.04

ASV0368	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	0	0	0	9	0	0	0.04
ASV0369	*Navicula perminuta*	*N/A*	Motile	0	14	0	0	0	0	0	0	0	0.06
ASV0373	*Unclassified*	*Amphora abludens*	Low profile	0	0	0	0	0	2	5	0	0	0.03
ASV0376	*Unclassified*	*Licmophora*	High profile	0	0	0	0	0	17	0	0	0	0.07
ASV0377	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	0	0	0	9	0	0	0.04
ASV0382	*Unclassified*	*Halamphora isumiensis*	Low profile	0	0	0	0	5	0	0	0	0	0.02
ASV0383	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	0	0	0	5	0	0	0.02
ASV0384	*Unclassified*	*Amphora*	Low profile	0	0	0	9	0	0	2	0	0	0.04
ASV0385	*Extubocellulus spinifer*	*N/A*	TychoPlanktonic	0	0	0	0	3	0	0	0	0	0.01
ASV0387	*Unclassified*	*Protokeelia*	Undetermined	0	0	0	0	0	10	0	0	0	0.04
ASV0390	*Unclassified*	*Nitzschia*	Motile	0	0	0	8	0	1	0	0	0	0.04
ASV0391	*Serratifera andersonii*	*N/A*	Low profile	0	0	0	0	7	0	0	0	0	0.03
ASV0393	*Coscinodiscus radiatus*	*N/A*	Planktonic	0	0	0	0	4	0	0	0	0	0.02
ASV0394	*Unclassified*	*Hyalosynedra*	Low profile	0	0	0	0	5	0	0	0	0	0.02
ASV0397	*Nitzschia spathulata*	*N/A*	Motile	0	0	8	0	0	0	0	0	0	0.03
ASV0399	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	0	4	0	0	0.02
ASV0400	*Unclassified*	*Navicula*	Motile	0	8	0	0	0	0	0	0	0	0.03
ASV0401	*Nitzschia spathulata*	*N/A*	Motile	0	0	9	0	0	0	0	0	0	0.04
ASV0402	*Unclassified*	*Nitzschia*	Motile	0	0	7	0	0	0	0	0	0	0.03
ASV0403	*Unclassified*	*Lyrella*	Low profile	0	0	0	6	0	0	0	0	0	0.02
ASV0404	*Unclassified*	*Nitzschia*	Motile	0	0	0	8	0	0	0	0	0	0.03
ASV0405	*Unclassified*	*Diploneis vacillans*	Low profile	0	0	0	0	3	0	0	0	0	0.01

ASV0406	*Nitzschia inconspicua*	*N/A*	Motile	0	0	0	0	3	0	0	0	0	0.01
ASV0409	*Unclassified*	*Entomoneis*	Motile	0	0	0	0	0	11	0	0	0	0.04
ASV0413	*Unclassified*	*Dimeregramma*	Low profile	0	0	0	7	0	0	0	3	0	0.04
ASV0415	*Unclassified*	*Psammodictyon sp.*	Motile	0	0	0	0	0	2	0	0	0	0.01
ASV0417	*Unclassified*	*Cylindrotheca sp.*	TychoPlanktonic	0	0	0	0	0	0	8	0	0	0.03
ASV0418	*Unclassified*	*Licmophora*	High profile	0	8	0	0	0	0	0	0	0	0.03
ASV0419	*Unclassified*	*Opephora sp.*	Low profile	0	0	0	0	4	3	0	0	0	0.03
ASV0420	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	5	0	0	0	0.02
ASV0421	*Unclassified*	*Pleurosigma*	Motile	0	0	0	0	0	0	5	0	0	0.02
ASV0423	*Unclassified*	*Navicula*	Motile	0	0	0	8	0	0	0	0	0	0.03
ASV0425	*Diploneis sp.*	*N/A*	Low profile	0	0	0	0	0	0	6	0	0	0.02
ASV0426	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	0	8	0	0	0.03
ASV0428	*Unclassified*	*Entomoneis*	Motile	0	0	10	0	0	0	0	0	0	0.04
ASV0429	*Minutocellus polymorphus*	*N/A*	TychoPlanktonic	0	0	0	7	0	0	0	0	0	0.03
ASV0432	*Unclassified*	*Nitzschia*	Motile	0	0	6	4	0	0	0	0	0	0.04
ASV0434	*Unclassified*	*Plagiogramma*	High profile	0	0	0	0	0	8	0	0	0	0.03
ASV0438	*Unclassified*	*Nitzschia*	Motile	0	6	0	0	0	0	0	0	0	0.02
ASV0440	*Chaetoceros tenuissimus*	*N/A*	Planktonic	0	0	5	0	0	0	0	0	0	0.02
ASV0441	*Unclassified*	*Nitzschia*	Motile	0	0	0	3	0	0	0	0	0	0.01
ASV0442	*Unclassified*	*Haslea*	Motile	0	0	0	0	2	0	0	0	0	0.01
ASV0443	*Unclassified*	*Amphora*	Low profile	0	0	0	0	2	0	0	0	0	0.01
ASV0444	*Unclassified*	*Plagiotropis*	Motile	0	0	0	0	8	0	0	0	0	0.03

ASV0445	*Unclassified*	*Simonsenia*	Undetermined	0	0	0	0	7	0	0	0	0	0.03
ASV0447	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	0	8	0	0	0.03
ASV0448	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	0	0	0	1	0	0	0.00
ASV0449	*Unclassified*	*Halamphora sp.*	Low profile	0	0	10	0	0	0	0	0	0	0.04
ASV0452	*Nitzschia spathulata*	*N/A*	Motile	0	0	0	0	5	0	0	0	0	0.02
ASV0454	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	0	4	0	0	0.02
ASV0456	*Unclassified*	*Papiliocellulus*	TychoPlanktonic	0	0	0	0	0	0	0	5	0	0.02
ASV0457	*Nitzschia spathulata*	*N/A*	Motile	0	0	7	0	0	0	0	0	0	0.03
ASV0460	*Unclassified*	*Hyalosynedra*	Low profile	0	0	0	0	5	0	0	0	0	0.02
ASV0465	*Unclassified*	*Entomoneis*	Motile	0	0	0	0	3	0	0	0	0	0.01
ASV0467	*Plagiotropis sp.*	*N/A*	Motile	0	0	0	0	6	0	0	0	0	0.02
ASV0471	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	9	0	0	0	0	0	0	0	0	0.04
ASV0473	*Unclassified*	*Meuniera*	Undetermined	0	0	0	5	0	0	0	0	0	0.02
ASV0474	*Unclassified*	*Caloneis*	Motile	0	0	0	4	0	0	0	0	0	0.02
ASV0475	*Pleurosigma sp.*	*N/A*	Motile	0	0	0	4	0	0	0	0	0	0.02
ASV0476	*Unclassified*	*Achnanthes*	High profile	0	0	0	7	0	0	0	0	0	0.03
ASV0477	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	0	5	0	0	0	0	0.02
ASV0478	*Unclassified*	*Planothidium*	Low profile	0	0	0	0	0	7	0	0	0	0.03
ASV0480	*Unclassified*	*Cylindrotheca*	TychoPlanktonic	0	0	0	0	0	0	5	0	0	0.02
ASV0481	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	0	0	0	14	0.06
ASV0484	*Unclassified*	*Pleurosigma*	Motile	0	0	6	0	0	0	0	0	0	0.02
ASV0485	*Unclassified*	*Navicula sp.*	Motile	0	0	0	5	0	0	0	0	0	0.02
ASV0486	*Unclassified*	*Halamphora*	Low profile	0	0	0	5	0	0	0	0	0	0.02

ASV0487	*Unclassified*	*Diploneis*	Low profile	0	0	0	0	1	0	0	0	0	0.00
ASV0490	*Unclassified*	*Pleurosigma*	Motile	0	0	0	0	0	5	0	0	0	0.02
ASV0491	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	0	6	0	0	0.02
ASV0494	*Haslea pseudostrearia*	*N/A*	Motile	0	5	0	0	0	0	0	0	0	0.02
ASV0495	*Unclassified*	*Amphora proteus*	Low profile	0	6	0	0	0	0	0	0	0	0.02
ASV0496	*Unclassified*	*Haslea*	Motile	0	0	0	3	0	0	0	0	0	0.01
ASV0498	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	0	6	0	0	0	0	0	0.02
ASV0499	*Unclassified*	*Planothidium*	Low profile	0	0	0	0	3	0	0	0	0	0.01
ASV0501	*Unclassified*	*Navicula*	Motile	0	0	0	0	0	1	0	0	0	0.00
ASV0502	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	5	0	0	0	0.02
ASV0505	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	0	3	0	0	0.01
ASV0507	*Entomoneis paludosa*	*N/A*	Motile	0	0	0	0	0	0	0	1	0	0.00
ASV0508	*Unclassified*	*Cocconeis*	Low profile	0	0	0	0	0	0	0	3	0	0.01
ASV0509	*Entomoneis sp.*	*N/A*	Motile	0	5	0	0	0	0	0	0	0	0.02
ASV0510	*Unclassified*	*Diploneis*	Low profile	0	0	1	0	0	0	0	0	0	0.00
ASV0512	*Unclassified*	*Navicula*	Motile	0	0	0	0	2	0	0	0	0	0.01
ASV0514	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	2	0	0	0	0	0.01
ASV0516	*Unclassified*	*Amphora sp.*	Low profile	0	0	0	0	0	0	1	0	0	0.00
ASV0517	*Biddulphia alternans*	*N/A*	High profile	0	0	0	0	0	0	4	0	0	0.02
ASV0518	*Unclassified*	*Nitzschia sp.*	Motile	0	0	0	0	0	0	6	0	0	0.02
ASV0520	*Nitzschia microcephala*	*N/A*	Motile	4	0	0	0	0	0	0	0	0	0.02
ASV0521	*Unclassified*	*Amphora*	Low profile	2	0	0	0	0	0	0	0	0	0.01

ASV0522	*Unclassified*	*Amphora*	Low profile	0	2	0	0	0	0	0	0	0	0.01
ASV0528	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	0	0	3	0	0.01
ASV0530	*Unclassified*	*Nitzschia*	Motile	5	0	0	0	0	0	0	0	0	0.02
ASV0531	*Unclassified*	*Seminavis*	Motile	0	4	0	0	0	0	0	0	0	0.02
ASV0534	*Unclassified*	*Cocconeis*	Low profile	0	0	0	0	2	0	0	0	0	0.01
ASV0538	*Unclassified*	*Navicula*	Motile	0	0	0	0	3	0	0	0	0	0.01
ASV0539	*Nitzschia frustulum*	*N/A*	Motile	0	0	0	0	2	0	0	0	0	0.01
ASV0540	*Unclassified*	*Navicula*	Motile	0	0	0	0	4	0	0	0	0	0.02
ASV0541	*Unclassified*	*Diploneis*	Low profile	0	0	0	0	0	4	0	0	0	0.02
ASV0542	*Arcocellulus mammifer*	*N/A*	TychoPlanktonic	0	0	0	0	0	1	0	0	0	0.00
ASV0543	*Unclassified*	*Diploneis*	Low profile	0	0	0	0	0	0	0	2	0	0.01
ASV0546	*Unclassified*	*Halamphora*	Low profile	0	0	0	0	0	0	0	0	12	0.05
ASV0548	*Unclassified*	*Amphora*	Low profile	0	0	0	0	0	0	0	0	6	0.02
ASV0549	*Nitzschia spathulata*	*N/A*	Motile	4	0	0	0	0	0	0	0	0	0.02
ASV0551	*Divergita toxoneides*	*N/A*	High profile	0	0	3	0	0	0	0	0	0	0.01
ASV0553	*Unclassified*	*Halamphora*	Low profile	0	0	0	2	0	0	0	0	0	0.01
ASV0556	*Unclassified*	*Pleurosigma sp.*	Motile	0	0	0	3	0	0	0	0	0	0.01
ASV0557	*Hyalosira delicatula*	*N/A*	High profile	0	0	0	0	0	0	1	0	0	0.00
ASV0558	*Neosynedra provincialis*	*N/A*	High profile	0	0	0	0	0	0	0	3	0	0.01
ASV0560	*Unclassified*	*Amphora*	Low profile	3	0	0	0	0	0	0	0	0	0.01
ASV0561	*Striatella unipunctata*	*N/A*	High profile	0	1	0	0	0	0	0	0	0	0.00
ASV0565	*Unclassified*	*Tetramphora*	Motile	0	0	0	4	0	0	0	0	0	0.02

ASV0568	*Unclassified*	*Amphora*	Low profile	0	0	0	0	4	0	0	0	0	0.02
ASV0569	*Unclassified*	*Tryblionella cf. compressa*	Motile	0	0	0	0	1	0	0	0	0	0.00
ASV0570	*Unclassified*	*Thalassiosira*	Planktonic	0	0	0	0	0	1	0	0	0	0.00
ASV0572	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	3	0	0	0	0.01
ASV0573	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	1	0	0	0	0.00
ASV0574	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	4	0	0	0	0.02
ASV0576	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	2	0	0	0	0.01
ASV0580	*Unclassified*	*Diploneis*	Low profile	0	0	0	0	0	0	0	1	0	0.00
ASV0581	*Unclassified*	*Schizostauron*	Undetermined	0	0	0	0	0	0	0	0	8	0.03
ASV0584	*Unclassified*	*Grammatophora*	High profile	0	3	0	0	0	0	0	0	0	0.01
ASV0585	*Unclassified*	*Sellaphora*	Motile	0	2	0	0	0	0	0	0	0	0.01
ASV0586	*Cylindrotheca closterium*	*N/A*	TychoPlanktonic	0	0	3	0	0	0	0	0	0	0.01
ASV0587	*Unclassified*	*Proschkinia*	Low profile	0	0	1	0	0	0	0	0	0	0.00
ASV0588	*Striatella unipunctata*	*N/A*	High profile	0	0	2	0	0	0	0	0	0	0.01
ASV0589	*Unclassified*	*Auricula*	Low profile	0	0	0	2	0	0	0	0	0	0.01
ASV0591	*Unclassified*	*Planothidium*	Low profile	0	0	0	0	4	0	0	0	0	0.02
ASV0592	*Asterionellopsis guyunusae*	*N/A*	Planktonic	0	0	0	0	0	1	0	0	0	0.00
ASV0598	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	2	0	0	0	0.01
ASV0603	*Unclassified*	*Planothidium*	Low profile	0	0	0	0	0	0	0	5	0	0.02
ASV0605	*Unclassified*	*Navicula*	Motile	0	3	0	0	0	0	0	0	0	0.01
ASV0606	*Unclassified*	*Navicula*	Motile	0	0	0	2	0	0	0	0	0	0.01
ASV0611	*Chaetoceros diversus*	*N/A*	Planktonic	0	0	0	3	0	0	0	0	0	0.01

ASV0614	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	1	0	0	0	0	0.00
ASV0615	*Unclassified*	*Proschkinia*	Low profile	0	0	0	0	1	0	0	0	0	0.00
ASV0624	*Unclassified*	*Protokeelia*	Undetermined	0	0	0	0	0	0	0	2	0	0.01
ASV0628	*Unclassified*	*Nitzschia*	Motile	0	0	0	0	0	0	0	0	4	0.02
ASV0631	*Unclassified*	*Chaetoceros*	Planktonic	0	2	0	0	0	0	0	0	0	0.01
ASV0633	*Unclassified*	*Nitzschia*	Motile	0	1	0	0	0	0	0	0	0	0.00
ASV0635	*Unclassified*	*Pinnularia*	Undetermined	0	0	2	0	0	0	0	0	0	0.01
ASV0653	*Chaetoceros decipiens*	*N/A*	Planktonic	0	0	0	0	0	4	0	0	0	0.02
ASV0658	*Unclassified*	*Pseudostauros ira*	Undetermined	0	0	0	0	0	1	0	0	0	0.00
ASV0660	*Unclassified*	*Licmophora*	High profile	0	0	0	0	0	0	1	0	0	0.00
ASV0664	*Unclassified*	*Achnanthes*	High profile	1	0	0	0	0	0	0	0	0	0.00
ASV0667	*Unclassified*	*Halamphora*	Low profile	0	3	0	0	0	0	0	0	0	0.01
ASV0669	*Unclassified*	*Navicula*	Motile	0	2	0	0	0	0	0	0	0	0.01
ASV0673	*Cyclotella sp.*	*N/A*	TychoPlanktonic	0	1	0	0	0	0	0	0	0	0.00

Supplementary table 3. List of all diatom species recorded in this study showing the specific method or methods that identified (indicated by √) each of them. Note that *Thalassiosira profunda* is not listed in the table as detected by LM because it was not found during the count of 300-400 valves but it was detected by a more exhaustive examination of the slides after analysing the metabarcoding data.

Species	**Detected exclusively by LM**	**Detected exclusively by DNA metabarcoding**	**Detected by both methods**
Achnanthes brevipes		√	
Achnanthes longipes			√
Achnanthes meridionalis	√		
Achnanthes pseudogroenlandica	√		
Achnanthes sanctipauli	√		
Actinoptychus octonarius		√	
Amicula specululum	√		
Amphora abludens		√	
Amphora caroliniana	√		
Amphora exilitata	√		
Amphora fusca		√	
Amphora helenensis			√
Amphora immarginata	√		
Amphora inconspicua	√		
Amphora incrassata	√		
Amphora kolbei	√		
Amphora marina	√		
Amphora micrometra	√		
Amphora pannucea	√		
Amphora proteus		√	
Amphora sulcata		√	
Arcocellulus mammifer		√	
Ardissonea crystallina	√		
Asterionellopsis guyunusae		√	
Berkeleya fennica			√
Berkeleya rutilans	√		
Berkeleya scopulorum	√		
Biddulphia alternans		√	
Caloneis formosa var. densestriata	√		
Caloneis liber	√		
Chaetoceros decipiens		√	
Chaetoceros diversus		√	
Chaetoceros socialis		√	
Chaetoceros tenuissimus		√	

Cocconeis barleyi	√			
Cocconeis costata	√			
Cocconeis diaphana	√			
Cocconeis dirupta	√			
Cocconeis distans	√			
Cocconeis euglypta	√			
Cocconeis krammeri	√			
Cocconeis molesta	√			
Cocconeis neothumensis var marina	√			
Cocconeis pelta	√			
Cocconeis peltoides	√			
Cocconeis scutellum	√			
Cocconeis septentrionalis	√			
Coscinodiscus radiatus		√		
Craspedostauros constricta		√		
Cyclophora tenuis	√			
Cyclotella choctawhatcheeana				√
Cylindrotheca closterium		√		
Delphineis livingstonii	√			
Diploneis decipiens var. parallela	√			
Diploneis vacillans				√
Ditylum intricatum		√		
Divergita toxoneides				√
Entomoneis decussata	√			
Entomoneis infula		√		
Entomoneis paludosa		√		
Extubocellulus spinifer		√		
Fallacia floriniae	√			
Fallacia forcipata	√			
Gedaniella guenter-grassii	√			
Gedaniella panicellus		√		
Grammatophora marina	√			
Grammatophora oceanica				√
Gyrosigma coelophilum	√			
Halamphora coffeaeformis	√			
Halamphora acutiuscula	√			
Halamphora banzuensis		√		
Halamphora isumiensis		√		
Halamphora kolbei	√			
Halamphora maritima		√		
Halamphora semperpalorum	√			
Halamphora spriggerica	√			
Halamphora tenerrima	√			

Halamphora yundangensis	√			
Haslea howeana		√		
Haslea ostrearia		√		
Haslea pseudostrearia		√		
Hyalodiscus scoticus		√		
Hyalosira delicatula		√		
Hyalosynedra laevigata	√			
Hyalosynedra lanceolata		√		
Hyalosynedra parietina	√			
Hyalosynedra sub-laevigata	√			
Licmophora abbreviata				√
Licmophora debilis	√			
Licmophora flabellata	√			
Licmophora oedibus	√			
Licmophora paradoxa				√
Lunella ghalebii	√			
Mastogloia acutiuscula	√			
Mastogloia binotata	√			
Mastogloia biocellata	√			
Mastogloia crucicula	√			
Mastogloia cuneata	√			
Mastogloia erythraea	√			
Mastogloia ovata	√			
Mastogloia ovulum	√			
Mastogloia pusilla	√			
Mastogloia robusta	√			
Melosira nummuloides				√
Minutocellus polymorphus		√		
Nanofrustulum shiloi				√
Navicula avium		√		
Navicula bipustulata	√			
Navicula cancellata	√			
Navicula gregaria	√			
Navicula hippodontafallax		√		
Navicula johnsonii	√			
Navicula normalis	√			
Navicula normaloides	√			
Navicula pavillardi	√			
Navicula perminuta				√
Navicula phyllepta	√			
Navicula phylleptosoma	√			
Navicula ramosissima	√			
Navicula rhynchocephala var. hankensis		√		

Navicula salinarum	√			
Navicula subagnita	√			
Navicula veneta	√			
Neosynedra provincialis				√
Nitzschia adhaerens		√		
Nitzschia angularis	√			
Nitzschia composita	√			
Nitzschia dalmatica		√		
Nitzschia dissipata	√			
Nitzschia fontifuga		√		
Nitzschia frustulum				√
Nitzschia grossestriata	√			
Nitzschia inconspicua				√
Nitzschia insignis	√			
Nitzschia liebetruthii	√			
Nitzschia linkei	√			
Nitzschia martiana	√			
Nitzschia microcephala				√
Nitzschia navicularis	√			
Nitzschia ovalis		√		
Nitzschia paleacea	√			
Nitzschia sigma	√			
Nitzschia socialis	√			
Nitzschia spathulata		√		
Nitzschia traheaformis		√		
Nitzschia vidovichi	√			
Nitzschia volvendirostrata		√		
Odontella mobiliensis		√		
Opephora horstiana	√			
Opephora pacifica	√			
Papiliocellulus simplex		√		
Paralia sulcata		√		
Parlibellus berkeleyi				√
Parlibellus hamulifer		√		
Plagiogramma minus	√			
Plagiogramma nanum	√			
Plagiotropis pusilla	√			
Proschkinia browderiana	√			
Psammodictyon coarctata	√			
Pteroncola marina	√			
Rhopalodia acuminata	√			
Seminavis robusta				√
Serratifera andersonii		√		

Skeletonema costatum		√	
Striatella unipunctata			√
Tabularia investiens	√		
Tabularia laevis		√	
Tetramphora ostrearia	√		
Tetramphora sulcata	√		
Thalassionema bacillaris	√		
Thalassionema nitzschioides var. lanceolata	√		
Thalassiosira angulata		√	
Thalassiosira profunda		√	
Toxarium undulatum	√		
Trachyneis aspera	√		
Tryblionella apiculata	√		

Supplementary Table 4. Comparison of Sørensen similarity index values between the LM and DNA metabarcoding methods. Index values are represented in a reddish (highest values) to bluish scale (lowest values).

LM	**E5 - Crassostrea gigas**	**E8 - Crassostrea gigas**	**E9 - Pinna nobilis biofilm**	**E10 - Pinna nobilis sediment**	**E11 - Pinna nobilis biofilm**	**E12 - Pinna nobilis biofilm**	**E13 - Pinna nobilis sediment**	**E14 - Cymodocea nodosa**
E8 - Crassostrea gigas	0.32							
E9 - Pinna nobilis biofilm	0.19	0.35						
E10 - Pinna nobilis sediment	0.24	0.29	0.39					
E11 - Pinna nobilis biofilm	0.26	0.37	0.37	0.47				
E12 - Pinna nobilis biofilm	0.22	0.33	0.34	0.32	0.48			
E13 - Pinna nobilis sediment	0.21	0.33	0.34	0.49	0.48	0.35		
E14 - Cymodocea nodosa	0.24	0.39	0.31	0.30	0.32	0.43	0.30	
E15 - Caulerpa prolifera	0.3	0.37	0.39	0.39	0.40	0.43	0.41	0.50

DNA metabarcoding	**E5 - Crassostrea gigas**	**E8 - Crassostrea gigas**	**E9 - Pinna nobilis biofilm**	**E10 - Pinna nobilis sediment**	**E11 - Pinna nobilis biofilm**	**E12 - Pinna nobilis biofilm**	**E13 - Pinna nobilis sediment**	**E14 - Cymodocea nodosa**
E8 - Crassostrea gigas	0.60							
E9 - Pinna nobilis biofilm	0.55	0.63						
E10 - Pinna nobilis sediment	0.71	0.63	0.73					
E11 - Pinna nobilis biofilm	0.51	0.57	0.62	0.54				
E12 - Pinna nobilis biofilm	0.53	0.55	0.58	0.54	0.67			
E13 - Pinna nobilis sediment	0.38	0.4	0.48	0.45	0.47	0.39		
E14 - Cymodocea nodosa	0.38	0.42	0.52	0.50	0.48	0.39	0.70	
E15 - Caulerpa prolifera	0.60	0.62	0.65	0.66	0.61	0.61	0.43	0.47

Supplementary Table 5. List of non-diatom ASVs, giving the algal class to which each is assigned by a combination of blastn search and (for ochrophyte ASVs) phylogenetic analysis (Fig. 4), together with (1) the number of reads of each ASV, (2) the number of samples in which it was recorded, (3) the identity, accession number, and % similarity of the top hit in GenBank, (4) the distribution of reads among the nine samples studied, (5) the relative abundances (between in parentheses next to the number of reads) of the ASVs in the nine samples. Also included are the total numbers of reads for each sample and the total number of non-diatom ASVs in each sample. Colours used for representing phyla and classes correspond to the colours code given in Fig. 4.

ASV id	Phylum or (Ochrophyta) Class	identification of closest Blastn hit	GenBank accession	% similarity to Blastn hit	E5 - Crassostrea gigas	E8 - Crassostrea gigas	E9 - Pinna nobilis biofilm	E10 - Pinna nobilis sediment	E11 - Pinna nobilis biofilm	E12 - Pinna nobilis biofilm	E13 - Pinna nobilis sediment	E14 - Cymodocea nodosa	E15 - Caulerpa prolifera
0078	Phaeophyceae	*Nemacystus decipiens*	LC382528	98,85	-	-	-	-	-	316 (1,91%)	-	-	7 (0,13%)
0128	Phaeophyceae	*Streblonema maculans*	AY157694	100	48 (0,35%)	134 (0,89%)	-	-	-	-	-	-	-
0158	Pelagophyceae	*Aureoumbra geitleri*	MT469981	95,06	-	-	-	-	-	137 (0,83%)	-	-	-
0183	Eustigmatophyceae	uncultured phytoplankton clone	KJ471775	100	-	-	4 (0,03%)	8 (0,05%)	62 (0,33%)	9 (0,05%)	28 (0,14%)	-	-
0206	Pelagophyceae	*Chrysoreinhardia giraudii*	MF927464	100	-	-	64 (0,43%)	-	-	29 (0,18%)	-	-	-
0260	Phaeophyceae	*Elachista stellaris*	LC016514	98,86	-	-	-	-	-	41 (0,25%)	-	20 (0,09%)	-
0282	Phaeophyceae	*Nemacystus decipiens*	LC382528	97,34	-	-	-	-	-	51 (0,31%)	3 (0,02%)	-	-
0287	Pelagophyceae	*Chrysoreinhardia muelleri*	MF927466	96,96	-	-	-	-	-	51 (0,31%)	-	-	-
0303	Rhodophyta	*Grania efflorescens*	KC134334	94,3	-	-	-	-	-	44 (0,27%)	-	-	-
0326	Pelagophyceae	*Aureoumbra geitleri*	MT469981	95,82	-	-	-	-	-	39 (0,24%)	-	-	-
0334	Phaeophyceae	*Myrionema balticum*	AY157694	97,72	-	-	-	-	-	-	-	38 (0,18%)	-
0350	Chlorophyta	*Pseudendoclonium akinetum*	AY835431	89,31	-	-	-	-	-	34 (0,21%)	-	-	-
0371	Phaeophyceae	*Sphacelaria tribuloides*	AJ287891	97,34	-	-	-	-	28 (0,15%)	-	-	-	-

0407	Chlorophyta	*Umbraulva yunseulla*	MT978110	93,18	-	-	-	-	22 (0,12%)	-	-	-	-
0416	Chlorophyta	*Umbraulva yunseulla*	MT978110	95,45	-	-	-	-	-	21 (0,13%)	-	-	-
0450	Pelagophyceae	*Chrysoreinhardia giraudii*	MF927464	91,63	-	-	-	9 (0,06%)	-	-	7 (0,04%)	-	-
0461	Chlorophyta	*Blidingia marginata*	HQ603480	88,35	-	-	-	-	15 (0,08%)	-	-	-	-
0463	Chlorophyta	*Pseudendoclonium akinetum*	AY835431	87,79	-	-	-	-	-	15 (0,09%)	-	-	-
0472	Chlorophyta	*Strombidium sp.*	AY257112	96,99	14 (0,1%)	-	-	-	-	-	-	-	-
0479	Synchromophyceae	*Synchroma pusillum*	JN004156	95,4	-	-	-	-	-	13 (0,08%)	-	-	-
0511	Chrysomeridophyceae	*Chrysowaernella hieroglyphica*	HQ710595	89,66	-	-	-	10 (0,07%)	-	-	-	-	-
0550	Chlorophyta	*Ulvella leptochaete*	MN515040	96,62	-	8 (0,05%)	-	-	-	-	-	-	-
0555	Dictyochophyceae	uncultured phytoplankton clone	AF381735	97,32	-	-	-	7 (0,05%)	-	-	-	-	-
0578	Chrysomeridophyceae	*Chrysowaernella hieroglyphica*	HQ710595	89,96	-	-	-	-	-	-	6 (0,03%)	-	-
0594	Phaeophyceae	*Hincksia sordida*	MT469950	96,2	-	-	-	-	-	5 (0,03%)	-	-	-
0599	Ochrophyta	*Ophiocytium sp.* (but see also *Chattonella)*	MK482704	89,96	-	-	-	-	-	-	5 (0,03%)	-	-
0600	Pelagophyceae	*Aureococcus anophagefferens*	HQ710615	88,12	-	-	-	-	-	-	5 (0,03%)	-	-
0612	Dictyochophyceae	*Apedinella radians*	HQ710599	99,24	-	-	-	4 (0,03%)	-	-	-	-	-
0621	Phaeophyceae	*Padina fraseri*	AB690274	87,16	-	-	-	-	-	-	4 (0,02%)	-	-
0626	Phaeophyceae	*Myrionema balticum*	AY157694	97,34	-	-	-	-	-	-	-	-	4 (0,07%)

0627	Chlorophyta	*Blidingia marginata*	HQ603480	88,35	-	-	-	-	-	-	-	-	4 (0,07%)
0638	Raphidophyceae	*Chattonella subsalsa*	HQ710628	97,72	-	-	-	3 (0,02%)	-	-	-	-	-
0639	Phaeophyceae	*Elachista stellaris*	LC016514	98,48	-	-	-	3 (0,02%)	-	-	-	-	-
0640	Chlorophyta	*Picochlorum sp.*	KM202138	97,37	-	-	-	3 (0,02%)	-	-	-	-	-
0647	Pinguiophyceae	*Pinguiococcus pyrenoidosus*	AF438319	90,49	-	-	-	-	-	3 (0,02%)	-	-	-
0651	Pelagophyceae	*Aureococcus anophagefferens*	HQ710615	89,19	-	-	-	-	-	3 (0,02%)	-	-	-
0652	Chrysophyceae?	uncultured phytoplankton clone	KJ471838	93,54	-	-	-	-	-	3 (0,02%)	-	-	-
0657	Chlorophyta	*Tetraselmis contracta*	MK482405	79,7	-	-	-	-	-	3 (0,02%)	-	-	-
0677	Rhodophyta	*Acrochaetium plumosum*	MF543840	94,27	-	-	-	-	-	-	-	2 (0,01%)	-
0678	Synchromophyceae	*Synchroma pusillum*	JN004156	96,96	-	-	-	-	-	-	-	2 (0,01%)	-
0679	Rhodophyta	*Pneophyllum sp.*	KM369158	95,44	-	-	-	-	-	-	-	-	2 (0,04%)

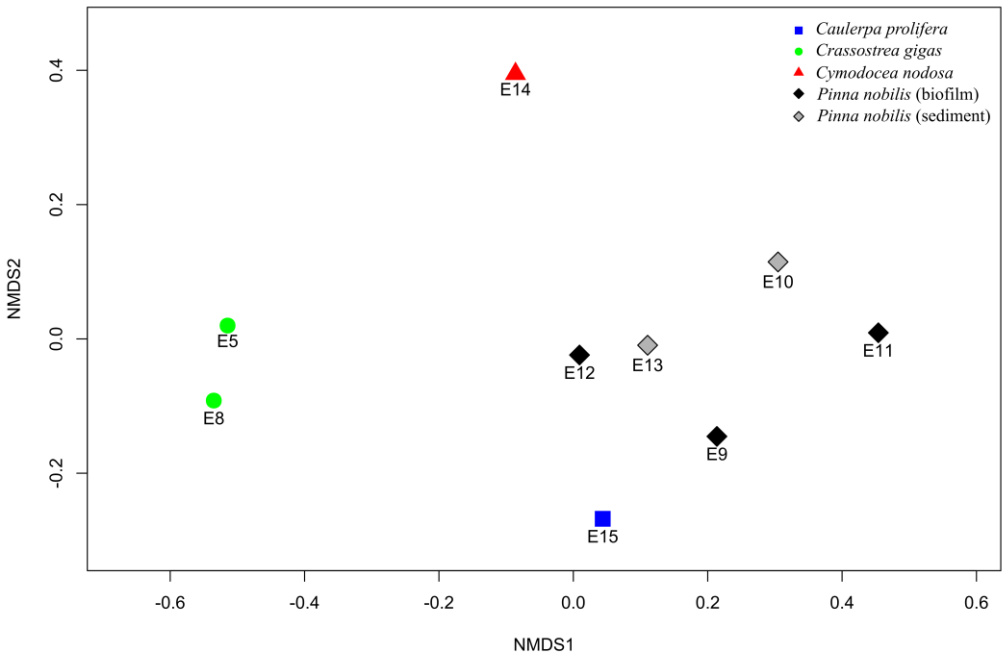
Total reads per sample	**13841**	**14976**	**14734**	**14987**	**18753**	**16546**	**19389**	**21162**	**5427**
Total non-diaom ASVs per sample	**2**	**2**	**2**	**8**	**4**	**18**	**7**	**4**	**4**

Supplementary Table 6. Comparison of Shannon diversity index values obtained by DNA metabarcoding and LM methods for the samples analysed in this study and for the samples corresponding to two biomonitoring campaigns held in Catalan rivers (NE Spain) during 2017 (ACA2017) and 2018 (ACA 2018). The indexes were computed on diatom taxa relative abundance and were based on natural logarithms. In the case of DNA ACA 2017 and ACA 2018 inventories, diatom taxa constituted those ASVs that were classified into diatoms taxa with a percentage of bootstrapping value ≥ 85%. Molecular inventories of ACA17 and ACA18 were rarefied into the minimum number of reads detected in a sample from this study (i.e. 5418 reads).

Inventory	Method	Shannon index Average	Shannon Index Maximum	Shannon Index Minimum	Shannon index Standard deviation
This study	LM	3.29	3.74	2.61	0.41
This study	DNA metabarcoding	2.50	3.04	1.59	0.50
ACA 2017	LM	2.13	3.34	0.21	0.56
ACA 2017	DNA metabarcoding	2.11	3.09	0.85	0.51
ACA 2018	LM	1.99	3.24	1.07	0.40
ACA 2018	DNA metabarcoding	2.17	3.15	0.98	0.44

Supplementary Table 7. Identifiability analyses for the ASVs obtained in the samples analysed in this study and in the samples corresponding to the biomonitoring campaigns held in Catalan rivers (NE Spain) during 2017 (ACA2017) and 2018 (ACA 2018). The table shows the total number of ASVs (Column *C*) that match at different similarity levels, from 100% to 95% (Column *B*), with reference sequences included in Diat.barcode v9 and the extent to which this number of ASVs accounts for both the total number of ASVs classified as diatoms (Column *D*) and the diatom relative abundance recorded in each inventory (Column *E*). Note that ASVs considered as diatoms were those that were classified by the Näive Bayesian classifier into the Bacillariophyta with a bootstrapping value equal to 100%. To allow inter-sample comparisons between inventories, data were rarefied to the minimum number of reads recorded in a sample from this study (i.e. 5427 reads). ACA2017 and ACA2018 inventories were composed by 162 and 125 samples respectively.

A) Inventory	**B) Percentage of similarity**	**C) Number of ASVs sharing the similarity percentage (Column *B*) with reference sequences included in Diat.barcode v9**	**D) Contribution to the total number of diatom ASVs**	**E) Contribution to diatom relative abundance (%)**
This study	100 %	40	6.86%	8.75%
This study	≥ 99 %	72	12.35%	32.22%
This study	≥ 98 %	125	21.44%	38.57%
This study	≥ 97 %	183	31.39%	52.05%
This study	≥ 96 %	270	46.31%	66.79%
This study	≥ 95 %	348	59.69%	76.66%
ACA 2017	100 %	317	16.75%	65.00%
ACA 2017	≥ 99 %	828	43.76%	81.71%
ACA 2017	≥ 98 %	1131	59.78%	89.01%
ACA 2017	≥ 97 %	1271	67.18%	91.14%
ACA 2017	≥ 96 %	1420	75.05%	94.30%
ACA 2017	≥ 95 %	1563	82.61%	96.96%
ACA 2018	100 %	332	18.20%	59.88%
ACA 2018	≥ 99 %	672	36.84%	79.01%
ACA 2018	≥ 98 %	918	50.33%	87.77%
ACA 2018	≥ 97 %	1057	57.95%	90.65%
ACA 2018	≥ 96 %	1220	66.88%	94.10%
ACA 2018	≥ 95 %	1390	76.21%	96.84%



Supplementary Fig 1. Non-metric multidimensional scaling of the Bray-Curtis dissimilarity calculated on the relative abundance of the diatom ASVs inventory that did not include the ASV taxonomically assigned as *Thalassiora profunda*.

Chapter 3

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

Pérez-Burillo, J., Mann, D. G., Trobajo, R.

Chemosphere (Under Review)

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

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

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

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

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

Abstract

Two short diatom *rbcL* barcodes, 331-bp and 263-bp in length, have frequently been used in diatom metabarcoding studies. They overlap in a common 263-bp region but differ in the presence or absence of a 68-bp tail at the 5′ end. Though the effectiveness of both has been demonstrated in separate biomonitoring and diversity studies, the impact of the 68-bp non-shared region has not been evaluated. Here we compare the two barcodes in terms of the values of a biotic index (IPS) and the ecological status classes derived from their application to an extensive metabarcoding dataset from United Kingdom rivers; this comprised 1703 samples and was produced using the 331-bp primers. In addition, we assess the effectiveness of each barcode for discrimination of genetic variants around and below the species level. The strong correlation found in IPS values between barcodes indicates that the choice of the barcode does not have major implications for current WFD ecological assessments, although a very few sites were downgraded from WFD acceptable classes ("good" / "high") to unacceptable classes ("moderate" / "poor" / "bad"). Analyses of the taxonomic resolution of the two barcodes indicate that for many ASVs, the use of either marker – 263-bp and 331-bp – gives unambiguous assignations at the species level though with differences in bootstrap confidence values. Such differences are caused by the stochasticity involved in the naïve Bayesian classifier used and by the fact that genetic distance, regarding closely related species, is increased when using the 331-bp barcode. However, in some specific cases, species differentiation fails with the shorter marker, leading to underestimates of species diversity. For example, use of the longer barcode allows classification of three ASVs as *Surirella brebisonii*, *Halamphora montana* and *Fragiliara agnesiae*, whereas each of these ASVs are identical to several different species with the 263-bp marker, some of them associated with different ecological preferences. Use of the shorter marker can sometimes lead to false positives when the extent and nature of infraspecific variation are poorly known.

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

1. Introduction

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

In this context, two similar barcodes of the *rbcL* gene have been developed independently by different research groups for diatom metabarcoding. One of these barcodes covers a region of 263-bp and is amplified by the primer pair Diat_rbcL_708F (Stoof-Leichsenring et al., 2012) and R3 (Bruder & Medlin 2007). These primers were further degenerated by Vasselon et al. (2017), in order to cover a wider diversity of diatoms, resulting in three forward primers (Diat_rbcL_708F1, Diat_rbcL_708F2 and Diat_rbcL_708F3) and two reverse primers (R3_1 and R3_2). The second barcode includes the same 263-bp region as the previous one but has an extra tail of 68-bp located at the 5′ end. This latter, developed by Kelly et al. (2018, 2020), therefore comprises 331 bp and is amplified by the primer pair rbcL-646F and rbcL-998R. Thus, although both barcodes overlap in the shared region of 263-bp, they could potentially differ in their ability to discriminate between species, which would be relevant for biodiversity analyses but also for the monitoring and management of freshwater rivers covered by the Water Framework Directive (WFD), since the diatom indices computed

for such purposes, such as the Indice de Polluosensibilité Spécifique (IPS; Cemagref, 1982), rely on species composition and relative abundance. Both barcodes (hereafter referred to as the 263- and 331-bp markers) have been demonstrated to be effective for biomonitoring and diversity analyses (e.g. Kang et al., 2021; Kelly et al., 2018, 2020; Rimet et al., 2018b; Rivera et al., 2020). Nevertheless, we might hypothesize that the 68-bp tail might confer an advantage for species assignment in two ways. On the one hand, it might be possible that related species are identical in the 263-bp shared region but differ at variable sites in the extra 5′ tail. On the other hand, the accuracy of some automated methods commonly applied for classifying metabarcoding data increases as the length of the query sequence increases (Porter et al., 2014; Karim & Abid, 2021). In this regard, it might be expected that the longer 331-bp barcode could increase the effectiveness of the Naïve Bayesian classifier (Wang et al. 2007), a Kmer-based method that is one of the most commonly implemented classifiers for assigning reads to named taxa in metabarcoding studies.

These two aspects have not yet, to our knowledge, been explored for the two similar diatom *rbcL* markers. Therefore, this study aimed to (1) compare the effect of choosing one or the other marker on WFD ecological assessments through the comparison of IPS scores: is there any significant advantage in using the longer marker? (2) assess the effectiveness of the two markers for discriminating genetic variants at or below the species level. For achieving these aims, we used a large dataset of environmental samples collected during several biomonitoring campaigns in UK rivers (Kelly et al. 2018, 2020).

2. Material and Methods

2.1 Dataset and bioinformatics analyses

The dataset used in this study comprised 1703 benthic diatom samples that were originally taken as part of routine WFD biomonitoring programmes of UK rivers held in 2014, 2016 and 2017 (Kelly et al., 2018, 2020). High-throughput sequencing (HTS) of these samples was based on the 331-bp *rbcL* marker amplified by the rbcL-646F and rbcL-998R primers, and we were supplied with the fastq files from MiSeq output. Further details about the preparation of samples for HTS are described in Kelly et al. (2018, 2020). We conducted bioinformatics analyses on the forward (R1) and reverse (R2) reads to generate the Amplicon Sequence Variants (ASVs) that constituted the fundamental units on which further examinations were carried out. ASVs were generated using the R package *DADA2* (Callahan et al., 2016) and the different runs (a total of 10)

were analysed separately. The rbcL-646F and rbcL-998R primers were removed from R1 and R2 reads using cutadapt (Martin, 2011). Then, the R1 and R2 reads were truncated to 220–240 and 160–180 nucleotides respectively, based on their quality profiles (median quality score < 30), and those reads with ambiguities or showing an expected error (maxEE) higher than 2 were removed. The DADA2 denoising algorithm was then applied to determine an error rates model in order to infer amplicon sequence variants (ASVs). Finally, ASVs detected as chimeras were discarded using the DADA2 function "removeBimeraDenovo". Since the ASVs generated were based on the 331-bp *rbcL* marker, they also contained the 263-bp region targeted by the three forward primers Diat_rbcL_708F1, Diat_rbcL_708F2 and Diat_rbcL_708F3 and the two reverse primers R3_1 and R3_2. To avoid any incongruence during the comparative analyses of the two markers, the only ASVs selected for further analyses were those in which the forward primers Diat_rbcL_708F1, Diat_rbcL_708F2 or Diat_rbcL_708F3 were also identified. For this, cutadapt was applied again, this time on the 331-bp ASVs already generated, to unambiguously identify and remove these primers specifically designed for the 263-bp marker. Thus, two datasets with the same number of ASVs were finally generated, one containing ASVs with a total length of 331-bp (i.e. those based on the rbcL-646F and rbcL-998R primers) and a second one including the same ASVs but truncated to a length of 263-bp.

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

2.2 Reference library preparation and taxonomic assignment.

A custom-made reference library composed of 331-bp sequences was used for performing the taxonomic assignment of the ASVs generated. By controlling the reference sequence length (rather than using reference sequences that have not been trimmed to the same length), it is easier to evaluate how the different marker lengths are affecting the taxonomic assignment. The custom-made library consisted of all the sequences from the curated diatom reference library Diat.barcode v10 (Rimet et al., 2019) that cover the full 331-bp *rbcL* marker. It was created by extracting a small subset of diatom *rbcL* sequences from Diat.barcode v10 that covered the 331-bp marker, aligning them (using MUSCLE: Edgar, 2004), and truncating them to the target 331-bp region using MegaX (Kumar et al., 2018). Then, all the remaining *rbcL* diatom sequences included in Diat.barcode v10 were extracted and aligned against the aligned subset using the *align.seqs* function implemented in Mothur software (Schloss et al., 2009), with default parameters. The resulting alignment of 331-bp diatom sequences was further filtered with Mothur (using the *screen. seqs* function*)* to keep only sequences without ambiguities. The taxonomic assignment of 263-bp and 331-bp ASVs was performed using two methods: 1) the naïve Bayesian classifier method (Wang et al., 2007) using the "assignTaxonomy" function from *DADA2* and 2) the Basic Local Alignment Search Tool (BLAST). Prior to the next analyses, and in order to remove non-diatom variants that likely occurred in our dataset, only ASVs classified into Bacillariophyta and receiving 100% bootstrap support by the Bayesian classifier were kept for downstream analyses. As a result, a total of 2933 ASVs were used in this study.

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

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

The ecological status of each sample was determined by applying the IPS diatom index, since this is adopted in many EU countries for WFD bioassessment of rivers. For each sample, the IPS was calculated twice, one using the species inventory derived from the 263-bp ASVs, and the other using the inventory from the 331-bp ASVs. IPSS and IPSV

values for each species were extracted from OMNIDIA software v5.5 (Lecointe et al., 1993). Comparisons of the IPS values were performed using ASVs that had a species assignment bootstrap value ≥ 85%, since thresholds from 80% to 85% are commonly applied for diatom biomonitoring assessments (e.g. Rivera et al., 2020; Mortágua et al., 2019; Vasselon et al., 2017). The WFD ecological status class for each sample was assigned by applying the following boundaries (Afnor, 2007): High (17 ≤ IPS ≤ 20), Good (13 ≤ IPS < 17), Moderate (9 ≤ IPS < 13), Poor (5 ≤ IPS < 9), Bad (1 ≤ IPS < 5).

2.4 In-depth analyses on species discrepancies

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

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

In order to compare and illustrate the nucleotide and amino-acid variability of the extra 68-bp region provided by the 331-bp marker, Shannon's entropy values were calculated from both the reference sequences from the 331-bp custom reference library and the 331-bp ASVs obtained. Before calculating Shannon entropy values on ASVs, several filter steps were applied in order to remove likely artefacts. For this, only ASVs with 331-bp length were kept and those showing an abundance lower than 10 reads and/or occurring in only 1 sample were also removed. The resulting ASVs were aligned against the custom 331-bp reference library and those with gaps and/or stop codons were further discarded. In addition, duplicated sequences from the custom reference library (i.e.

sharing the 331-bp marker) were removed. Shannon entropy was thus calculated on a total of 2617 ASVs and 1886 reference sequences. Entropy values were computed using the "MolecularEntropy" function implemented in the R package *HDMD* (McFerrin, 2013) and the values were standardized to 4 and 20 for nucleotides and amino acids respectively.

3. Results

3.1 Effects of the marker on taxonomic assignment

The number of ASVs assigned at the species level by the naïve bayesian classifier was always higher when using the longer marker, regardless of the bootstrap confidence threshold applied (Table 1). On the other hand, BLAST analyses indicated that for the 263-bp marker, a total of 536 different ASVs (18.3%) had at least one identical match (identical matches considered only when query ASV sequences were fully covered) with reference sequences included in Diat.barcode while this number was reduced to 426 ASVs (14.5%) when considering the full 331-bp marker. In addition, 29 ASVs based on the 331-bp marker were identical to reference sequences from more than 1 species and these ambiguous assignments corresponded to a total of 62 different species but to a total of 74 species when considering only the 263-bp marker (Supplementary Table 1). These ambiguous assignations at the species level were exemplified, among others, in some ASVs classified into the genera *Fragilaria* (ASVs 59, 131 and 346; Fig. 4), *Iconella* (ASVs 270 and 361), *Surirella* (ASV 26; Fig. 3) and *Gomphonema* (ASVs 6, 148, 216, 274 and 610) (Supplementary Table 1).

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

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

3.2 Effects of the marker choice on ecological status assessment

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

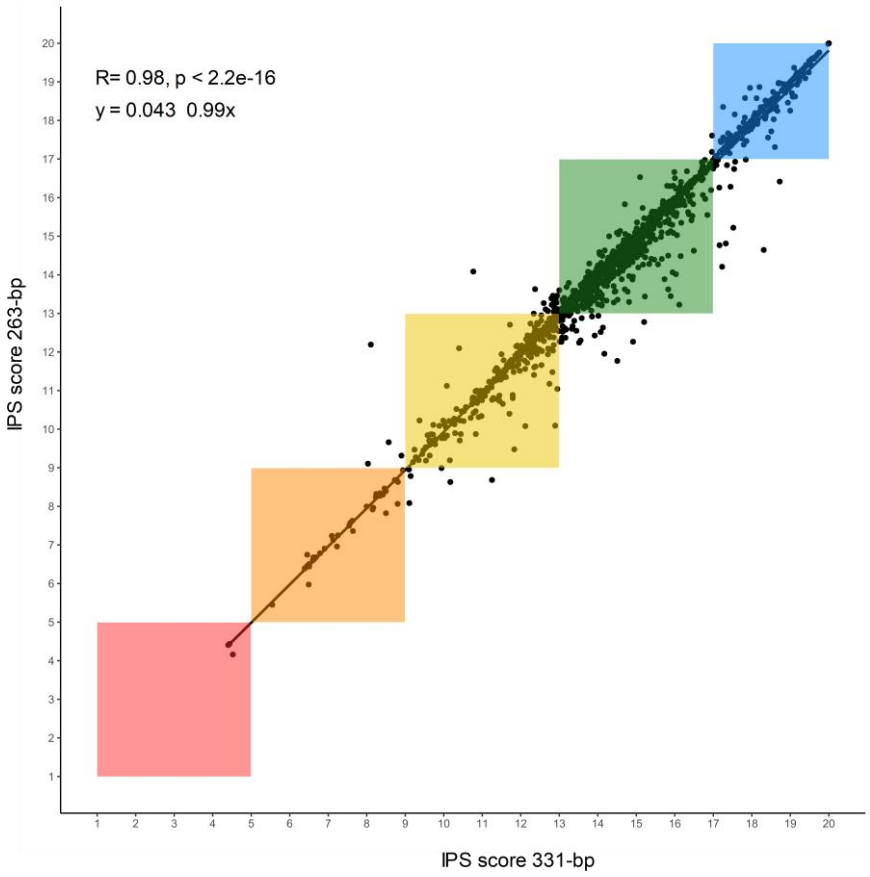


Fig. 1 Correlation of IPS values derived from 263-bp and 331-bp markers considering the total 1703 samples analyzed. Pearson's coefficient (R) and p-value are given. Coloured squares represent boundaries for the different WFD ecological status classes: blue=high (17 ≤ IPS ≤ 20); green=good (13 ≤ IPS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

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

The species showing the greatest dissimilarities in relative abundance between markers are listed in Fig. 2.

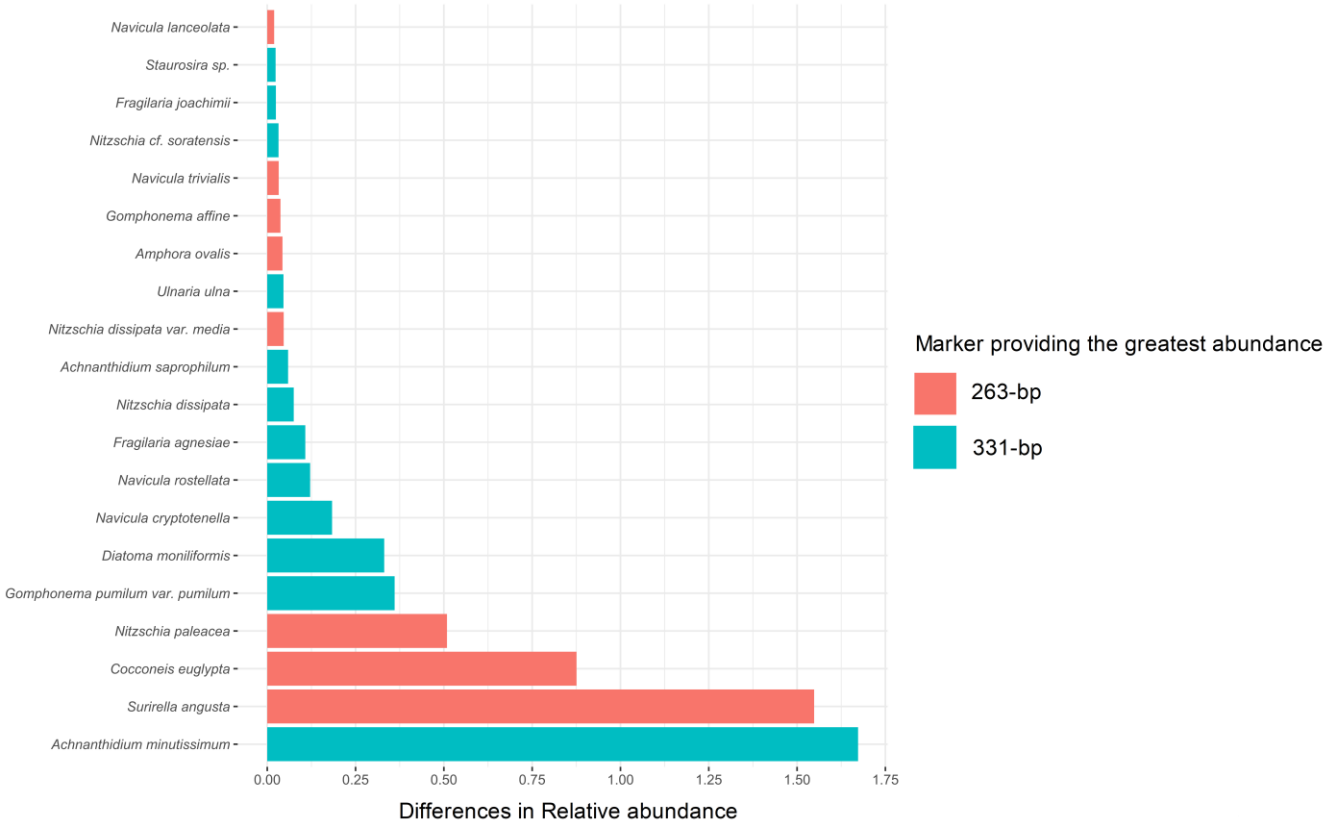


Fig. 2. Top 15 species showing the greatest differences in relative abundance between 263-bp and 331-bp markers considering the total 1703 samples analyzed. Bars in red and blue represent species for which the greatest relative abundance was provided by the 263-bp and 331-bp respectively.

Examination of bootstrap support values and BLAST outputs for both 263-bp and 331-bp ASVs of these species revealed there are three main reasons for the abundance dissimilarities:

i) False negatives: Some ASVs were classified into the same species by both the 263-bp and 331-bp markers but the identifications could be rejected for one or other marker because bootstrap support values did not reach the confidence threshold (i.e. bootstrap values ≥ 85), ultimately causing differences between markers in species' relative abundance. Some false negatives arose when the assignments of 263-bp ASVs received much lower bootstrap support values than their 331-bp counterparts. This occurred when the genetic distance between ASVs and closely related reference sequences (as measured by the number of base-pair mismatches between ASVs and reference sequences reported by BLAST analyses) decreased when using the shorter marker compared to the longer one. In this regard, the most important cases were detected in ASVs from the *Achnanthidium*

minutissimum complex (observed in ASVs closely related to *A. jackii* and *A. pyrenaicum*, such as ASV909, ASV1420, ASV7083), *Nitzchia perminuta* (detected in ASVs assigned to this species but similar also to *N. acidoclinata*, for instance, ASV2288), *Encyonema ventricosum* (ASVs also similar to *E. minutum*, such as ASV929), *Diatoma moniliformis* (ASVs also similar to *D. tenuis*, e.g. ASV73, ASV403 and ASV1159) or *Navicula rostellata* (ASV200 and ASV721, two ASVs similar to reference sequences classified as *Navicula* sp. and *Haslea howeana*) (Supplementary Data 1 & 2). By contrast, other false negatives were detected without being recorded an increase in the genetic distance between ASVs and closely related reference sequences. This was particularly evident in ASV33 and ASV136, two abundant ASVs belonging to *Cocconeis euglypta* and *Gomphonema affine* respectively (Supplementary Data 1 & 2)

ii) Some ASVs were unambiguously classified at the species level based on the 331-bp marker, but not based on the 263-bp marker. This was seen in ASVs in *Surirella* (ASV17), *Fragilaria* (ASV140) and *Halamphora* (ASV1784). Within *Surirella*, ASV17 had identical matches with reference sequences from *Surirella brebissonii* (including *S. brebissonii* var. *kuetzingii*) when the ASV was based on the 331-bp marker and could therefore be identified unambiguously. The effect of reducing the barcode marker to the 263-bp region was to make ASV17 identical to reference sequences belonging to 10 different taxa (i.e. *S. angusta, Surirella* sp*., S. cf. pinnata, S. brightwellii, S. ovalis var. apiculata, S. cf. minuta, S. minuta,* and *S. lacrimula*, as well as the two that are identical over the whole of the 331-bp marker, *S. brebissonii* and *S. brebissonii var. kuetzingii*). A haplotype network for these and other *Surirella* species and related ASVs is given in Fig. 3 and shows the changes in assignment and relationships when the marker length is reduced from 331 bp (Fig. 3a) to 263 bp (Fig. 3b). In the case of *Fragilaria* species, ASV140 matched only one species (*Fragilaria agnesiae*) based on the 331-bp marker (Fig. 4a), but was identical to three species, *F. agnesiae*, *Fragilaria* sp. and *Fragilaria* cf. *nanoides*, with the 263-bp marker (Fig. 4b). A third case (not graphed) was ASV1784, which shared the full 263-bp marker with reference sequences from *Halamphora montana* and *H. banzuensis* species but differed from the latter by two mutations located at the 30th and 34th positions of the 331-bp marker.

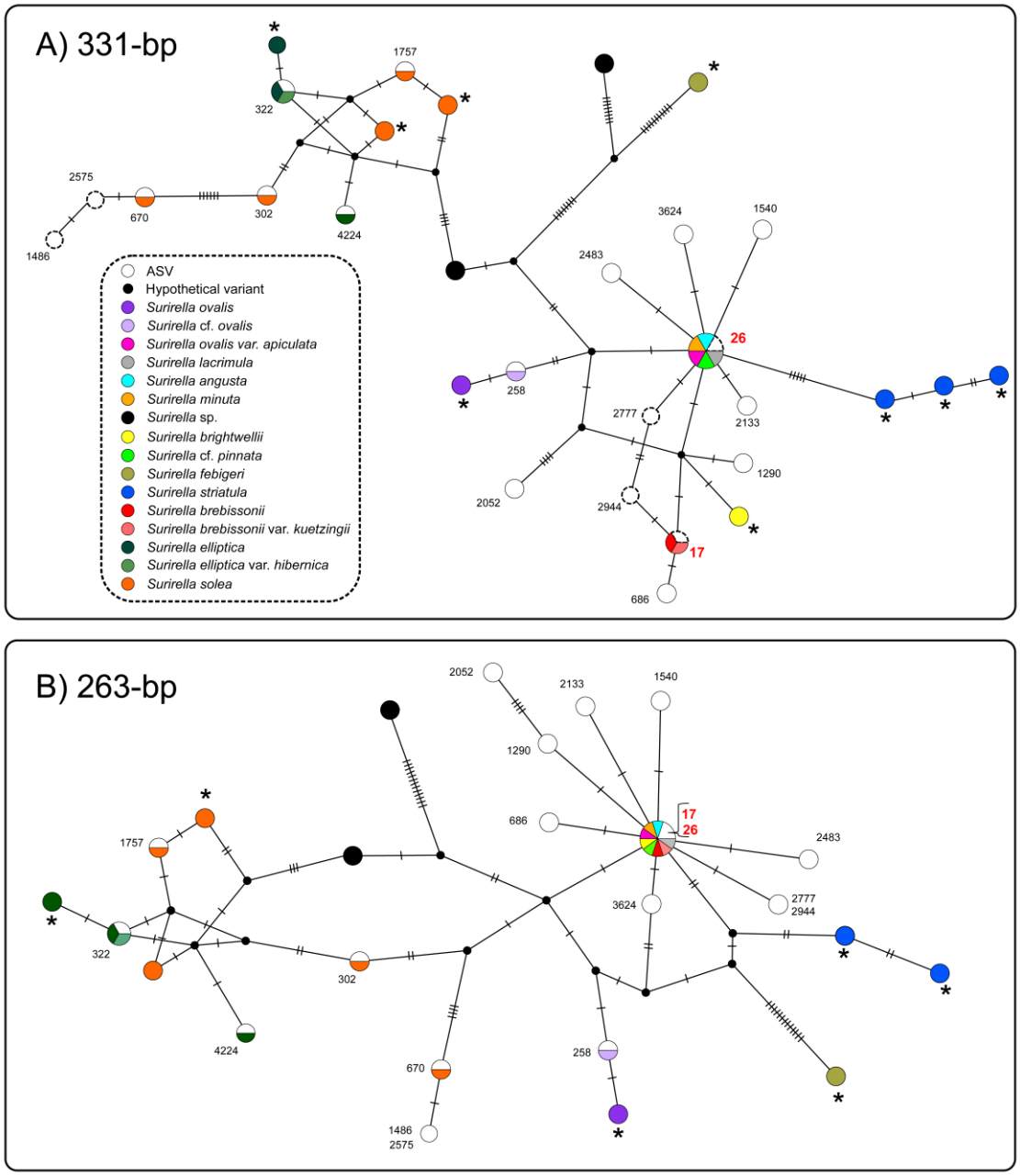


Fig. 3. TCS haplotype networks of *Surirella* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) *rbcL* markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition and share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

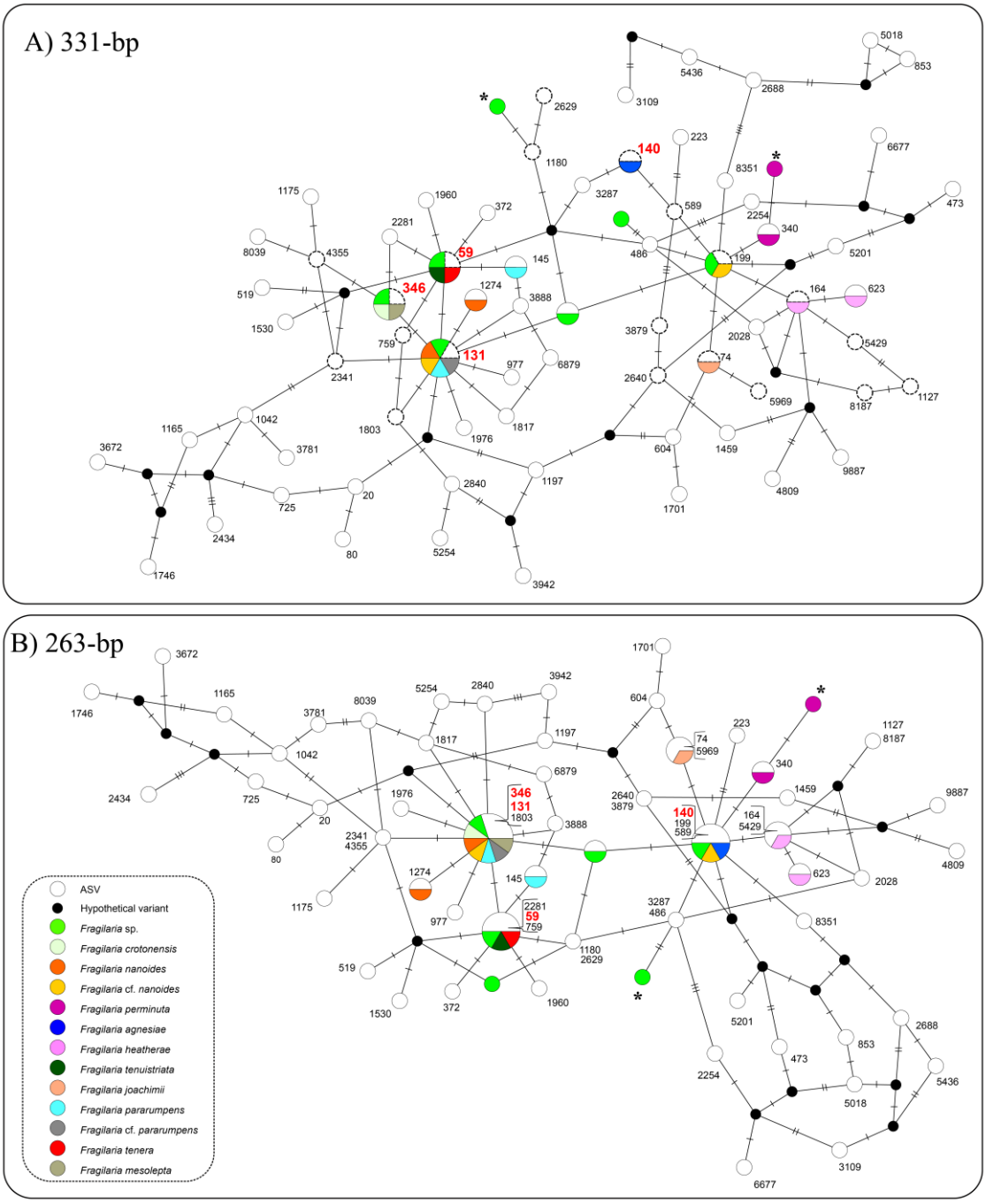


Fig. 4. TCS haplotype networks of several *Fragilaria* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) *rbcL* markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition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

iii) A third group comprised ASVs that could not be identified to species with either marker: they were identical to reference sequences from more than one taxon for both the 263-and the 331-bp marker. In these cases, differences in species' relative abundance between markers occurred when the taxonomic classification provided by one marker did not reach the selected confident threshold (i.e. bootstrap values ≥ 85) but this threshold was reached when using the other marker. This pattern is likely associated with the random component of the naïve Bayesian classifier and it was observed in ASVs classified into the genera and *Achnanthidium* (ASV12) and *Iconella* (ASV 361) (Supplementary Data 3).

A more complex and particularly instructive case illustrating the potential complexities of interpretating the metabarcoding data, is given by *Nitzschia* ASVs 1690 and 3022. These two haplotypes shared the full 263-bp marker with reference sequences from *Nitzschia dissipata* var*. media* and *N. heufleriana*, respectively, and therefore seemed securely identified, ASV 3022 as *N. dissipata var. media* and ASV 1690 as *N. heufleriana* (Fig. 5b). However, when considering the full 331-bp marker these ASVs were not identical to the same two reference sequences and had no exact match in the reference dataset. Instead, each of them differed by 1 nucleotide from both *N. dissipata* var*. media* and *N. heufleriana*, making identification impossible at species level (Fig 5a).

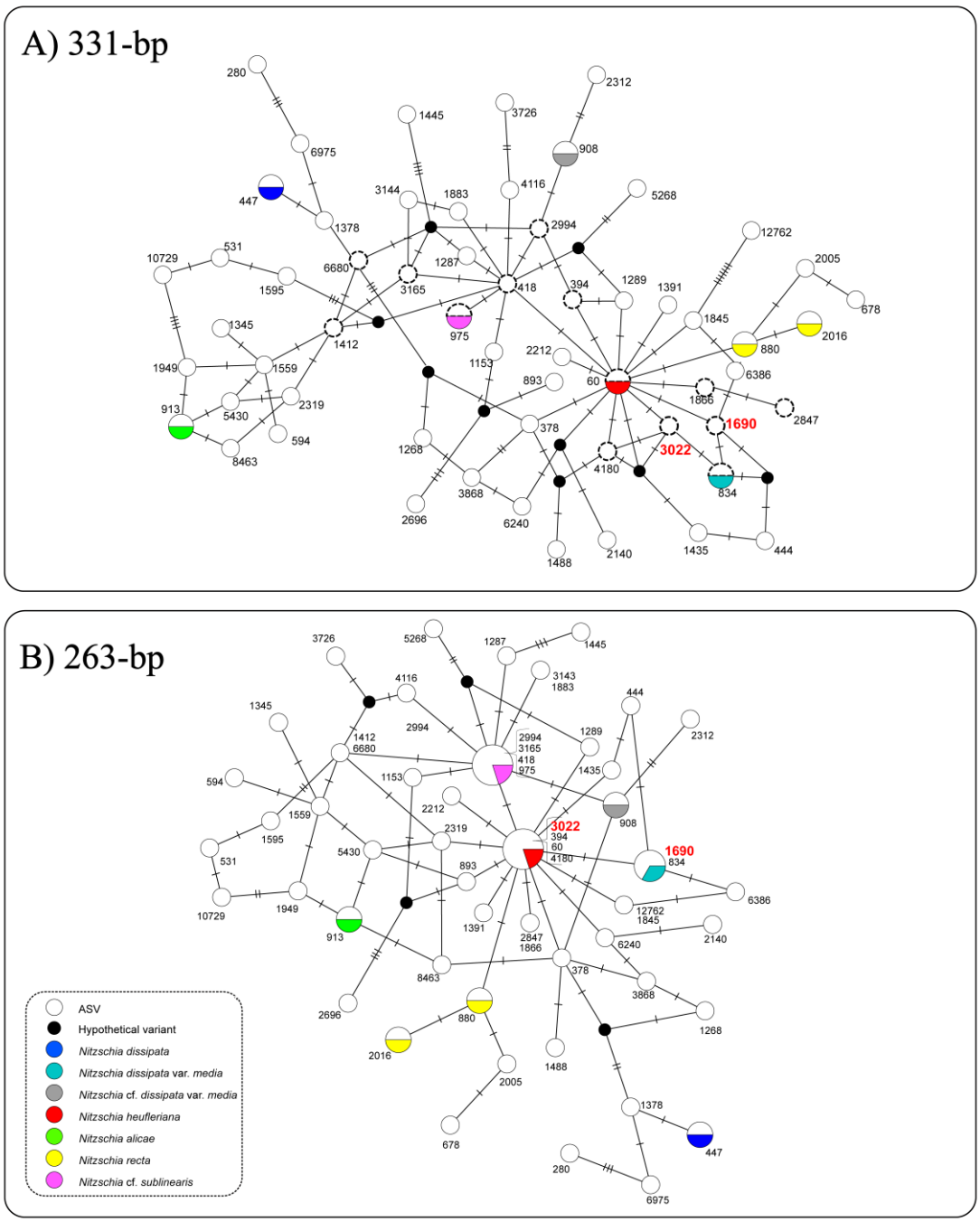


Fig. 5. TCS haplotype networks of several *Nitzchia* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) *rbcL* markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition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

3.4 Nucleotide and amino-acid variability.

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

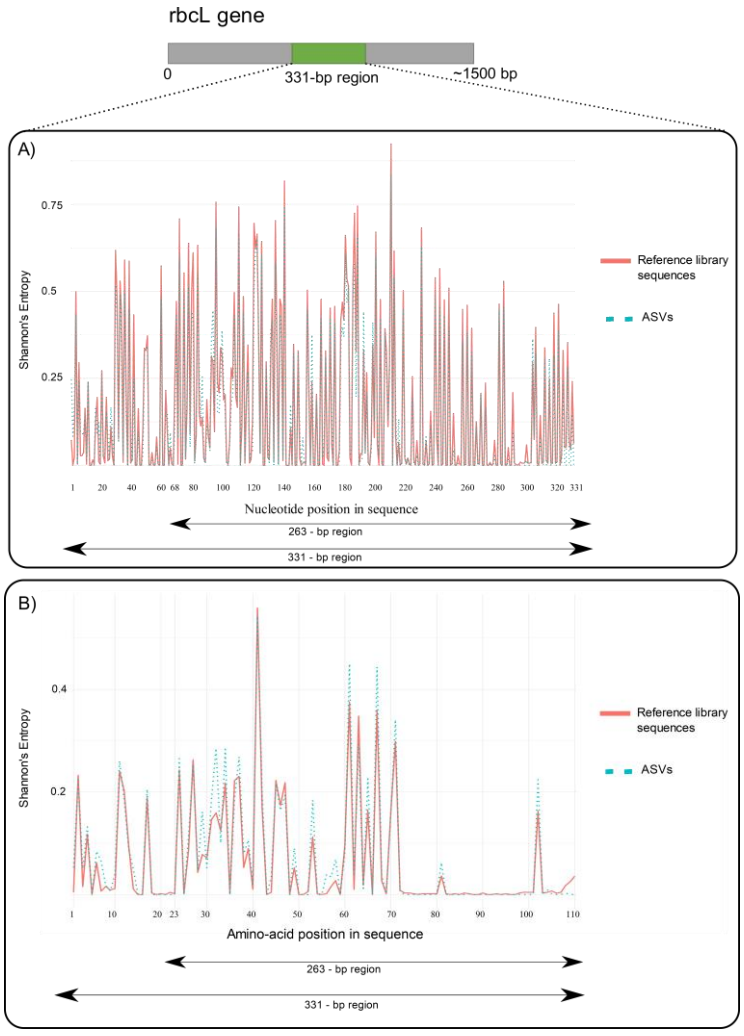


Fig. 6 Shannon's entropy per nucleotide (figure a) and amino-acid (figure b) position obtained for 1886 reference sequences of 331-bp from Diat.barcode v10 (represented by a red line) and a total of 2617 ASVs obtained in this study (represented by a blue dashed line). ASVs included for computing entropy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

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

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

4. Discussion

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

The extra length of the 331-bp marker means that it inevitably provides more information on genetic diversity, given the variability of the extra 68-bp tail (Fig. 6). Our results indicate, however, that the choice of the 263-bp or 331-bp *rbcL* marker has no important effects on WFD ecological status assessments, since IPS scores derived from both markers were very highly correlated (i.e. Pearson's R = 0.98 and intercept close to 0) and the vast majority of sites were classified into the same ecological status class regardless of the marker used (i.e. 95.2%). In addition, out of the sites that differed in the ecological status assignment, most of them correspond to absolute deviations in the IPS scores of < 1. However, the overall number of sites classified into "Moderate", "Poor" and "Bad" status differed with the marker chosen, and this number was higher when using the 263-bp one. As a consequence, some particular sites were assigned to the "Good" or "High" ecological status when using one marker, but they were assigned instead to the "Moderate", "Poor" or "Bad" status when using the other (observed in a total of 55 out of 1073 samples studied). Though the proportion of such samples is very low, they should not be overlooked since the WFD demands remedial actions for those aquatic systems that fail to reach at least the "good" ecological status.

At first, it might be interpreted that the discrepancies in IPS values for those sites that alter their ecological status from acceptable (i.e. "Good"/"High") to unacceptable ("Bad"/"Poor"/"Moderate") classes are brought about by differences in species' relative abundances caused by the higher taxonomic resolution of the 331-bp marker (i.e. the 331-bp marker can unambiguously classify some ASVs at the species level that 263-bp marker cannot). However, our results indicated that the choice of the marker was

decisive for discriminating taxa at species level in only three ASVs (discussed further in section 4.2) and more importantly, these ASVs were scarcely represented in most of the samples: only ASV17 (*Surirella brebissonii*) contributed at least 10% of reads' relative abundance in 7 samples (supplementary Data 4). Thus, most of the discrepancies observed between markers in species' relative abundance, and hence in WFD ecological status assignations, cannot be attributed to differences in taxonomic resolution between markers. Instead they are likely due to other factors such as the stochasticity involved in the Bayesian classifier (Wang et al., 2007) and false negatives. In this regard, our results showed that the use of the extra 68-bp region can reduce the number of false negatives by increasing the genetic distance between ASVs and closely related taxa and therefore if initiating a new metabarcoding study, the 331-bp marker could be preferable.

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

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

Other cases where the 331-bp marker is decisive for species identification include *Halamphora montana* vs *H. banzuensis* (ASV1784), two species with very different

habitat requirements. *H. montana* occurs in intermittently wet terrestrial microhabitats and eutrophic freshwaters (Lange-Bertalot et al., 2017) and is characterized by intermediate IPS sensitivity values (IPSS=2.9). In contrast, *H. banzuensis* is a marine species (recently described by Stepanek & Kociolek, 2018) and hence has no associated IPS indicator values. The little variation found between both 263-bp and 331-bp *rbcL* markers for these species is not exceptional within *Halamphora*, as other examples of close phylogenetic relationships between freshwater and marine species can be found within the genus (Stepanek & Kociolek, 2019). Similarly, *F. agnesiae* (ASV140) cannot be identified using the 263-bp marker, but in this case the effects are unclear: *F. agnesiae* is a recently described species without a full ecological characterization (Kahlert et al., 2019).

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

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

Overall, the 331-bp marker is superior in that the diversity that can be detected is greater and the proportion of ambiguous identifications is lower. Sometimes too, an apparently straightforward identification with the shorter marker is deceptive. Particularly instructive in this regard is the example of *Nitzschia* ASVs 1690 and 3022, which seem to be identifiable confidently and unambiguously with the 263-bp marker (100% matches with *N. dissipata* var. *media* and *N. heufleriana* reference sequences, respectively) but not with the 331-bp marker: the two ASVs cannot be identified from the 331-bp versions since they are not identical to either of the reference sequences that are available but separated from each of them by the same genetic distance. In this case, to interpret the metabarcoding datasets fully in terms of nominal species and varieties, much more information would be needed about the correspondence between *rbcL* variation and morphology.

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

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

Most of the genetic variants examined were not represented in the reference library: out of the 2933 ASVs separated by the 331-bp marker, identical matches with reference sequences were found for only 426 (14.5%) and 536 ASVs (18.3%) respectively for the 331- and 263-bp markers. To some extent, this is because of the lack of reference sequences for many nominal species, but it also reflects the high intraspecific diversity that characterizes diatom species, at least as these are currently circumscribed (e.g. Amato et al., 2007; Perez-Burillo et al. 2021; Pinseel et al., 2017; Souffreau et al., 2013). The question that arises is whether the intraspecific diversity detected by the two *rbcL* markers is only 'genetic noise', or whether it contains information on ecological or biogeographical differentiation and therefore needs to be recorded and analysed. First indications are that, while closely related species often share a similar ecology (Keck et al., 2018), closely related ASVs can differ in ecological preferences and distribution (Pérez-Burillo et al., 2021). Therefore, while it will always be important to relate the ASVs of metabarcoding datasets to formal morphology-based taxonomy – e.g. to ensure continuity with previous studies and allow cross-talk with fields where DNA-based approaches are limited in their application (e.g. stratigraphical or palaeoecological studies) – degrading analysis to the level of nominal species is suboptimal. For example, from a biomonitoring perspective it will mean that diatom indexes are being computed

using only a part of the information from the total captured, especially when strict confidence thresholds are applied. In particular, we found that around 70% of the ASVs were not assigned to a species by the naïve Bayesian classifier when the confidence threshold was ≥ 99%. Hence an attractive alternative to the present approach, if environmental data are available for an extensive set of metabarcoded samples, is a direct calibration of the environmental preferences of ASVs or OTUs, as suggested by other studies (e.g. Apothéloz-Perret-Gentil et al., 2017; Feio et al., 2020; Smucker et al., 2020; Tapolczai et al., 2019). Microscopy-based approaches remain important, however, since they give opportunities to study traits that are not or only partially taxon-related, such as life-history stage or, in the case of some marine diatoms, existence as endosymbionts (Pérez-Burillo et al., 2022).

Conclusions

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

Acknowledgements

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

References

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

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

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., Pawlowski, J., 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Mol. Ecol. Resour. 17, 1231-1242. https://doi.org/10.1111/1755-0998.12668.

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

Bailet, B., Apothéloz-Perret-Gentil, L., Baricevic, A., Chonova, T., Franc, A., Frigerio, J.-M., Kelly, M., Mora, D., Pfannkuchen, M., Proft, S., Ramon, M., Vasselon, V., Zimmermann, J., Kahlert, M., 2020. Diatom DNA metabarcoding for ecological assessment: comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. Sci. Total Environ. 745, 140948. https://doi.org/10. 1016/j.scitotenv.2020.140948.

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

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

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

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

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

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

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

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

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

Kahlert, M., Kelly, M.G., Mann, D.G., Rimet, F., Sato, S., Bouchez, A., Keck, F., 2019. Connecting the morphological and molecular species concepts to facilitate species identification within the genus *Fragilaria* (Bacillariophyta). J. Phycol. 55, 948-970. https://doi.org/10.1111/jpy.12886

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría Galindo, P., Vences, M., Wang, J., Schwalba, A., 2021. Diatom Metabarcoding and Microscopic Analyses from Sediment Samples at Lake Nam Co, Tibet: The Effect of Sample-Size and Bioinformatics on the Identified Communities. Ecol. Indic. 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070.

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

Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18, 1299-1309. https://doi.org/10.1111/1755-0998.12919.

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

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

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

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

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

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

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

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

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

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

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

Pérez-Burillo, J., Trobajo, R., Leira, M., Keck, F., Rimet, F., Sigró, J., Mann, D.G., 2021. DNA metabarcoding reveals differences in distribution patterns and ecological preferences among

genetic variants within some key freshwater diatom species. Sci. Total Environ. 728, 149029 https://doi.org/10.1016/j.scitotenv.2021.149029

Pérez-Burillo, J., Valoti, G., Witkowski, A., Prado, P., Mann, D. G., Trobajo, R., 2022. Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters. Mar. Pollut. Bull. 174, 113183. https://doi.org/10.1016/j.marpolbul.2021.113183.

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. Eur. J. Phycol. 52, 264-280. https://doi.org/10.1080/09670262.2017.1283540.

Porter, T.M, Gibson, J.F., Shokralla, S., Baird, D.J., Golding, G.B., Hajibabaei, M., 2014. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome coxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. Mol. Ecol. Resour. 14,929-942. https://doi.org/10.1111/1755-0998.12240.

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

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

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

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

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

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

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

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

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

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

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

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

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., Vasselon, V., 2019. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. Front. Ecol. Evol. 7, 407 https://doi.org/10.3389/fevo.2019.00409.

Urbánková, P., Veselá, J., 2013. DNA-barcoding: A case study in the diatom genus *Frustulia* (Bacillariophyceae). Nova Hedwigia. 142, 147-162.

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

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. Appl. Environ. Microbiol. 73, 5261-5267. https://doi.org/10.1128/AEM.00062-07.

Supplementary material

Supplementary Table 1. List of the 29 ASVs that shared the full 331-bp region with reference sequences from more than 1 species. Corresponding species matching the 263-bp and 331-bp markers are shown.

ASV	Identical match for the 331-bp marker	Identical match for the 263-bp marker
ASV1064	*Stephanodiscus neoastraea* *Stephanodiscus agassizensis*	*Stephanodiscus neoastraea* *Stephanodiscus agassizensis*
ASV12	*Achnanthidium pyrenaicum* *Achnanthidium minutissimum*	*Achnanthidium pyrenaicum* *Achnanthidium minutissimum*
ASV1277	*Pinnularia gibba* *Pinnularia microstauron* *Pinnularia parvulissima*	*Pinnularia gibba* *Pinnularia microstauron* *Pinnularia parvulissima*
ASV131	*Fragilaria nanoides* *Fragilaria pararumpens*	*Fragilaria pararumpens* *Fragilaria mesolepta* *Fragilaria nanoides* *Centronella reicheltii* *Fragilaria crotonensis*
ASV140	*Fragilaria taeniavaucheriae* *Fragilaria agnesiae*	*Fragilaria taeniavaucheriae* *Fragilaria agnesiae*
ASV148	*Gomphonema truncatum* *Gomphonema capitatum*	*Gomphonema truncatum* *Gomphonema capitatum*
ASV1518	*Discostella pseudostelligera* *Discostella woltereckii*	*Discostella pseudostelligera* *Discostella woltereckii*
ASV168	*Thalassiosira pseudonana* *Thalassiosira delicatula*	*Thalassiosira pseudonana* *Thalassiosira delicatula*
ASV1769	*Stauroneis heinii* *Stauroneis gracilis*	*Stauroneis heinii* *Stauroneis gracilis*
ASV1783	*Sellaphora capitata* *Sellaphora pupula*	*Sellaphora capitata* *Sellaphora pupula*
ASV197	*Neidium productum* *Neidium dubium*	*Neidium productum* *Neidium dubium*
ASV216	*Gomphonema coronatum* *Gomphonema brebissonii* *Gomphonema clavatum* *Gomphonema acuminatum*	*Gomphonema coronatum* *Gomphonema brebissonii* *Gomphonema clavatum* *Gomphonema acuminatum*
ASV218	*Nitzschia gracilis* *Nitzschia acicularis*	*Nitzschia gracilis* *Nitzschia acicularis*
ASV228	*Achnanthidium jackii* *Achnanthidium minutissimum*	*Achnanthidium jackii* *Achnanthidium minutissimum*
ASV26	*Surirella minuta* *Surirella ovalis* var. *apiculata* *Surirella angusta* *Surirella lacrimula*	*Surirella minuta* *Surirella ovalis* var. *apiculata* *Surirella angusta* *Surirella lacrimula* *Surirella brightwellii* *Surirella brebissonii*
ASV270	*Iconella levanderi* *Iconella spiralis* *Iconella hibernica*	*Iconella levanderi* *Iconella spiralis* *Iconella hibernica* *Iconella linearis* var. *Helvetica*
ASV274	*Gomphonema clavatum*	*Gomphonema clavatum*

	Gomphonema acuminatum	*Gomphonema acuminatum*
ASV3089	*Pinnularia grunowii* *Pinnularia mesolepta*	*Pinnularia grunowii* *Pinnularia mesolepta*
ASV3366	*Pinnularia neglectiformis* *Pinnularia viridiformis*	*Pinnularia neglectiformis* *Pinnularia viridiformis*
ASV346	*Centronella reicheltii* *Fragilaria crotonensis* *Fragilaria mesolepta*	*Centronella reicheltii* *Fragilaria crotonensis* *Fragilaria mesolepta* *Fragilaria nanoides*
ASV557	*Stephanodiscus minutulus* *Stephanodiscus hantzschii* *Stephanodiscus binderanus*	*Stephanodiscus minutulus Stephanodiscus hantzschii Stephanodiscus binderanus*
ASV59	*Fragilaria tenuistriata* *Fragilaria tenera*	*Fragilaria tenuistriata* *Fragilaria tenera*
ASV5909	*Placoneis paraelginensis* *Placoneis elginensis*	*Placoneis paraelginensis* *Placoneis elginensis*
ASV597	*Staurosira construens* *Gedaniella flavovirens*	*Staurosira construens* *Gedaniella flavovirens*
ASV6	*Gomphonema parvulum* *Gomphonema exilissimum*	*Gomphonema parvulum* *Gomphonema exilissimum*
ASV610	*Gomphonema saprophilum* *Gomphonema parvulum*	*Gomphonema saprophilum* *Gomphonema parvulum*
ASV696	*Pinnularia subgibba* *Pinnularia australogibba* var.*subcapitata*	*Pinnularia subgibba* *Pinnularia australogibba* var*. subcapitata*
ASV835	*Brachysira microcephala* *Brachysira neoexilis*	*Brachysira microcephala* *Brachysira neoexilis*
ASV839	*Psammothidium helveticum* *Psammothidium bioretii*	*Psammothidium helveticum Psammothidium bioretii*

Chapter 4

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

Pérez-Burillo, J., Trobajo, R., Leira, M., Keck, F., Rimet, F., Sigró, J., Mann, D.G., 2021.

Sci. Total Environ. 728, 149029
https://doi.org/10.1016/j.scitotenv.2021.149029.

Science of the Total Environment 798 (2021) 149029

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

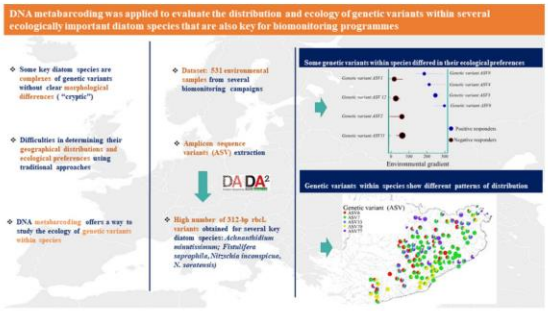
Javier Pérez-Burillo [a,b], Rosa Trobajo [a,*], Manel Leira [c,d], François Keck [e,f], Frédéric Rimet [g,h], Javier Sigró [b], David G. Mann [a,i]

[a] 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
[b] Center for Climate Change (C3), Departament de Geografia, Universitat Rovira i Virgili, C/Joanot Martorell 15, E43500 Vila-seca, Tarragona, Spain
[c] BioCost Research Group, Facultade de Ciencias and Centro de Investigacións Científicas Avanzadas (CICA), Universidade de A Coruña, 15071 A Coruña, Spain
[d] Biodiversity and Applied Botany Research Group, Departmento de Botánica, Facultade de Bioloxía, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
[e] Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[f] Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland
[g] INRAE, UMR Carrtel, 75 av. de Corzent, FR-74203 Thonon les Bains cedex, France
[h] University Savoie Mont-Blanc, UMR CARRTEL, FR-73370 Le Bourget du Lac, France
[i] Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, Scotland, UK

HIGHLIGHTS

- Diatom infraspecific genetic variation is detected via DNA-metabarcoding.
- Genetic variants within species show different patterns of distribution.
- Some genetic variants within species differ in their ecological preferences.
- DNA-metabarcoding facilitates the development of more accurate biological indices.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 April 2021
Received in revised form 16 June 2021
Accepted 9 July 2021
Available online 20 July 2021

Editor: Sergi Sabater

Keywords:
ASV
Environmental DNA
Water framework directive
rbcL

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

* Corresponding author.
E-mail address: rosa.trobajo@irta.cat (R. Trobajo).

https://doi.org/10.1016/j.scitotenv.2021.149029

Ecological preferences
Species distribution

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.

1. Introduction

Diatoms play a crucial role in aquatic systems due, among 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; European Commission, 2016) 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, 2016b, 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., 2018). 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., 2016), 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).

2.2. Physicochemical and biotic parameters

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

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	99.5–13,341.33 (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 (°C)	330	130	200	1 ± 0.00	4.26 ± 6.44	5–26 (12.45 ± 3.17)	7.13–24.5 (18.12 ± 3.99)

diatom indices IBD ("Indice Biologique Diatomées") and IPS ("Indice de Polluosensibilité spécifique") were retrieved respectively for French and Catalan rivers sites analysed.

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

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

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

The influence of DNA extraction methods on the diatom inventory produced by DNA metabarcoding has been evaluated by Vasselon et al. (2017a). In this study, the authors evaluated five different extraction methods, including the two methods used in our study. They found some slight differences in relative abundance between methods for some particular species, but the differences in community composition caused were far less than the differences attributable to habitat. Importantly for the current analyses, the slight differences did not affect species richness ("Regardless of taxonomic level (OTU or species), the taxonomic composition of the community represented in the extracts was not affected by DNA extraction methods…": Vasselon et al., 2017a). Furthermore, (1) the taxa contributing more than >1% of the dissimilarities between diatom communities obtained with the GenElute and MN-Soil protocols did not include either *Achnanthidium* or *Fistulifera* (Vasselon et al., 2017a, Table 3), and (2) all our comparisons were made among ASVs belonging to the same species complex and hence of diatoms with very similar physical characteristics (frustule shape, size and robustness). Thus, we can expect that the different extraction methods used for the French 2016 and 2017 datasets will not have greatly affected either the presence/absence of ASVs (especially since a high threshold of abundance was set for inclusion in the analyses) or relative abundances across the combined dataset. Comparisons across wider ranges of species (e.g. all *Achnanthidium*, all *Navicula*, etc.) might have been more seriously affected.

2.4. Bioinformatic analysis

The sequencing facilities performed the demultiplexing of all the samples providing two fastq files per sample, one corresponding to forward reads (R1) and one to reverse reads (R2). All the demultiplexed Miseq reads were treated together using the R package *DADA2*, following the method described by Callahan et al. (2016). Primers were removed from the R1 and R2 reads using cutadapt (Martin, 2011). The resulting R1 and R2 reads were truncated to 200 and 170 nucleotides respectively, based on to their quality profile (median quality score < 30) and those reads with ambiguities or an expected error (maxEE) higher than 2 were discarded. The DADA2 denoising algorithm was applied to determine an error rates model in order to infer amplicon sequence variants (ASVs); ASVs detected as chimeras were discarded using the function

"removeBimeraDenovo" implemented in DADA2. Finally, the taxonomic affiliation of the ASVs was determined using the database "A ready-to-use database for DADA2: Diat.barcode_rbcL_312bp_DADA2" (Chonova et al., 2020), which is derived from the curated diatom reference library Diat.barcode v9 (Rimet et al., 2019, available at https://www6.inra.fr/carrtel-collection_eng/Barcoding-database and at https://data.inrae.fr/file.xhtml?persistentId=doi:10.15454/TOMBYZ/IEGUXB&version=10.0), and the naïve Bayesian classifier method (Wang et al., 2007); 50% was set as the minimum confidence threshold (the default in DADA2). In this study we focused on ASVs that were assigned by the pipeline to *Nitzschia inconspicua*, *N. soratensis*, *Achnanthidium minutissimum* 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 and Bork, 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 scales 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 change points, which are the values of the environmental gradient where the greatest change in taxon abundance and frequency occurs within the observed samples. TITAN standardizes the magnitude of responses as z scores in order to facilitate cross-taxa comparison. Z scores reflect the type of response, positive (+ z scores) or negative (- z scores), of a particular taxon (ASV in this case) along the environmental gradient and the sum of the z scores (sum z) gives information about the assemblage responses, either negative (sum −z) or positive (sum +z), along the gradient, the maximum z score occurring at the point at which change in assemblage composition is greatest (Baker and King, 2010). We conducted TITAN analyses for each of the environmental parameters and using ASV relative abundance. Number of permutations was set to 250, number of bootstrap replicates used was 500, the minimum number of observations required on each side of a candidate change point was 5 and the TITAN filtering metrics of uncertainty "purity" and "reliability", used to separate reliable responders from stochastic noise along the gradient, were set to 0.95. Z scores obtained for those ASVs whose responses fulfilled purity and reliability criteria for at least 4 environmental variables were hierarchically clustered and visualized through heatmaps in order distinguish groups of ASVs with similar response patterns for environmental data. For that, Euclidean distance and ward. D functions (Ward, 1963) were used to compute dissimilarity distance and hierarchical clustering respectively. On the other hand, Kruskal–Wallis (Hollander and Wolfe, 1973) tests with post hoc Dunn's test (Dunn, 1964) were performed to determine environmental data statistically significant ($p < 0.05$) among the sites where species and ecological groupings occurred. We used the implementation available in the R packages *TITAN2* (Baker et al., 2019), *gplots* (Warnes et al., 2020), *stats* (R Core Team, R, 2020) and *dunn.test* (Dinno, 2017) to conduct the TITAN analyses, heatmaps, Kruskal–Wallis test and Dunn's test respectively.

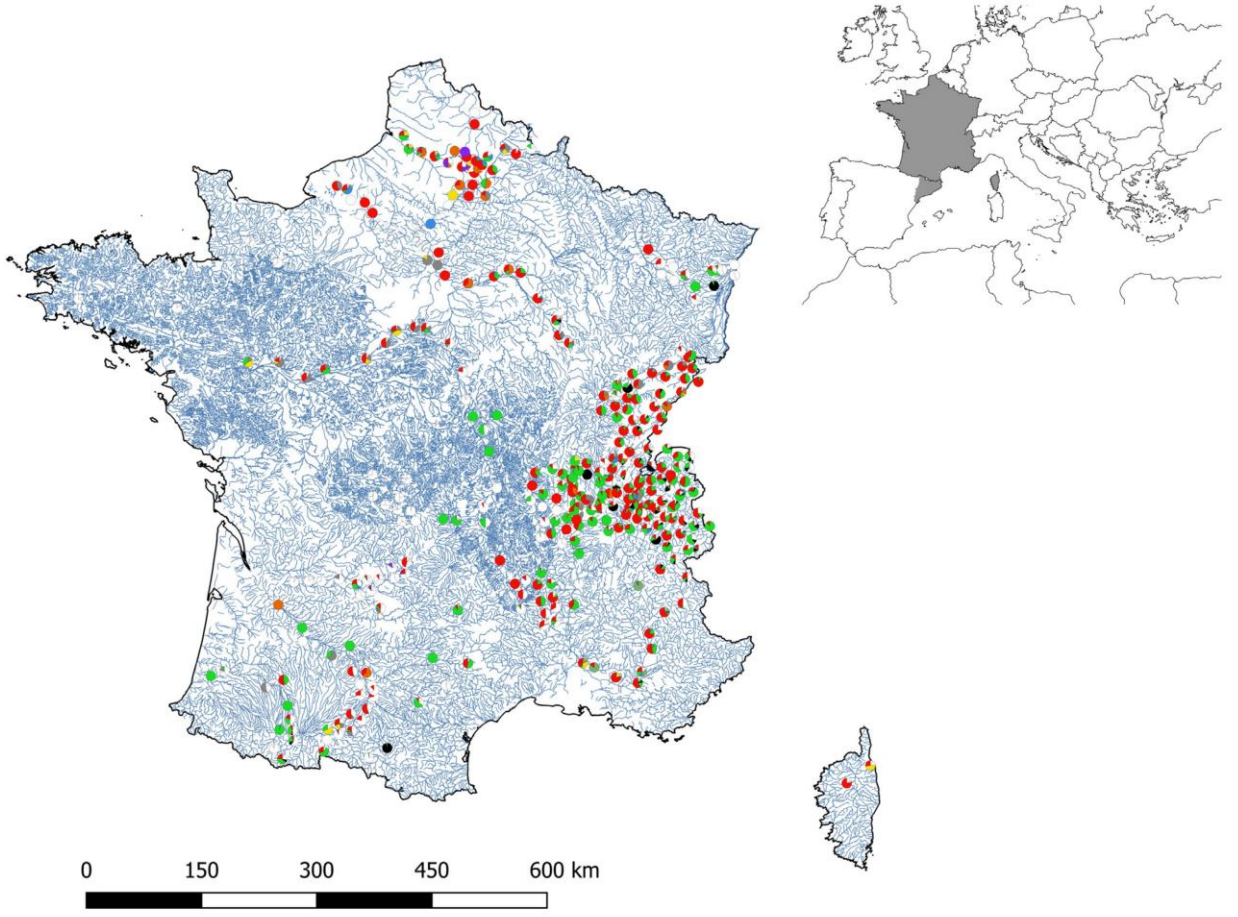
2.6.4. Boosted regression tress

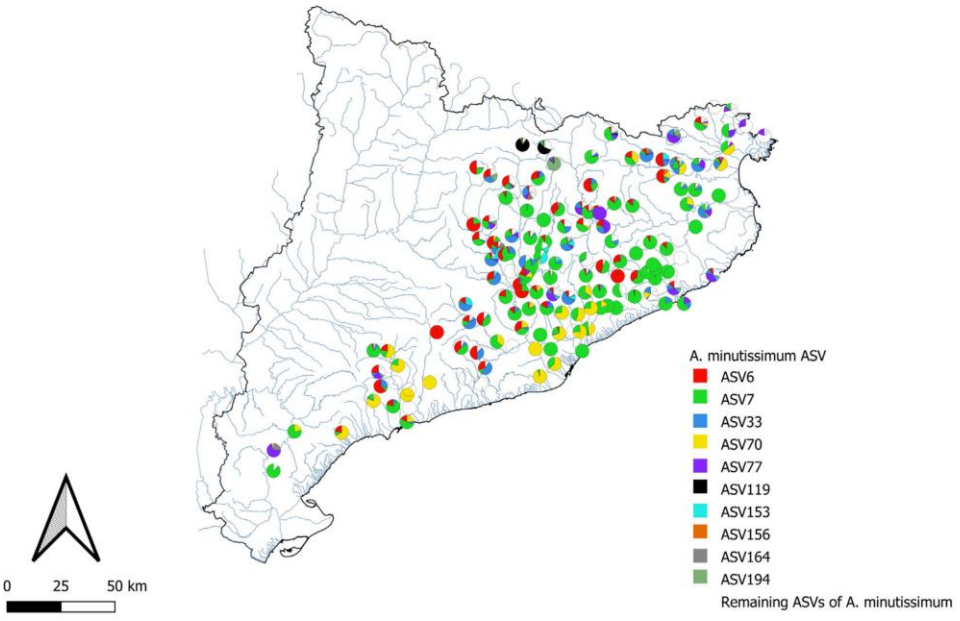
Relationships of the groups of ASVs, defined after TITAN analysis, with environmental variables were additionally evaluated using boosted regression trees (BRT). BRT is a machine learning model that uses a boosting algorithm to combine large numbers of decision trees for improving model accuracy (Elith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

3. Results

3.1. Metabarcoding data

30,251,272 reads were obtained by Miseq Illumina sequencing of a total of 610 samples from Catalan and French rivers. After quality


0
150
300
450
600 km


A. minutissimum ASV
ASV6
ASV7
ASV33
ASV70
ASV77
ASV119
ASV153
ASV156
ASV164
ASV194
Remaining ASVs of A. minutissimum
0
25
50 km

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

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

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

These plots reflected a positive relationship of FSAP EG1 and FSAP EG2 with SO_4^{2-} and a negative one of FSAP EG3 with SO_4^{2-}. This was observed in the increasing response of both FSAP EG1 and EG2 (though intermittently in the former case) along the gradient between 0 and 500 mg/L and in the large decreasing response of FSAP EG3 from 0 to 20 mg/L (Fig. 4). Partial dependence plots also indicated a negative relationship of ASVs from FSAP EG2 with altitude, since the plot depicted a large decrease in the response from 200 to 600 m (Fig. 4). In contrast, the response increased from 0 to 400 m for the FSAP EG1 group, while in the case of FSAP EG3, the response increased from 0 to 600 m and partially and gradually decreased from 700 to 1030 m (Fig. 4). BRT models explained 53.4%, 40.8%, 41.7% of the deviance for FSAP EG1, FSAP EG2 and FSAP EG3 respectively, and 37.2%, 24.3% and 21.6% of cross-validated deviance for FSAP EG1, FSAP EG2 and FSAP EG3 respectively.

3.5.3. Nitzschia species

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

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.

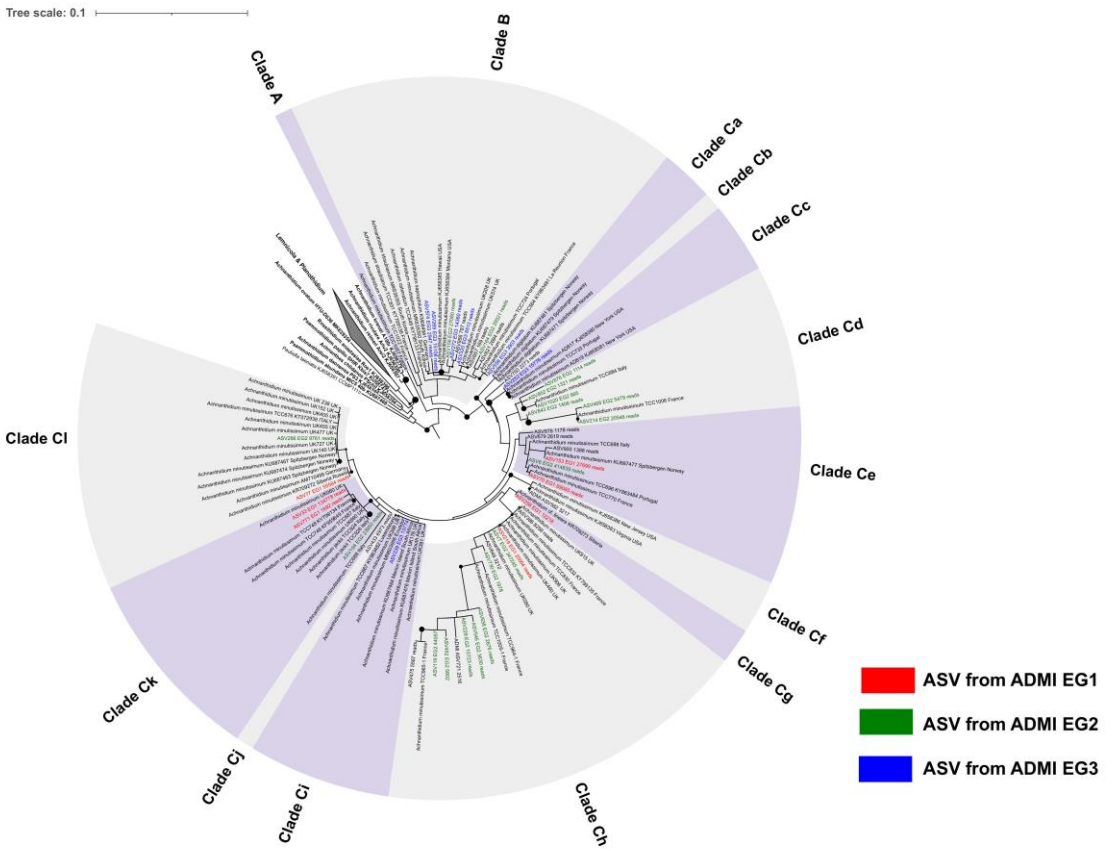


Fig. 2. Maximum likelihood phylogenetic tree of *Achnanthidium minutissimum* ASVs obtained in this study and related reference sequences extracted from Diat.barcode v9 and GenBank. The tree was obtained using raxmlGUI and a GRT-Gamma model with 1000 replicates for the bootstrap analyses. The tree was 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 (circle diameter is proportional to bootstrap value >50). Three major clades can be distinguished (A, B and C) and a number of subclades (Ca to Cl). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

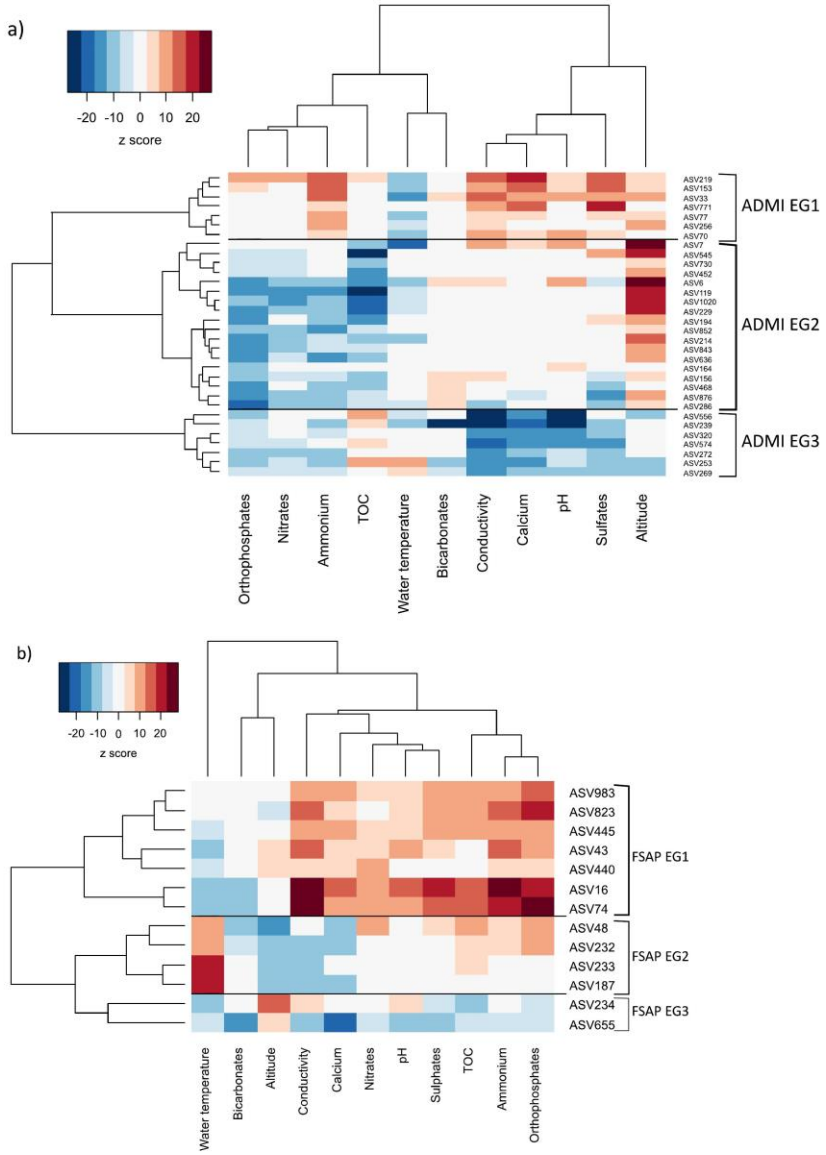


Fig. 3. Heatmap based on z scores obtained by the different TITAN analyses performed on ASVs from a) *Achnanthidium minutissimum* and b) *Fistulifera saprophila*. Euclidean distance and ward.D functions were used to compute dissimilarity distance and hierarchical clustering respectively on ASVs z scores 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 indicates negative responses. The magnitudes of the responses (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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distributed in waters with significant differences levels of calcium, conductivity, NH_4^+, NO_3^-, pH, PO_4^{3-}, SO_4^{2-} and TOC (Table 3).

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

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

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

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

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

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

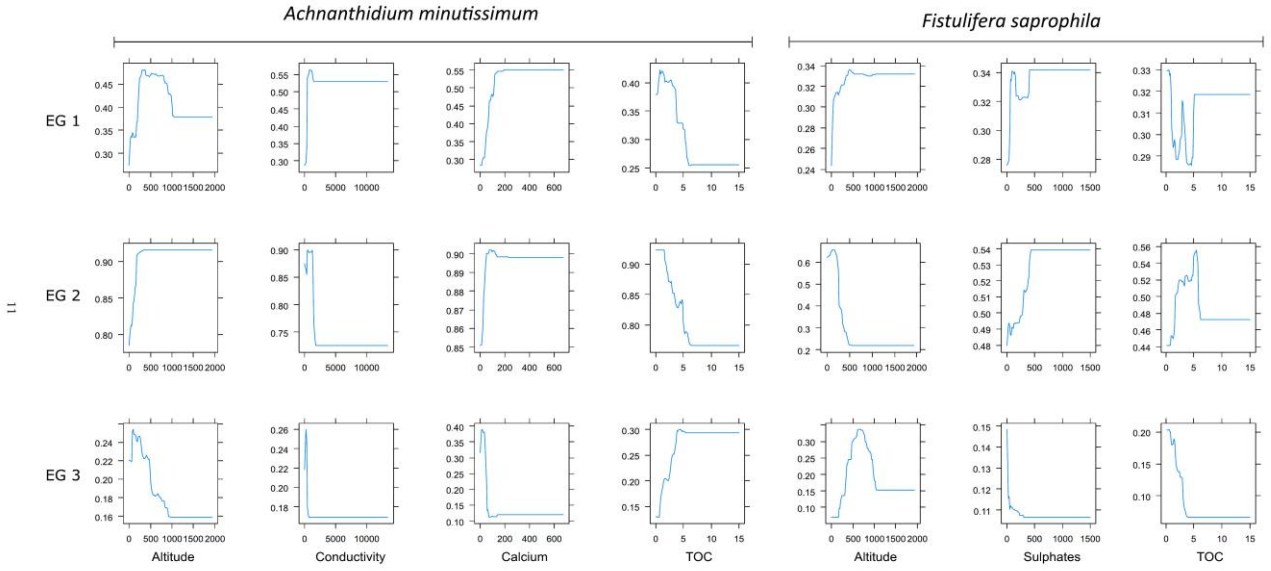


Fig. 4. Partial dependence plots generated by boosted regression trees analyses depicting the response of the ecological groupings of ASVs from *Achnanthidium minutissimum* to altitude (m), Calcium (mg/L), conductivity (μS/cm) and Total Organic Carbon (TOC, mg/L) and ecological groupings of ASVs from *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.

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]83.01-13,341.33(1077.06 ± 1452.74)	[a]48-2738(550.23 ± 452.36)	[b]30.67-2377.67(556.84 ± 480.72)	[a]77.5-13,341.33(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)

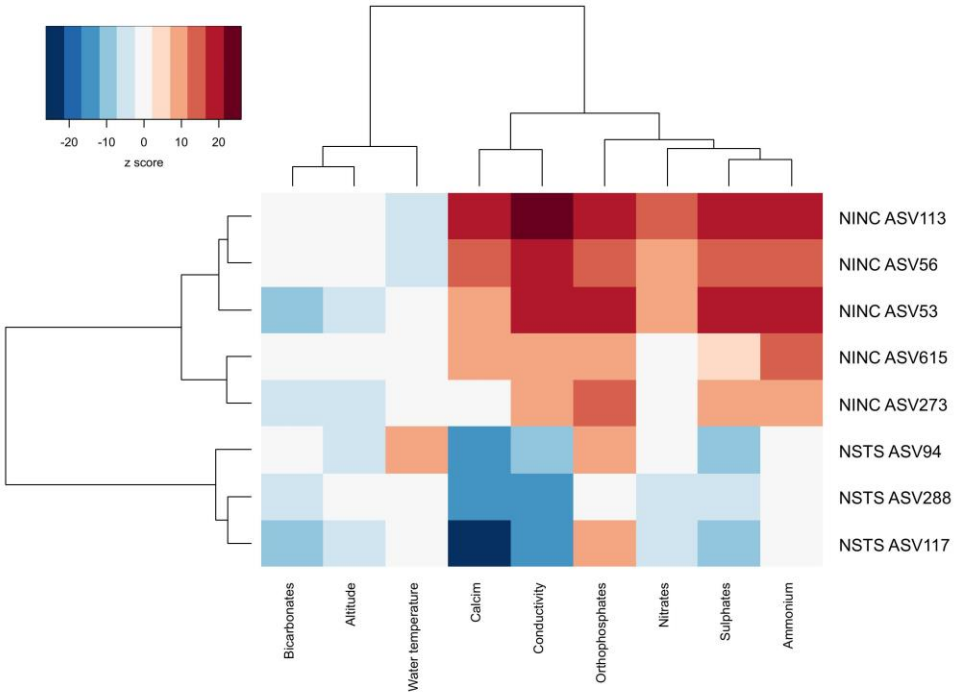


Fig. 5. Heatmap based on z score obtained by the different TITAN analyses performed on ASVs from *Nitzschia inconspicua* (NINC) and *N. soratensis* (NSTS). Only those ASVs with more than 3 responses that fulfilled purity and reliability criteria are represented. Red colour indicates positive responses 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Extra factors that need to be taken into account in interpreting the spatial structuring observed in some ASVs are i) spatial structuring of key environmental variables and ii) the possibility that important variables were not measured. Spatial structuring of the environment was particularly obvious in the case of calcium, conductivity and sulphates, whose levels were generally higher in Catalan rivers than French ones (Table 1). This could partly explain why ASVs characterized by a strong positive response to calcium and conductivity often predominated in Catalan rivers or were restricted there (e.g. ASV153; ASV219; ASVs from *N. inconspicua*), whereas ASVs that showed a strong negative response were often better represented in France (e.g. ASV269; ASVs from *N. soratensis*). Unmeasured environmental parameters - such as substrate composition, dissolved oxygen, turbidity, water flow, channel width or metals concentration – may also be influential (cf. Castro et al., 2019; Dalu et al., 2017; Keck et al., 2018a) accounting for the low amount of variance explained by the RDA model built from environmental data.

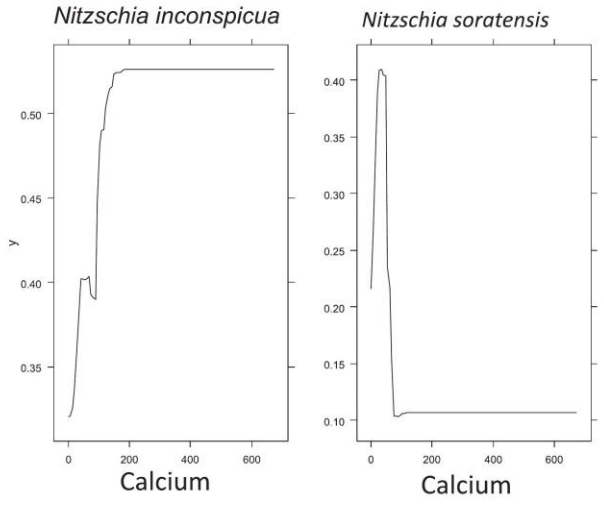


Fig. 6. Partial dependence plots generated by boosted regression trees analyses depicting the response of ASVs from *Nitzschia inconspicua* and *Nitzschia soratensis* to calcium (mg/L). Y axis represents fitted function.

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

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

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

In both species we found that two of the ecological groupings of ASVs were clearly separated by their opposite responses to calcium and conductivity, while in the case of *F. saprophila* a third ecological grouping (FSAP EG3) showed a preference for waters with low organic pollution. It might be argued that the type of response shown by this grouping corresponds better, within the genus *Fistulifera*, to *F. pelliculosa*, since this species is considered to occur from oligo to mesotrophic habitats

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

In a lesser way, deviation from the 'expected' ecology was also observed in *A. minutissimum*. Whereas one grouping of ASVs (ADMI EG3) was particularly restricted to low nutrient concentrations (i.e. PO_4^{3-}, SO_4^{2-}, NH_4^+ and NO_3^-), as might be expected from the characterization of *A. minutissimum* as an indicator of nutrient-poor, good quality waters (e.g. Potapova and Charles, 2007, especially Appendix A), the other two groupings of ASVs tolerated higher nutrient levels and would explain extension of the species complex into more nutrient-rich waters, creating the impression of a broad ecological tolerance – hence the characterization by Lange-Bertalot et al. (2017) "ecological amplitude apparently very wide" (see also Potapova and Hamilton, 2007; Snoeijs and Balashova, 1998; Round, 2004). The idea that *A. minutissimum* is a heterogeneous collection of lineages with different ecological preferences is not new. For example, Potapova and Hamilton (2007) were able to distinguish morphotypes within *A. minutissimum* and to associate them to some extent with different preferences for conductivity, pH and nutrients. However, the morphological differences between these variants (and between some of those documented by Pinseel et al., 2017) are very subtle and distinguishing them in LM-based assessments is arguably impractical. The metabarcoding approach not only aids identification but also allows vastly greater sampling of *A. minutissimum* across natural communities.

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

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

The preferences we obtained for the ASVs are based on correlations between their relative abundances in different samples and the environmental characteristics at the sites where the samples were obtained, exactly as has been done previously with microscopical cell counts to determine the preferences of morphologically defined species. These correlations likely reflect adaptations of the ASVs to different ecological conditions and ASVs that are closely related phylogenetically might be expected to share similar adaptations and belong to the same ecological grouping (Keck et al., 2016a, 2016b, 2018b). Overall, we did not find very strong evidence of a correlation between ecological and phylogenetic groupings, though there were some trends that could be observed in some cases. For instance, in *A. minutissimum*, the more distantly related ASVs generally belonged to the groupings that differed most (i.e. ADMI EG1 and ADMI EG3). And in *Fistulifera*, ASV74, which tolerated a high conductivity level (c. 9.000 μS/cm), was closely related to a sequence (HQ337547) from clone CCMP543, isolated from a brackish pond (in Massachusetts USA; this clone is often kept in fully marine medium), and clone TCC809, isolated from the River Arão estuary in Portugal (Rimet et al., 2019). However, the *F. saprophila* ASV recorded in the highest conductivity site in our dataset (c. 13.000 μS/cm) was ASV445, which is not closely related to ASV74 and belongs to a clade whose other members were recorded from freshwaters.

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

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

5. 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.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.149029.

CRediT authorship contribution statement

Javier Pérez-Burillo: Formal analysis, Investigation, Data curation, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Rosa Trobajo:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Manel Leira:** Formal analysis, Writing – review & editing. **François Keck:** Formal analysis, Writing – review & editing. **Frédéric Rimet:** Writing – review & editing. **Javier Sigró:** Writing – review & editing. **David G. Mann:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

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.

References

AFNOR, 2007. NF T90-354. Qualité de l'eau - Détermination de l'Indice Biologique Diatomées (IBD). AFNOR, pp. 1–79.
Alberdi, A., Aizpurua, O., Gilbert, M., Bohmann, K., 2017. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol. Evol. 9 (1), 134–147. https://doi.org/10.1111/2041-210X.12849.
Armbrust, E.V., 2009. The life of diatoms in the world's oceans. Nature 459, 185–192. https://doi.org/10.1038/nature08057.
Bailet, B., Apothéloz-Perret-Gentil, L., Baricevic, A., Chonova, T., Franc, A., Frigerio, J.-M., Kelly, M., Mora, D., Pfannkuchen, M., Proft, S., Ramon, M., Vasselon, V., Zimmermann, J., Kahlert, M., 2020. Diatom DNA metabarcoding for ecological assessment: comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. Sci. Total Environ. 745, 140948. https://doi.org/10.1016/j.scitotenv.2020.140948.
Baker, M.E., King, R.S., 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods Ecol. Evol. 1 (1), 25–37. https://doi.org/10.1111/j.2041-210X.2009.00007.x.
Baker, M.E., King, R.S., Kahle, D., 2019. TITAN2: threshold indicator taxa analysis. R package, version 2.4. https://CRAN.R-project.org/package=TITAN2.
Blanchet, F.G., Legendre, P., Borcard, D., 2008. Forward selection of explanatory variables. Ecology 89 (9), 2623–2632. https://doi.org/10.1890/07-0986.1.
Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high resolution sample inference from illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869.
Castro, E., Siqueira, T., Melo, A.S., Bini, L.M., Landeiro, V.L., Schneck, F., 2019. Compositional uniqueness of diatoms and insects in subtropical streams is weakly correlated with riffle position and environmental uniqueness. Hydrobiologia 842, 219–232. https://doi.org/10.1007/s10750-019-04037-8.
CEN, 2014. CEN_EN 13946: Water Quality - Guidance for the Routine Sampling and Preparation of Benthic Diatoms From Rivers and Lakes, pp. 1–22.
CEN, 2018. CEN/TR 17245: Water Quality –Technical Report for the Routine Sampling of Benthic Diatoms From Rivers and Lakes Adapted for Metabarcoding Analysis. CEN/TC 230/WG23 – Aquatic Macrophyte and Algae, pp. 1–8.
Chonova, T., Keck, F., Bouchez, A., Rimet, F., 2020. A ready-to-use database for DADA2: Diat.barcode_rbcL_263bp_DADA2 based on Diat.barcode v9. Portail Data INRAE, V2 https://doi.org/10.15454/QBLSXP.
Dalu, T., Wasserman, R.J., Magoro, M.L., Mwedzi, T., Froneman, P.W., Weyl, O.L.F., 2017. Variation partitioning of benthic diatom community matrices: effects of multiple variables on benthic diatom communities in an austral temperate river system. Sci. Total Environ. 601, 73–82. https://doi.org/10.1016/j.scitotenv.2017.05.162.
De Luca, D., Piredda, R., Sarno, D., Kooistra, W.H.C.F., 2021. Resolving cryptic species complexes in marine protists: phylogenetic haplotype networks meet global DNA metabarcoding datasets. ISME J. https://doi.org/10.1038/s41396-021-00895-0.
Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., Vere, N., Pfrender, M.E., Bernatchez, L., 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. 26 (21), 5872–5895. https://doi.org/10.1111/mec.14350.
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., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbours matrices (PCNM). Ecol. Model. 196 (3–4), 483–493. https://doi.org/10.1016/j.ecolmodel.2006.02.015.
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.
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 Res. 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. J. Anim. Ecol. 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. Ann. Stat. 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. Freshw. Biol. 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. Wiley, New York, NY, USA.
Jahn, R., Abarca, N., Gemeinholzer, B., Mora, D., Skibbe, O., Kulikovskiy, M., Gusev, E., Kusber, W.H., Zimmermann, J., 2017. *Planothidium lanceolatum* and *Planothidium frequentissimum* reinvestigated with molecular methods and morphology: four new species and the taxonomic importance of the sinus and cavum. Diatom Res. 32 (1), 75–107. https://doi.org/10.1080/0269249X.2017.1312548.
Keck, F., Bouchez, A., Franc, A., Rimet, F., 2016a. Linking phylogenetic similarity and pollution sensitivity to develop ecological assessment methods: a test with river diatoms. J. Appl. Ecol. 53 (3), 856–864. https://doi.org/10.1111/1365-2664.12624.
Keck, F., Rimet, F., Franc, A., Bouchez, A., 2016b. Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring. Ecol. Appl. 26 (3), 861–872. https://doi.org/10.1890/14-1966.
Keck, F., Franc, A., Kahlert, M., 2018a. Disentangling the processes driving the biogeographyof freshwater diatoms: a multiscale approach. J. Biogeogr. 45, 1582–1592. https://doi.org/10.1111/jbi.13239.
Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18 (6), 1299–1309. https://doi.org/10.1111/1755-0998.12919.
Kelly, M.G., Trobajo, R., Rovira, L., Mann, D.G., 2015. Characterizing the niches of two very similar *Nitzschia* species and implications for ecological assessment. Diatom Res. 30 (1), 27–33. https://doi.org/10.1080/0269249X.2014.951398.
Kelly, M.G., Juggins, S., Mann, D.G., Sato, S., Glover, R., Boonham, N., Sapp, M., Lewis, E., Hany, U., Kille, P., Jones, T., Walsh, K., 2020. Development of a novel metric for

evaluating diatom assemblages in rivers using DNA metabarcoding. Ecol. Indic. 118, 106725. https://doi.org/10.1016/j.ecolind.2020.106725.
Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35 (6), 1547–1549. https://doi.org/10.1093/molbev/msy096.
Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., 2017. Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment. English Edition With Updated Taxonomy and Added Species. English Edition With Updated Taxonomy and Added Species. Schmitten-Oberreifenberg, Koeltz Botanical Books.
Lanzén, A., Mendibil, I., Borja, Á., Alonso-Sáez, L., 2020. A microbial mandala for environmental monitoring: predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. Mol. Ecol. 1–19. https://doi.org/10.1111/mec.15489.
Lecointe, C., Coste, M., Prygiel, J., 1993. OMNIDIA—software for taxonomy, calculation of diatom indexes and inventories management. Hydrobiologia 269, 509–513. https://doi.org/10.1007/BF00028048.
Legendre, P., Legendre, L., 2012. Chapter thirteen – spatial analysis. Numerical Ecology, 3rd ed. Elsevier Science BV, Amsterdam, pp. 785–858.
Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256-W259. https://doi.org/10.1093/nar/gkz239.
Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A., Bowler, C., 2016. Insights into global diatom distribution and diversity in the world's ocean. Proc. Natl. Acad. Sci. U. S. A. 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. Mol. Phylogenet. Evol. 158, 106985. https://doi.org/10.1016/j.ympev.2020.106985.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 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. Ecol. Indic. 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. Sci. Total Environ. 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. Eur. J. Phycol. 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. Sci. Rep. 8, 18059. https://doi.org/10.1038/s41598-018-36345-9.
Pniewski, F., Friedl, T., Latala, A., 2010. Identification of diatom isolates from the Gulf of Gdansk: testing of species identifications using morphology, 18S rDNA sequencing and DNA barcodes of strains from the culture collection of Baltic algae (CCBA). Oceanol. Hydrobiol. Stud. 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. Ecol. Indic. 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. J. Phycol. 43 (3), 561–575. https://doi.org/10.1111/j.1529-8817.2007.00332.x.
Poulíčková, A., Špacková, 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.
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. Sci. Total Environ. 599–600, 820–833. https://doi.org/10.1016/j.scitotenv.2017.05.034.
R Core Team, R, 2020. A Language and Environment for Statistical Computing. Retrieved from. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
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. Org. Divers. Evol. 18, 51–62. https://doi.org/10.1007/s13127-018-0359-5.
Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M., Maltsev, Y., Mann, D.G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2019. Diat. barcode, an open-access curated barcode library for diatoms. Sci. Rep. 9, 1–12. https://doi.org/10.1038/s41598-019-51500-6.
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. Ecol. Indic. 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*. J. Eukaryot. Microbiol. 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 brightwellii*. Limnol. Oceanogr. 51 (3), 1249–1261. https://doi.org/10.4319/lo.2006.51.3.1249.
Smetacek, V., 1999. Diatoms and the ocean carbon cycle. Protist 150 (1), 25–32. https://doi.org/10.1016/S1434-4610(99)70006-4.
Smucker, N.J., Pilgrim, E.M., Nietch, C.T., Darling, J.A., Johnson, B.R., 2020. DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. Ecol. Appl. 30 (8), e02205. https://doi.org/10.1002/eap.2205.
Snoeijs, P., Balashova, N., 1998. Intercalibration and distribution of diatom dpecies in the Baltic Sea. Opulus Press, Uppsala.
Soininen, J., Jamoneau, A., Rosebery, J., Leboucher, T., Wang, J., Kokocinski, M., Passy, S.I., 2018. Stream diatoms exhibit weak niche conservation along global environmental and climatic gradients. Ecography 42 (2), 346–353. https://doi.org/10.1111/ecog.03828.
Souffreau, C., Vanormelingen, P., Van de Vijver, B., Isheva, T., Verleyen, E., Sabbe, K., Vyverman, W., 2013. Molecular evidence for distinct antarctic lineages in the cosmopolitan terrestrial diatoms *Pinnularia borealis* and *Hantzschia amphioxys*. Protist 164 (1), 101–115. https://doi.org/10.1016/j.protis.2012.04.001.
Stoof-Leichsenring, K.R., Pestryakova, L.A., Epp, L.S., Herzschuh, U., 2020. Phylogenetic diversity and environment form assembly rules for Arctic diatom genera—a study on recent and ancient sedimentary DNA. J. Biogeogr. 47 (5), 1166–1179. https://doi.org/10.1111/jbi.13786.
Styczynski, M.P., Jensen, K.L., Rigoutsos, I., Stephanopoulos, G., 2008. BLOSUM62 miscalculations improve search performance. Nat. Biotechnol. 26 (3), 274–275. https://doi.org/10.1038/nbt0308-274.
Tapolczai, K., Selmeczy, G.G., Szabó, B., B-Béres, V., Keck, F., Bouchez, A., Rimet, F., Padisák, J., 2021. The potential of exact sequence variants (ESVs) to interpret and assess the impact of agricultural pressure on stream diatom assemblages revealed by DNA metabarcoding. Ecol. Indic. 122, 107322. https://doi.org/10.1016/j.ecolind.2020.107322.
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* complex (Bacillariophyceae). Eur. J. Phycol. 45 (4), 413–425. https://doi.org/10.1080/09670262.2010.498586.
Trobajo, R., Rovira, L., Ector, L., Wetzel, C.E., Kelly, M., Mann, D.G., 2013. Morphology and identity of some ecologically important small *Nitzschia* species. Diatom Res. 28 (1), 37–59. https://doi.org/10.1080/0269249X.2012.734531.
Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., Yamanaka, H., 2019. Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: a case study using tank water. Environ. DNA 2 (1), 42–52. https://doi.org/10.1002/edn3.44.
Turon, X., Antich, A., Palacín, C., Præbel, K., Wangensteen, O.S., 2019. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol. Appl. 30 (2), e02036. https://doi.org/10.1002/eap.2036.
Vanormelingen, P., Evans, K.M., Mann, D.G., Lance, S., Debeer, A.-E., D'Hondt, S., Verstraete, T., De Meester, L., Vyverman, W., 2015. Genotypic diversity and differentiation among populations of two benthic freshwater diatoms as revealed by microsatellites. Mol. Ecol. 24 (17), 4433–4448. https://doi.org/10.1111/mec.13336.
Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A., 2017a. Application of highthroughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? Freshw. Sci. 36, 162–177. 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). Ecol. Indic. 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. Freshw. Sc. 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. Appl. Environ. Microbiol. 73 (16), 5261–5267. https://doi.org/10.1128/AEM.00062-07.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845.

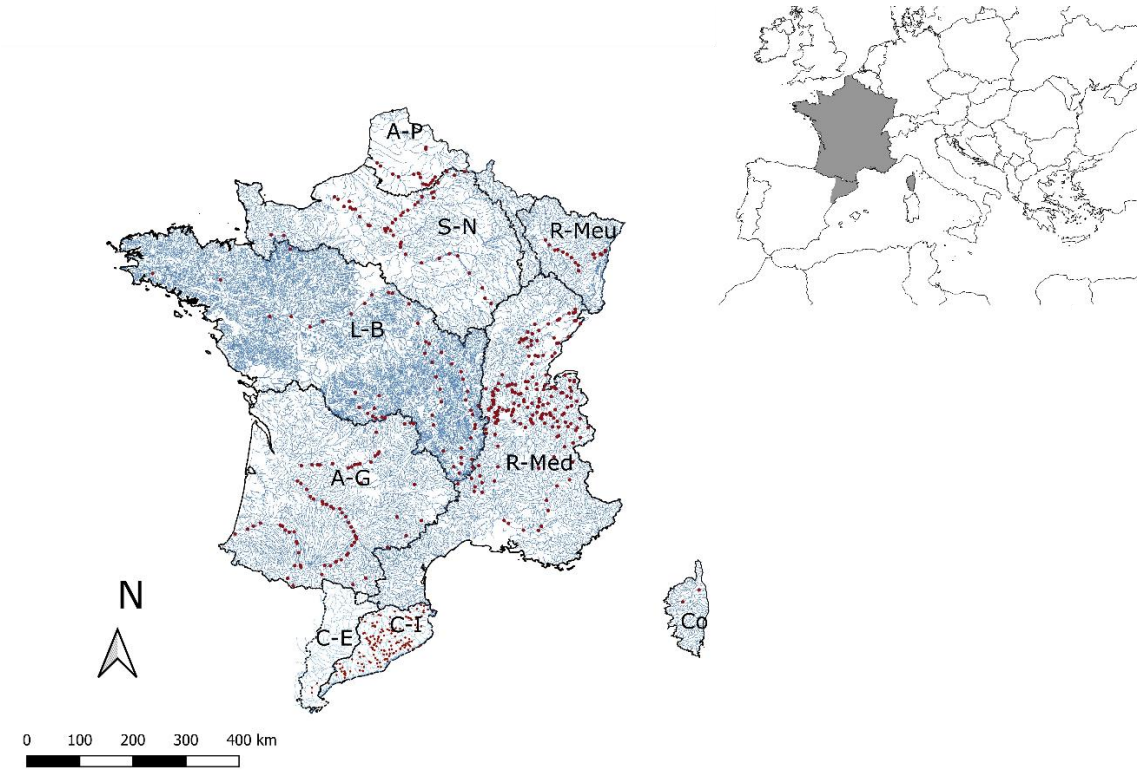
Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., 2020. Gplots: Various R programming tools for plotting data. R package, version 3.1.1. https://CRAN.R-project.org/package=gplots.

Wattier, R., Mamos, T., Copilas-Ciocianu, D., Jelic, 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). Sci. Rep. 10, 16536. https://doi.org/10.1038/s41598-020-73739-0.

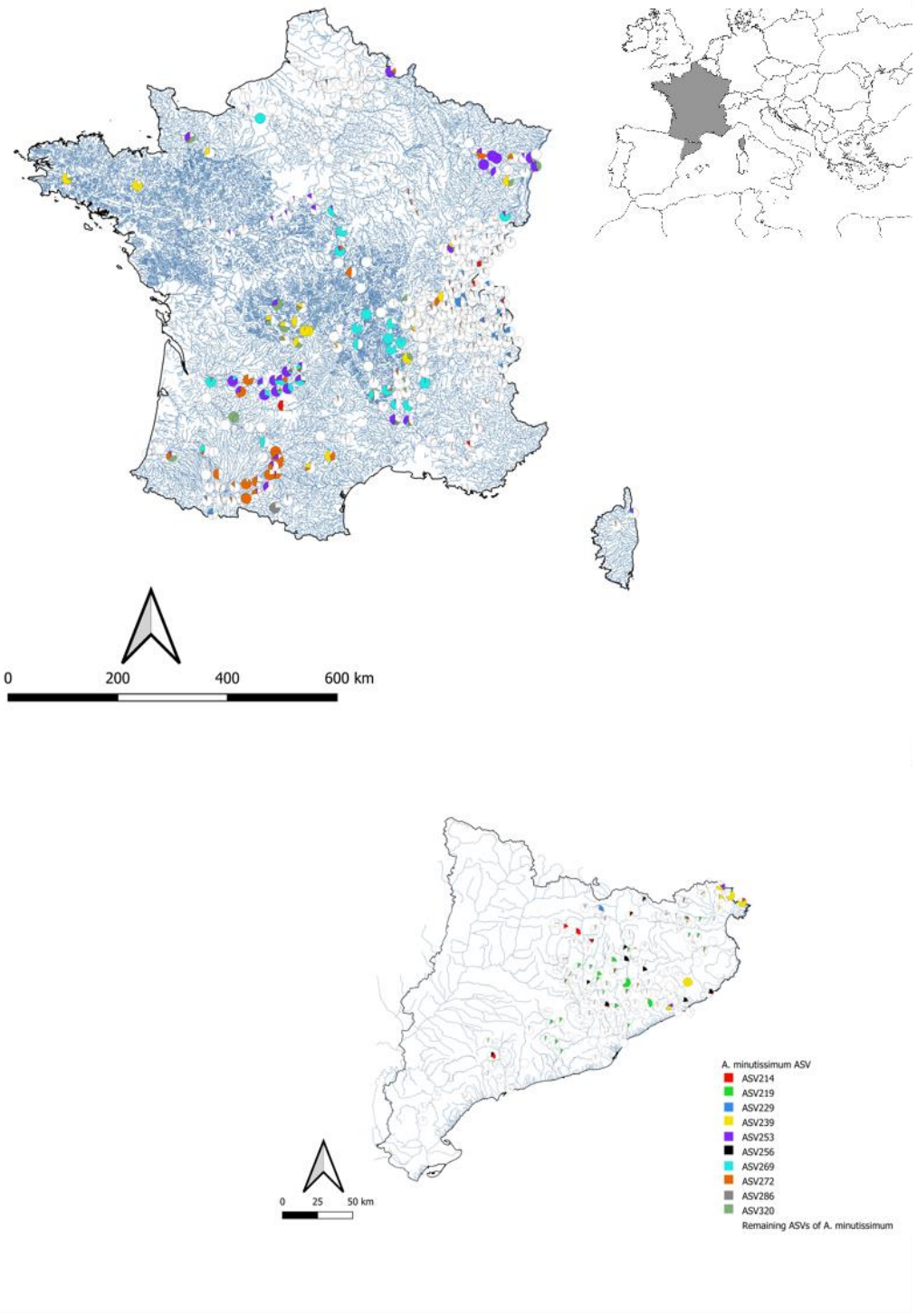
Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., Latala, A., 2013. Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. Diatom Res. 28 (4), 431–443. https://doi.org/10.1080/0269249X.2013.833136.

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 Metagenomics 4, e51925. https://doi.org/10.3897/mbmg.4.51925.

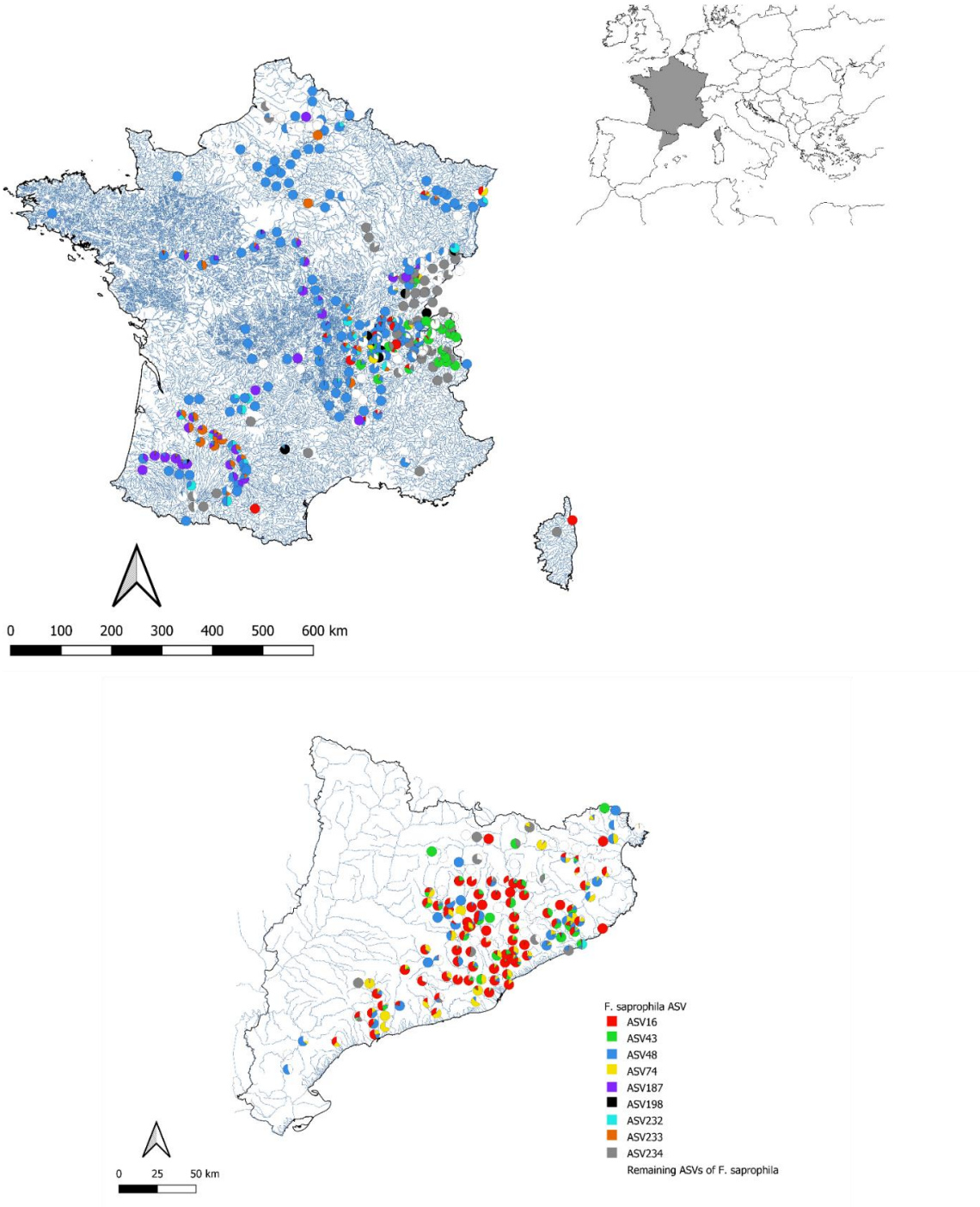
Supplementary material



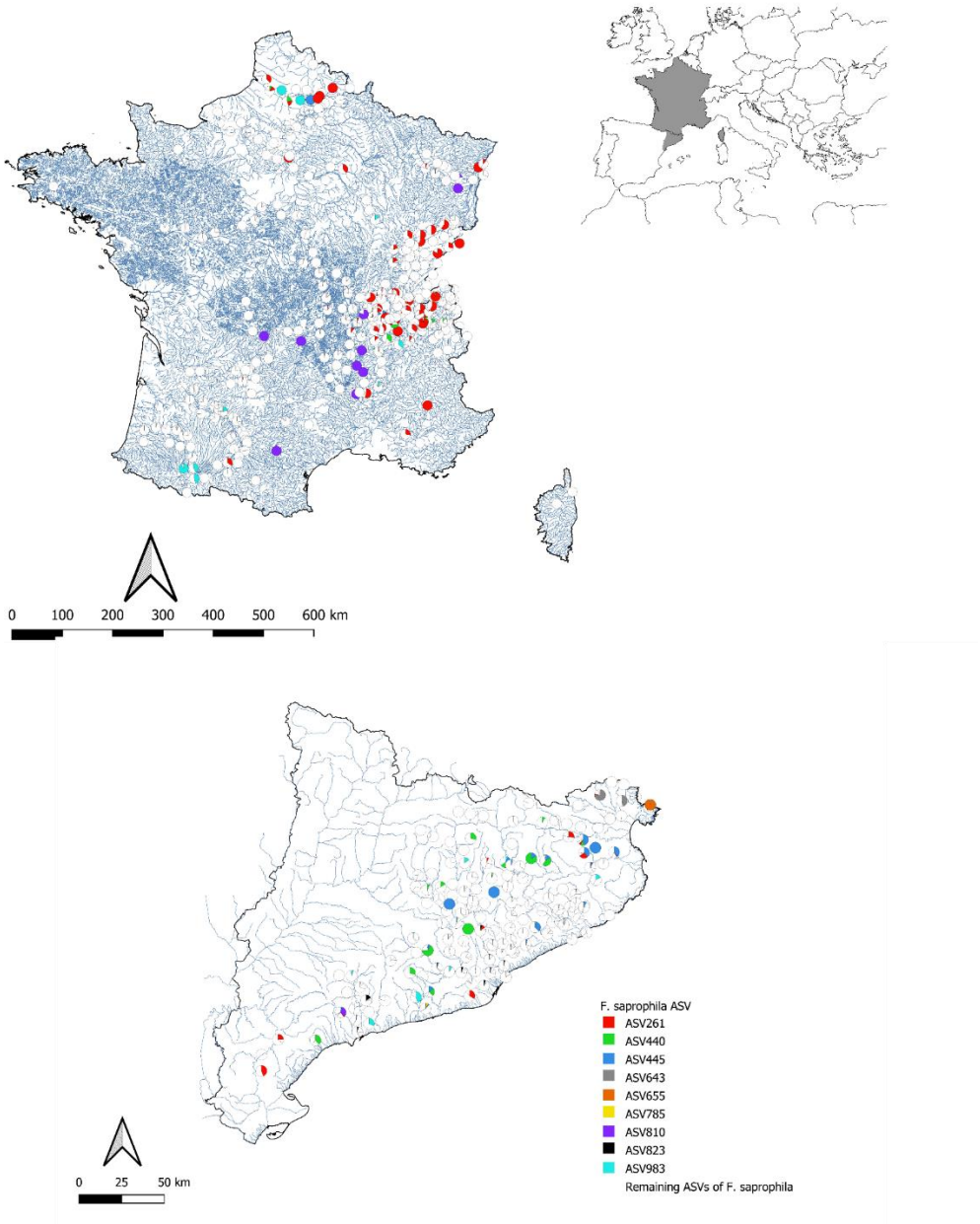
Supplementary Fig 1. Location of rivers sites analysed in this study for which, environmental variables were available. Main hydrographic basins from Catalonia (NE Spain) and France are delimitated and indicated as follow: CE (Catalan Interregonial basins); CI (Catalan Internal basins); A-G (Adour–Garonne basins); A-P(Artois–Picardie basins); L-B(Loire–Bretagne basins), R-Meu(Rhin–Meuse basins); R-Med(Rhône–Méditerranée basins); Corse (Co) and S-N(Seine-Normandie basins).



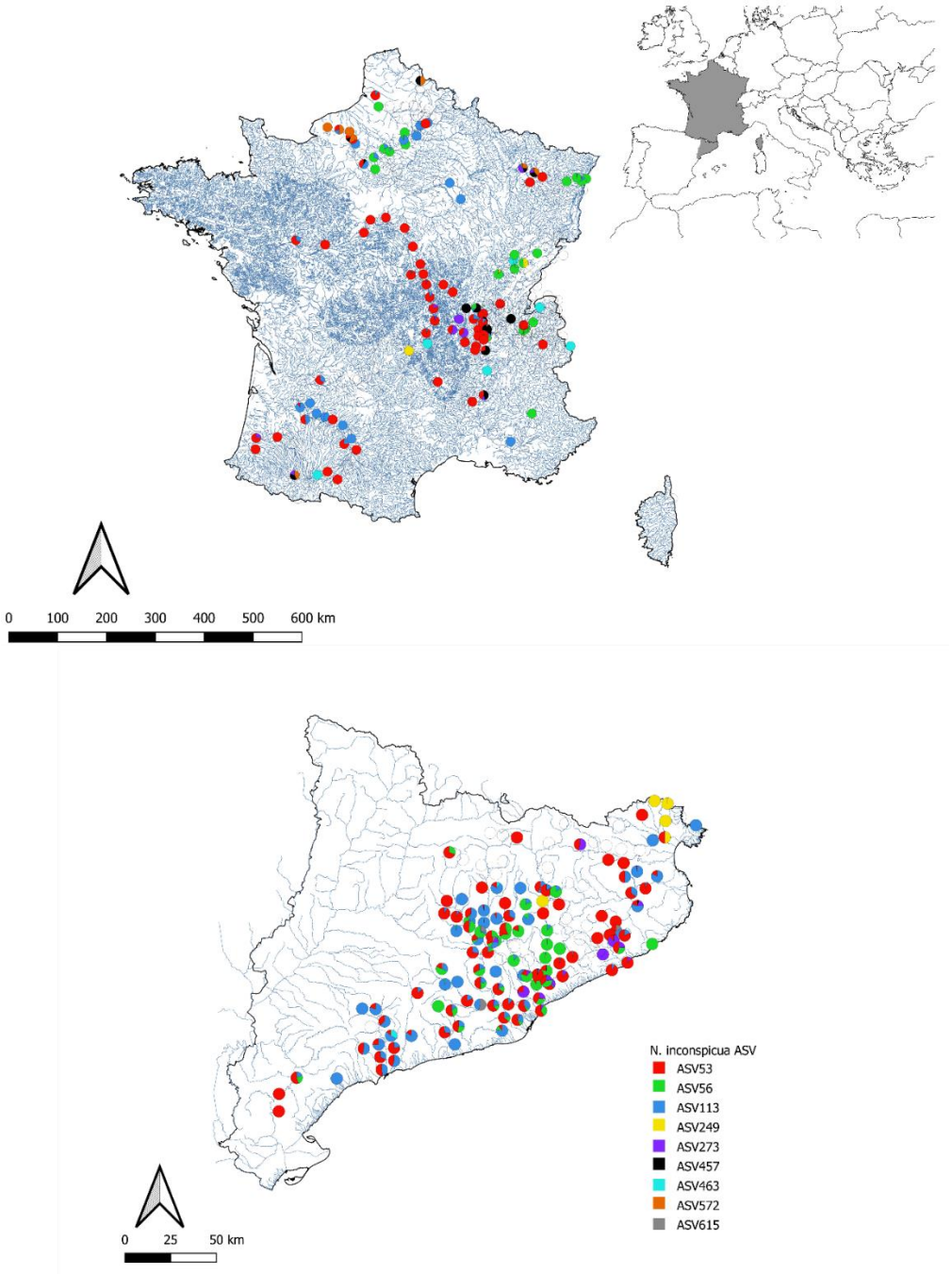
Supplementary Fig 2. Spatial distribution of the ASVs from *Achnanthidium minutissimum* in French and Catalan rivers. ASVs represented are the 11th to 20th most abundant of the species. Segments in each circle represent the proportion of *A. minutissimum* reads recorded in each sample site.



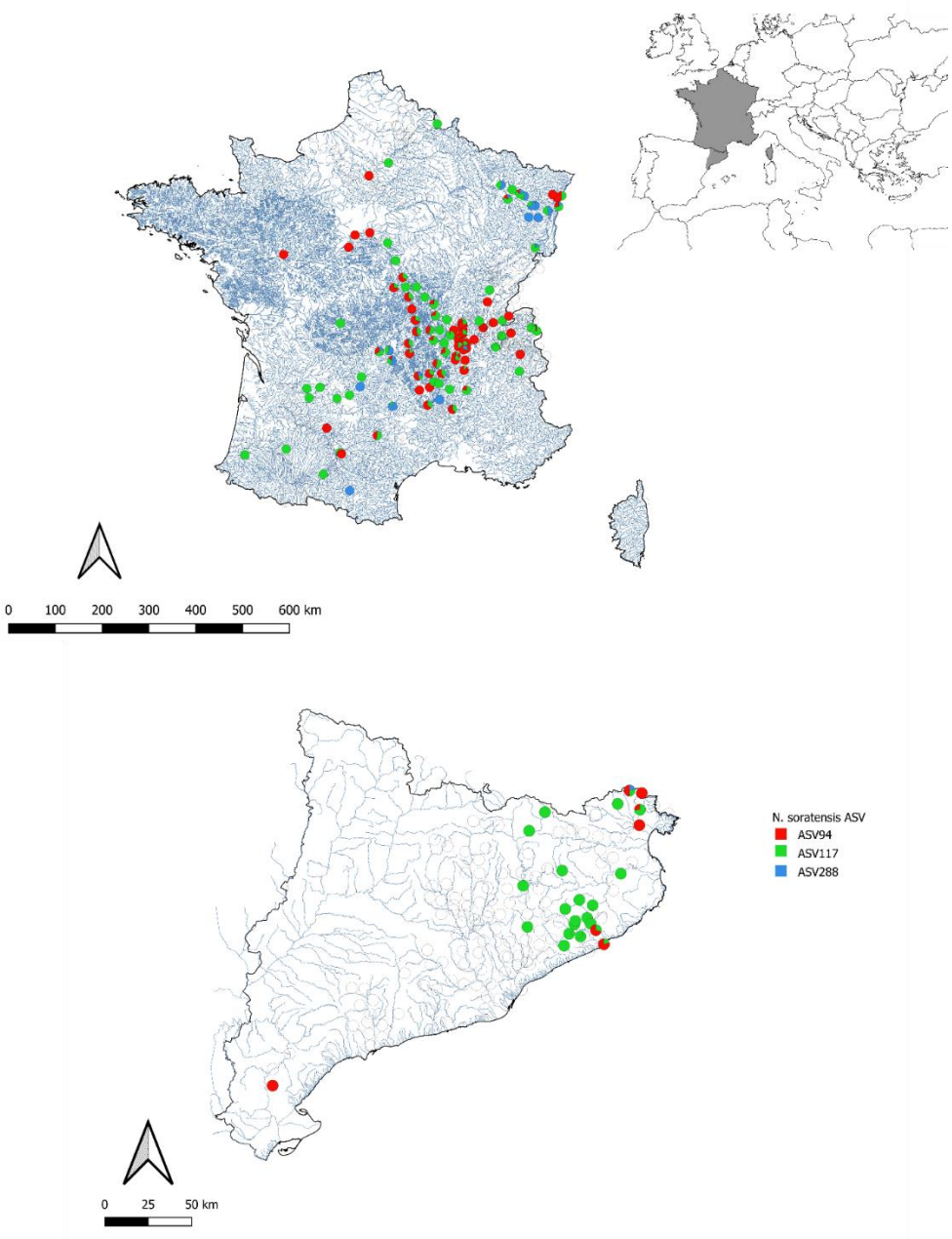
Supplementary Fig 3. Spatial distribution of the 10 most abundant ASVs from *Fistulifera saprophila* in French and Catalan rivers. Segments in each circle represent the proportion of *F. saprophila* reads recorded in each sample site.



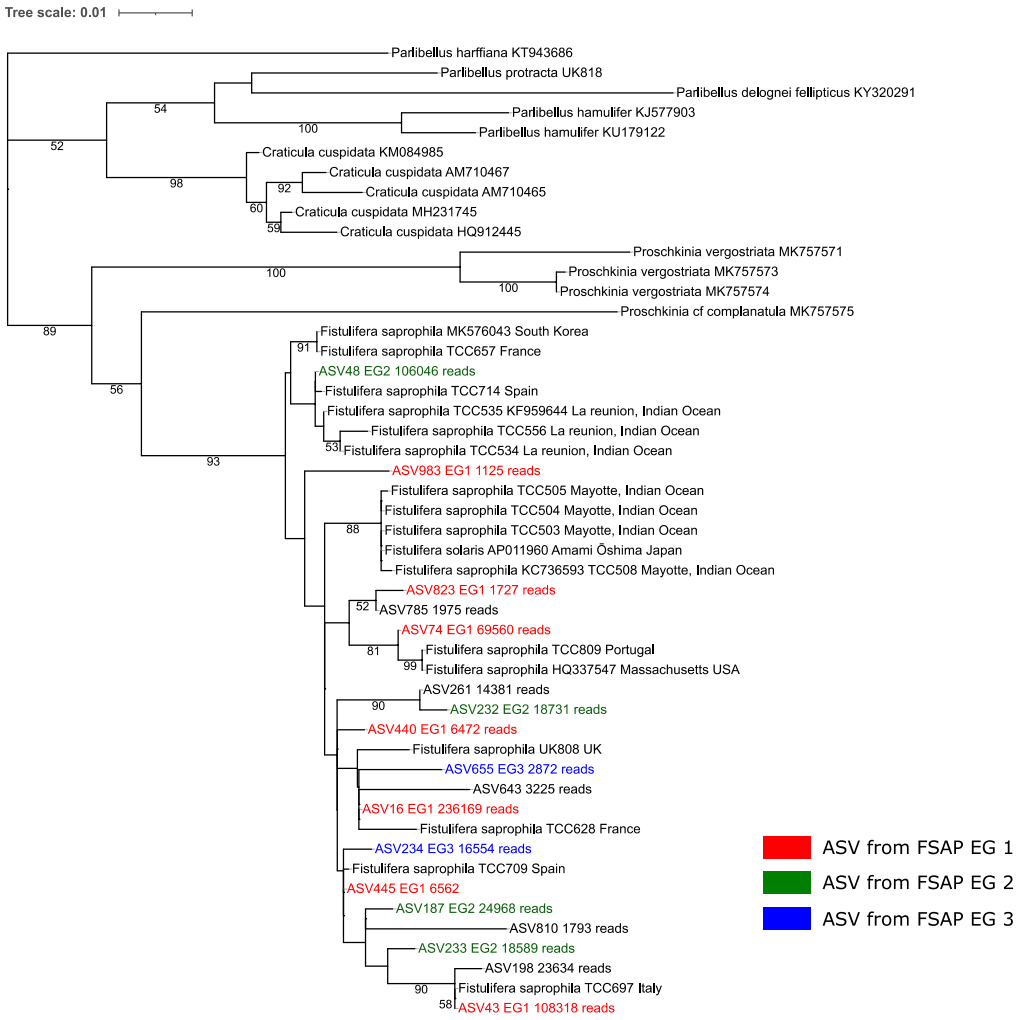
Supplementary Fig 4. Spatial distribution of the ASVs from *Fistulifera saprophila* in French and Catalan rivers. ASVs represented are the 11th to 20th most abundant of the species. Segments in each circle represent the proportion of *F. saprophila* reads recorded in each sample site.



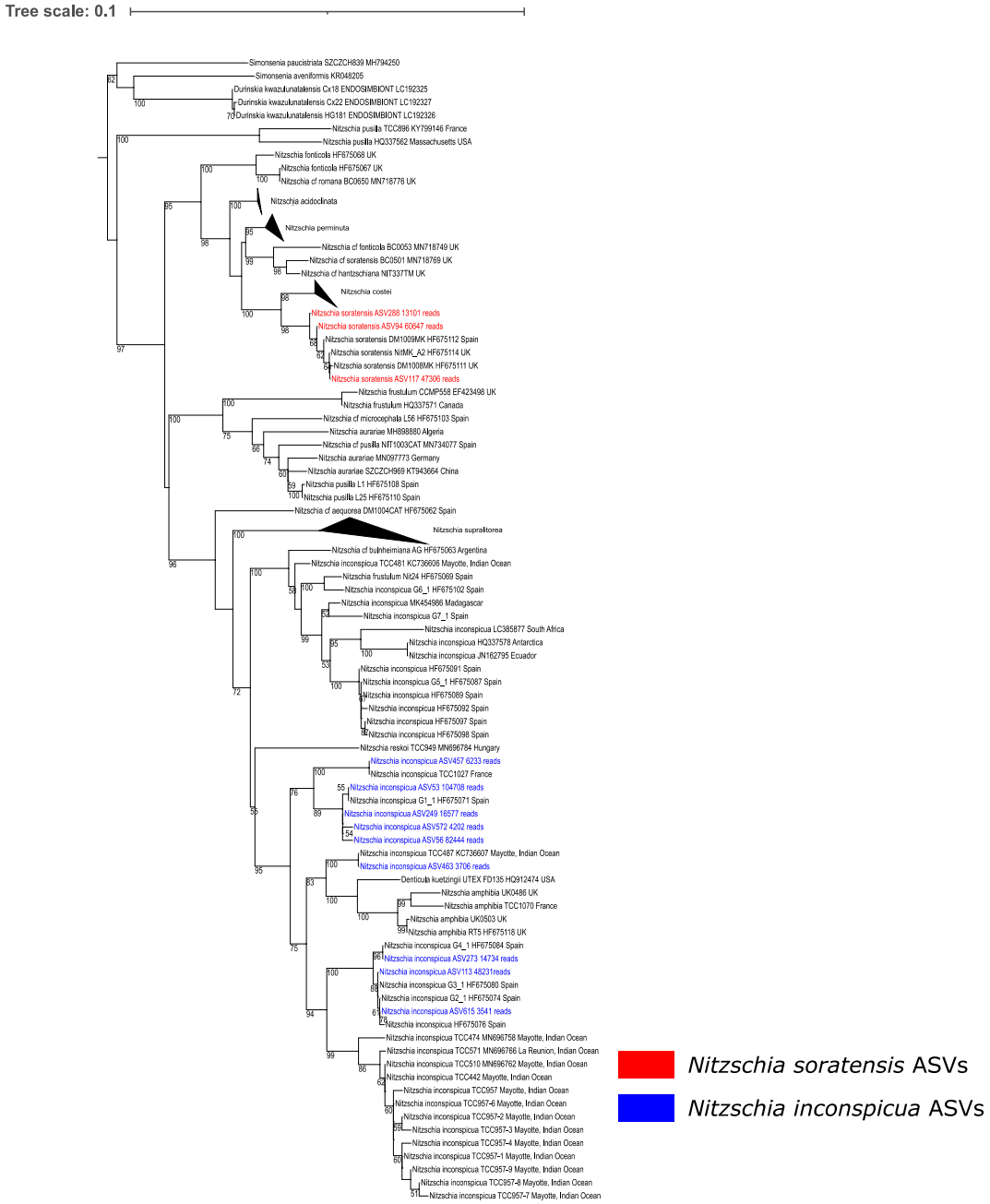
Supplementary Fig 5. Spatial distribution of the ASVs from *Nitzschia inconspicua* in French and Catalan rivers. Segments in each circle represent the proportion of *N. inconspicua* reads recorded in each sample site.



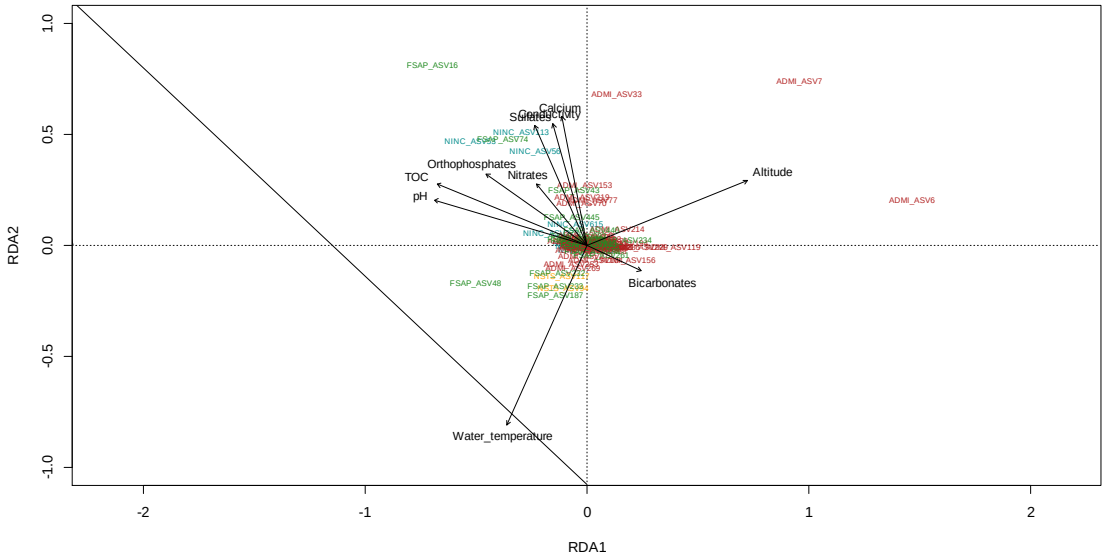
Supplementary Fig 6. Spatial distribution of the ASVs from *Nitzschia soratensis* in French and Catalan rivers. Segments in each circle represent the proportion of *N. soratensis* reads recorded in each sample site.



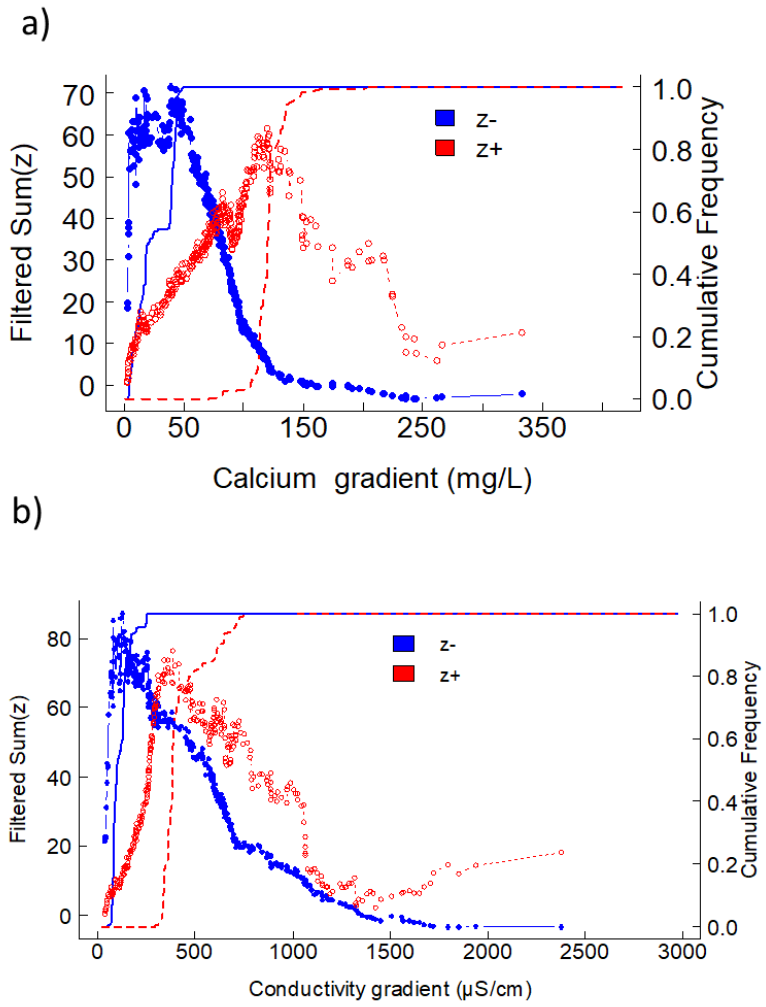
Supplementary Fig 7. Maximum likelihood phylogenetic tree based on *Fistulifera* saprophila ASVs obtained in this study and on sequences from *F. saprophila* and its closely related species extracted from Diat.barcode v9 and GenBank database. The tree was obtained using raxmlGUI and setting the GRT-Gamma 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. Bootstrap support values from 50 to 100 are represented.



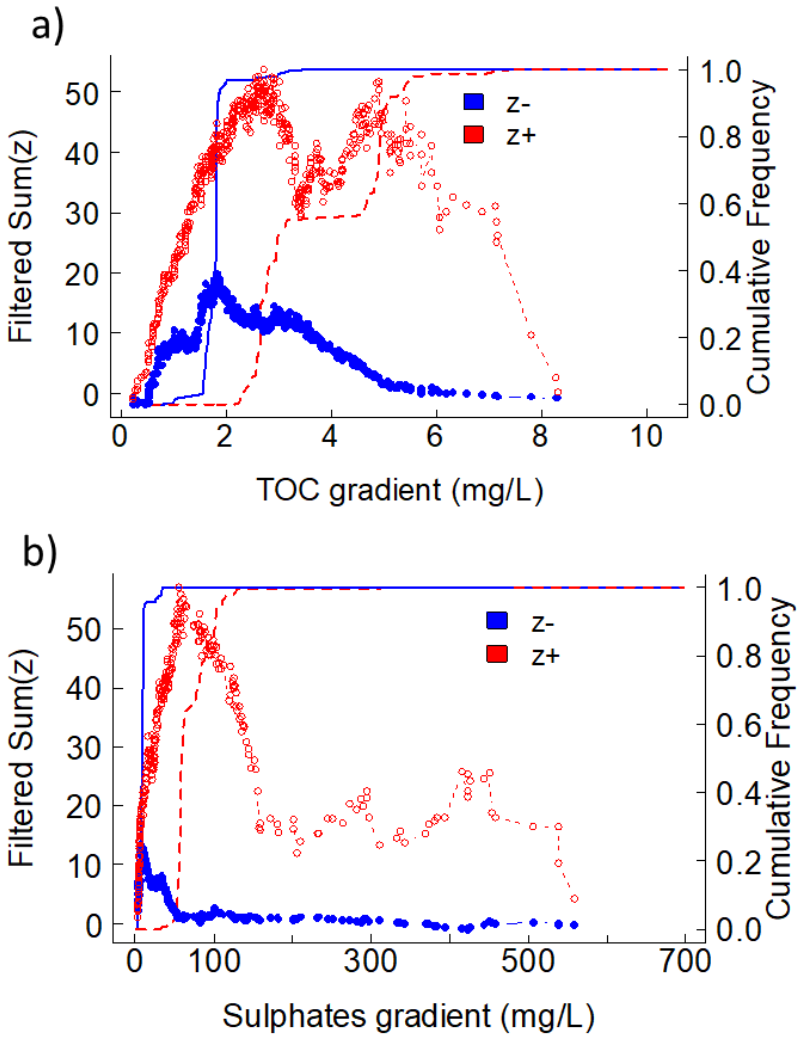
Supplementary Fig 8. Maximum likelihood phylogenetic tree based on *Nitzschia inconspicua and N. soratensis* ASVs obtained in this study and on sequences from both species and its closely related species extracted from Diat.barcode v9 and GenBank database. The tree was obtained using raxmlGUI and setting the GRT-Gamma model with 1000 replicates for the bootstrap analyses. The tree was drawn using iTOL. Bootstrap support values from 50 to 100 are represented.



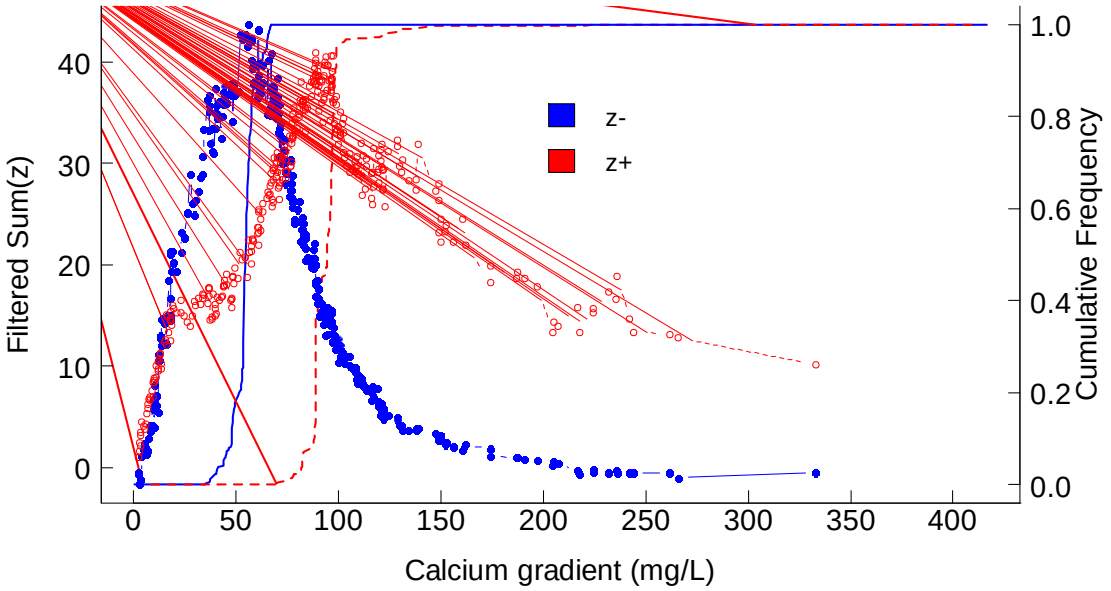
Supplementary Fig 9. Biplot from redundancy analysis based on ASVs from *Achnanthidium minutissimum* (ADMI; in red), *Fistulifera saprophila* (FSAP; in green), *Nitzschia inconspicua* (NINC; in blue) and *N. soratensis* (NSTS; in orange) and environmental variables selected as significant ($p < 0.05$) by forward selection and showing a Bonferroni adjusted p value < 0.05.



Supplementary Fig 10. TITAN analysis showing sum z scores of ASVs from *Achnanthidium minutissimum* for calcium (a) and conductivity (b). Left-Y axis represent the sum z scores of those ASVs that fulfilled pure and reliability criteria. Red circles correspond to sum z scores from positive responses and blue circles sum z scores from negative responses. Right-Y axis and dashed and continuous lines show the proportion of the distribution (cumulative frequency) of assemblage change points (given by the maximum sum z score) from 500 bootstrap replicates. Sum z scores indicated that the assemblage change point of ASVs with a negative response to calcium and conductivity (i.e., ASVs mainly from ADMI EG1 and ADMI EG3 since most of the ASVs that fulfilled both purity and reliability criteria for such responses belonged to ADMI EG1 and ADMI EG3) occurred at 39.5 mg/L (4.6–44.7, 5th–95th percentile) and at 130.2 µS/cm (72.2–199.3, 5^{th}–95^{th} percentile) respectively while they occurred at 119.7 mg/L (105.7–138.4, 5^{th}–95^{th} percentile) and at 384.9 µS/cm (335–700, 5^{th}–95^{th} percentile) respectively for positive responders



Supplementary Fig 11. TITAN analysis showing sum z scores of ASVs from *Fistulifera saprophila* for organic carbon (a) and sulphates (b). Left-Y axis represent the sum z scores of those ASVs that fulfilled pure and reliability criteria. Red circles correspond to sum z scores from positive responses and blue circles sum z scores from negative responses. Right-Y axis show the proportion of the distribution (cumulative frequency) of assemblage change points (given by the maximum sum z score) from 500 bootstrap replicates. Sum z scores indicated that the assemblage change point of ASVs with a positive response to TOC and SO_4^{2-} (FSAP EG1 and FSAP EG2) occurred at 2.7 mg/L TOC (2.3–5.4, 5th– 95th percentile) and at 55.2 mg/L SO_4^{2-} (52.5–110.1, 5th– 95th percentile) respectively, while it occurred at 1.8 mg/L TOC (1.5–1.9, 5th– 95th percentile) and at 7.7 mg/L SO_4^{2-} (2.9–12.1, 5th– 95th percentile) respectively for negative responders (FSAP EG3)



Supplementary Fig 12. TITAN analysis showing sum z scores of ASVs from *Nitzschia inconspicua* (NINC) and *N. soratensis* (NSTS) for Calcium. Left-Y axis represent the sum z scores of those ASVs that fulfilled pure and reliability criteria. Red circles correspond to sum z scores from positive responses (ASVs from NINC) and blue circles sum z scores from negative responses (ASVs from NSTS). Right-Y axis show the proportion of the distribution (cumulative frequency) of assemblage change points (given by the maximum sum z score) from 500 bootstrap replicates. Sum z scores regarding calcium identified assemblage changes points for NINC ASVs at 88.9 mg/L (82,8–99.3, 5th– 95th percentile) while they occurred at 56.4 mg/L (44–65.3, 5th– 95th percentile) for ASVs of NSTS.

Supplementary Table 1. Abundance and occurrence data of amplicon sequence variants (ASVs) from *Nitzschia inconspicua* (NINC), *N. soratensis* (NSTS), *Achnanthidium minutissimum* (ADMI) and *Fistulifera saprophila* (FSAP) species recorded throughout the total of 531 samples used for statistical analyses.

ASV id	Species	Abundance (reads)	Relative abundance	Occurrence (French rivers sites)	Occurrence (Catalan rivers sites)
ASV53	NINC	104708	4.27	61	87
ASV56	NINC	82444	3.36	26	44
ASV113	NINC	48231	1.97	41	81
ASV249	NINC	16577	0.68	3	6
ASV273	NINC	14734	0.6	11	19
ASV457	NINC	6233	0.25	14	1
ASV463	NINC	3706	0.15	10	4
ASV572	NINC	4202	0.17	10	0
ASV615	NINC	3541	0.14	0	8
ASV94	NSTS	60647	2.47	66	7
ASV117	NSTS	47306	1.93	80	21
ASV288	NSTS	13101	0.53	22	1
ASV6	ADMI	414839	16.92	257	86
ASV7	ADMI	347245	14.16	217	121
ASV33	ADMI	134778	5.5	71	82
ASV70	ADMI	69085	2.82	37	46
ASV77	ADMI	55584	2.27	21	47
ASV119	ADMI	44057	1.8	60	5
ASV153	ADMI	27699	1.13	0	31
ASV156	ADMI	27080	1.1	76	12
ASV164	ADMI	24961	1.02	62	21
ASV194	ADMI	23950	0.98	14	6
ASV214	ADMI	15464	0.63	38	18
ASV219	ADMI	20004	0.82	0	31
ASV229	ADMI	15723	0.64	26	3
ASV239	ADMI	13774	0.56	26	10
ASV253	ADMI	15776	0.64	54	6

ASV256	ADMI	12218	0.5	2	19
ASV269	ADMI	15108	0.62	35	0
ASV272	ADMI	14360	0.59	60	9
ASV286	ADMI	9761	0.4	30	3
ASV320	ADMI	8512	0.35	27	1
ASV386	ADMI	8258	0.34	21	3
ASV433	ADMI	6973	0.28	0	2
ASV452	ADMI	5802	0.24	10	0
ASV468	ADMI	5479	0.22	20	0
ASV475	ADMI	5567	0.23	2	1
ASV545	ADMI	3630	0.15	14	0
ASV556	ADMI	3903	0.16	12	1
ASV574	ADMI	2947	0.12	12	2
ASV582	ADMI	3217	0.13	4	5
ASV621	ADMI	2694	0.11	4	8
ASV636	ADMI	2676	0.11	11	0
ASV648	ADMI	3212	0.13	0	9
ASV657	ADMI	169	0.01	3	0
ASV679	ADMI	2619	0.11	5	1
ASV721	ADMI	2516	0.1	4	0
ASV730	ADMI	1978	0.08	12	0
ASV750	ADMI	2273	0.09	3	3
ASV771	ADMI	1892	0.08	0	8
ASV843	ADMI	1406	0.06	7	0
ASV852	ADMI	1321	0.05	6	0
ASV876	ADMI	1114	0.05	12	0
ASV878	ADMI	1178	0.05	3	5
ASV900	ADMI	1388	0.06	2	0
ASV956	ADMI	787	0.03	4	3
ASV1020	ADMI	885	0.04	14	3
ASV16	FSAP	236169	9.63	33	87
ASV43	FSAP	108318	4.42	44	50
ASV48	FSAP	106046	4.33	184	63

ASV74	FSAP	69560	2.84	17	78
ASV187	FSAP	24968	1.02	43	0
ASV198	FSAP	23634	0.96	15	0
ASV232	FSAP	18731	0.76	37	11
ASV233	FSAP	18589	0.76	43	0
ASV234	FSAP	16554	0.68	69	22
ASV261	FSAP	14381	0.59	66	18
ASV440	FSAP	6472	0.26	9	16
ASV445	FSAP	6343	0.26	4	25
ASV643	FSAP	3225	0.13	0	2
ASV655	FSAP	2872	0.12	10	1
ASV785	FSAP	1975	0.08	0	2
ASV810	FSAP	1779	0.07	12	1
ASV823	FSAP	1727	0.07	0	13
ASV983	FSAP	1125	0.05	0	17

Supplementary table 2. TITAN main outputs obtained for the responses analyses of ASVs from *Achnanthidium minutissimum (ADMI), Fistulifera saprophila* (FSAP)*, Nitzschia inconspicua* (NINC) and *N. soratensis* (NSTS) for calcium, conductivity, TOC sulphates and phosphates. Change point indicates the value of the environmental parameters at which the change point occurred (mg/L Calcium and µS/cm conductivity). The magnitude of the response is given by z score. The occurrence frequency of the ASVs are represented (Frequency). 5^{th} to 95^{th} percentiles indicate the frequency of distribution of change points for 500 bootstrap replicates given that an estimation of uncertainty associated to the change point. Only responses that fulfilled purity and reliability metrics (≥ 0.95) are represented.

Environmental Variable	Species ASV	Frequency	Change point	Z score	Response type	5^{th}	10^{th}	50^{th}	90^{th}	95^{th}
Calcium	ADMI ASV239	23	5.75	18.44	Negative	3.9	4.6	7.9	38.4	40
Calcium	ADMI ASV253	34	55.80	16.74	Negative	38.8	40.1	47.8	54.5	56.7
Calcium	ADMI ASV574	6	10.50	16.08	Negative	3.4	3.5	8.7	12.1	16.9
Calcium	ADMI ASV320	11	7.43	15.9	Negative	3.5	4.1	5.5	10	11.5
Calcium	ADMI ASV556	7	10.50	15.06	Negative	3.4	3.6	7.3	18.5	22
Calcium	ADMI ASV269	16	43.78	12.37	Negative	13.6	13.7	42.4	45.5	48
Calcium	ADMI ASV272	26	48.67	8.68	Negative	7.2	9.6	42.4	61.4	65.4
Calcium	ADMI ASV876	10	77.55	5.73	Negative	73.2	74	77	77.7	78.3
Calcium	ADMI ASV219	31	120.88	18.62	Positive	116.8	120	121.7	136.3	142.8
Calcium	ADMI ASV153	31	148.25	14.76	Positive	111	112.7	127.8	221.3	240.1
Calcium	ADMI ASV771	8	132.92	12.88	Positive	116	116.8	131.3	160.8	212.7
Calcium	ADMI ASV33	118	111.91	11.79	Positive	65.4	74.4	108.8	116	120

Calcium	ADMI ASV7	226	51.50	6.63	Positive	6.6	15.4	54.7	67.1	104.9
Calcium	ADMI ASV70	69	71.29	6.23	Positive	45.5	49.4	77.3	125.9	126.6
Calcium	ADMI ASV77	56	120.00	5.01	Positive	40	108.3	119.7	136.3	221.3
Calcium	ADMI ASV386	12	78.83	4.06	Positive	61.4	62.5	80.9	89.1	95
Conductivity	ADMI ASV556	13	70.19	24.53	Negative	49.5	57.1	71.5	112.1	133.4
Conductivity	ADMI ASV239	28	77.75	22.89	Negative	71.5	72.3	77.8	96.5	104.6
Conductivity	ADMI ASV574	11	130.25	17.95	Negative	62.3	65.8	86.3	130.3	134.5
Conductivity	ADMI ASV320	18	78.60	15.12	Negative	65.8	70.2	89.3	169	172.5
Conductivity	ADMI ASV272	45	306.33	14.34	Negative	252.5	255.2	301.9	339.8	352.9
Conductivity	ADMI ASV253	50	384.88	13.13	Negative	194.1	251	377.3	429	456.5
Conductivity	ADMI ASV269	27	254.50	12.62	Negative	191.1	194.6	232.4	330.8	337.8
Conductivity	ADMI ASV286	13	86.14	9.03	Negative	51.4	71.5	78.6	91.6	153.7
Conductivity	ADMI ASV33	99	342.75	16.19	Negative	314	333.4	384	480.6	537.4
Conductivity	ADMI ASV219	28	704.17	15.46	Positive	586	619.9	651.5	727	768.3
Conductivity	ADMI ASV153	26	648.83	10.53	Positive	520.5	608.3	993.5	2557.8	2557.8
Conductivity	ADMI ASV70	59	729.00	10.45	Positive	337.8	362	708	747.6	787.7
Conductivity	ADMI ASV771	7	997.50	8.97	Positive	707.7	728.8	931.3	1064.6	1318.5
Conductivity	ADMI ASV7	176	309.75	8.4	Positive	297.6	314	491.8	721	751.3
Conductivity	ADMI ASV6	172	362.00	7.11	Positive	307.2	309.8	335.1	362	382.9

Conductivity	ADMI ASV77	50	384.88	5.98	Positive	307.6	311.5	362	474.5	615.1
Conductivity	ADMI ASV156	31	384.88	5.89	Positive	266.3	267	377.7	414.3	421.7
Conductivity	ADMI ASV621	9	384.88	4.39	Positive	342	361.3	422.2	653	670
Conductivity	ADMI ASV256	19	295.21	3.55	Positive	289.9	292.6	475.5	2811.5	2811.5
Phosphates	FSAP ASV234	91	0.02	4.33	Negative	0.01	0.01	0.02	0.21	0.40
Phosphates	FSAP ASV655	11	0.14	4.44	Negative	0.07	0.09	0.11	0.13	0.13
Phosphates	FSAP ASV16	116	0.24	21.02	Positive	0.16	0.19	0.29	0.37	0.39
Phosphates	FSAP ASV232	47	0.11	7.83	Positive	0.10	0.11	0.17	0.24	0.26
Phosphates	FSAP ASV261	80	0.05	4.76	Positive	0.04	0.04	0.10	0.17	0.31
Phosphates	FSAP ASV43	91	0.39	12.41	Positive	0.08	0.10	0.30	0.44	0.60
Phosphates	FSAP ASV440	23	0.10	5.48	Positive	0.07	0.07	0.09	0.10	0.13
Phosphates	FSAP ASV445	24	0.68	11.51	Positive	0.10	0.31	0.66	1.40	1.46
Phosphates	FSAP ASV48	241	0.13	9.76	Positive	0.07	0.07	0.13	0.22	0.22
Phosphates	FSAP ASV74	90	0.50	25.78	Positive	0.31	0.32	0.47	0.54	0.58
Phosphates	FSAP ASV823	13	0.92	19.57	Positive	0.34	0.50	0.89	1.84	2.26
Phosphates	FSAP ASV983	17	1.46	16.04	Positive	0.16	0.17	1.47	2.64	3.35
Sulphates	FSAP ASV234	60	9.80	4.98	Negative	6.01	7.45	10.10	33.87	56.79
Sulphates	FSAP ASV655	7	7.77	8.70	Negative	2.75	2.93	7.05	10.10	26.60
Sulphates	FSAP ASV16	98	61.50	18.32	Positive	55.00	57.67	61.50	75.00	81.53

Sulphates	FSAP ASV43	72	88.50	5.13	Positive	15.52	20.25	39.03	90.35	128.17
Sulphates	FSAP ASV445	24	452.00	10.14	Positive	38.00	60.75	423.50	482.25	498.25
Sulphates	FSAP ASV48	123	4.00	3.18	Positive	4.10	4.39	15.18	296.25	408.75
Sulphates	FSAP ASV74	83	55.25	16.21	Positive	45.80	47.00	54.00	74.67	81.53
Sulphates	FSAP ASV823	13	84.50	9.20	Positive	82.67	84.00	93.67	125.50	133.76
Sulphates	FSAP ASV983	17	94.33	9.54	Positive	41.57	55.82	102.00	142.07	344.50
TOC	FSAP ASV234	90	1.82	11.41	Negative	1.54	1.58	1.82	3.00	3.20
TOC	FSAP ASV655	11	1.80	4.20	Negative	0.60	1.10	1.68	2.95	3.00
TOC	FSAP ASV810	13	1.80	5.37	Negative	0.69	0.70	1.70	1.80	1.86
TOC	FSAP ASV16	115	4.90	14.09	Positive	2.60	2.90	4.90	5.05	5.23
TOC	FSAP ASV232	47	1.70	5.37	Positive	1.55	1.60	1.78	2.56	4.49
TOC	FSAP ASV233	43	1.52	5.03	Positive	1.48	1.50	1.60	3.78	4.02
TOC	FSAP ASV445	23	2.88	8.57	Positive	2.58	2.60	3.14	5.60	7.14
TOC	FSAP ASV48	240	2.20	12.36	Positive	1.60	1.66	2.15	2.40	2.56
TOC	FSAP ASV74	89	4.90	16.07	Positive	2.40	2.56	4.90	5.40	5.50
TOC	FSAP ASV823	13	7.48	11.16	Positive	2.90	2.95	5.10	7.14	7.48
TOC	FSAP ASV983	17	5.90	10.83	Positive	2.75	2.80	5.95	6.70	7.10
Calcium	NINC ASV53	119	88.77	10.21	Positive	81.80	86.84	92.00	104.79	121.17
Calcium	NINC ASV56	60	89.00	12.96	Positive	74.50	85.84	89.10	97.26	122.40

Calcium	NINC ASV113	105	98.20	18.39	Positive	78.83	82.30	97.23	102.72	108.03
Calcium	NINC ASV249	7	20.27	3.24	Positive	18.00	18.48	23.10	94.29	101.04
Calcium	NINC ASV273	24	93.83	3.45	Positive	47.90	74.20	96.40	263.83	263.83
Calcium	NINC ASV457	10	243.17	6.04	Positive	71.10	82.32	244.00	263.83	288.50
Calcium	NINC ASV463	9	338.50	2.18	Positive	51.85	61.20	102.39	381.60	428.00
Calcium	NINC ASV615	8	338.50	7.16	Positive	70.20	74.18	239.88	338.50	356.00
Calcium	NSTS ASV94	46	70.10	12.63	Negative	48.00	53.43	57.68	67.34	68.00
Calcium	NSTS ASV117	60	56.42	21.36	Negative	42.99	45.46	54.35	63.00	65.40
Calcium	NSTS ASV288	16	37.25	12.96	Negative	9.94	13.00	37.00	54.27	56.85

Chapter 5

Phylogeographical patterns in freshwater diatoms revealed by DNA metabarcoding of a short *rbcL* marker

Pérez-Burillo, J., Trobajo, R., Mann, D.G.

In preparation

1.Introduction

The study of genetic variation in diatom populations has greatly increased our understanding of diatom biology. Particularly, these studies have broadened our knowledge about genetic diversity structure and connectivity of populations, evolutionary processes within species and populations and, speciation mechanisms (e.g. Casteleyn et al., 2009; Evans et el., 2009; Godhe & Härnström, 2010; Vanormelingen et al., 2015; Van den Wyngaert et al., 2015). Genetic information of diatom populations has traditionally been evaluated using different tools and markers such as amplified fragment length polymorphism (ALFP), microsatellites or Sanger sequencing. Despite the many valuable insights into diatom biology revealed by these studies, a common drawback of these approaches is the high time and effort required to reach a sampling size large enough to adequately cover the genetic diversity of the species. This is especially evident when dealing with rare species.

The arrival of Next-generation sequencing technologies has overcome in somehow this limitation since genetic information from multiple species and a large number of samples can be evaluated at a fraction of the cost and time demanded by traditional approaches (Dufresne et al., 2014). In particular, the cost of sequencing has been significantly reduced with the advent of Illumina technologies compared to previous sequencing platforms (454 Roche GS FLX System), which has significantly improved the affordability of these technologies (Reuter et al., 2015). Nevertheless, the impossibility of relating sequencing reads to individuals and the low phylogenetic resolution of the short markers used are some of the major factors that reduce the possibilities achievable via metabarcoding at the population level. Despite these limitations, metabarcoding can still be considered a complementary tool able to provide valuable insights into the genetic structure of diatom populations as some studies have already shown, for example in the case of marine diatom planktonic species using the V4 rRNA marker (De Luca et al., 2021; Ruggiero et al., 2022).

A crucial step for accurately measuring genetic diversity via metabarcoding is to identify and discard PCR and sequencing artefacts. For this aim, bioinformatic pipelines based on sequencing denoising algorithms, such as DADA2 (Callahan et al., 2016), have been demonstrated to be particularly efficient for separating real genetic variants from artefacts (e.g. Macé et al., 2022; Tsuji et al., 2019). Nevertheless, chapters 2, 3 and 4 evidenced that these algorithms still are subjected to errors as some ASVs denoted as real by DADA2 was very likely artefacts due to the presence of stop codons in their amino-acids sequences. This evidence the need for further analyses, such as

phylogenetic based analyses, for ensuring the reliability of genetic variants inferred by bioinformatics pipelines.

Once genetic variants have been successfully validated as real, reliable haplotype frequency data can be extracted which can provide comprehensive coverage of the genetic diversity of a large number of taxa. *In silico* analyses of large-scale haplotype frequency data derived from metabarcoding studies of different ecosystems and geographic areas can provide new insights into the phylogeography of species (Burki et al., 2021; Turron et al., 2020). Numerous diatom metabarcoding datasets have recently become publicly available and thus provide a good opportunity to study aspects of diatom genetic diversity and phylogeography about which very little is known.

As exemplified from our studies in freshwater environments (detailed in chapters 3 and 4), it seems that benthic diatom species differ greatly in the number of *rbcL* variants, with some species represented by a high number of *rbcL* variants whereas others are reduced to only 1 or a few. In addition, our analyses in chapter 4 clearly evidenced that at the regional scale, different patterns in the genetic structure (i.e. phylogeographic patterns) of the *rbcL* marker were perceived among the species analyzed. Thus, *Achnanthidium minutissimum* and *Fistulifera saprophila* showed a large number of *rbcL* variants widely but differently distributed in French and Catalan rivers and moreover, within each species, certain variants clearly differed in their environmental preferences. By contrast, *Nitzschia soratensis* showed a lower number of ASVs compared to the previous two complexes, with a more restricted distribution to certain regions and similar preferences for environmental conditions.

Thus, this study aimed firstly to characterize the intraspecific diversity (based on the short 263-bp *rbcL* diatom marker) of multiple species using a large metabarcoding dataset covering different and well-separated regions. By doing this we also describe the phylogeographic patterns observed for each species in order to characterise common types of patterns among the species analysed. To our knowledge, there is not a study that has attempted to characterize the diversity of the short *rbcL* marker used for metabarcoding of freshwater diatom species and this might be relevant from both an ecological and applied perspective. Thus, characterization of *rbcL* diversity and phylogeography of diatom species could shed light on some traditional ecological questions such as the amount of cosmopolitan diversity in microeukaryotic communities (e.g. Finlay et al., 2002; Finlay & Fenchel, 2004). In addition, these data can inform on patterns of dominance among genetic variants within species that ultimately reflect which genetic variants are playing a higher role in ecosystem functions. On the other hand, this

information is relevant in a biomonitoring context, as demonstrated in Chapter 4, genetic variants within the species complex may differ in their ecological preferences, therefore, mapping genetic variants occurring in a given hydrogeographic region tell us whether variants with or without the same ecological profiles are expected to be found, potentially making future biomonitoring campaigns more effective. Overall, this is a first exploratory attempt that will be useful and supportive for future studies based on more efficient technologies (such as long sequencing technologies) that will be able to provide a better characterisation of the genetic diversity of diatoms.

Second, aiming to understand the cause and significance of the differences observed in the intraspecific diversity among species, we correlated the *rbcL* diversity observed with several diatom traits:

a) *RbcL* diversity between pennate and centric species was compared as it has been shown that the rate of diversification of diatom species showing an oogamous reproductive mode (i.e. only reported in centric species) is lower than those with an isogamous mode (i.e. mainly observed in pennate species) (Nakov et al., 2018). In addition, the symmetry of valves could be reflected into a higher or lower diversity of the *rbcL* since plastid inheritance patterns differ between centric and pennate species. Thus, a uniparental and maternal inheritance has been reported for most of the centric diatoms studied, while chloroplast in pennate species seems to be inherited exclusively biparentally (Mann, 1996; Round et al., 1990). The pattern of inheritance (biparental vs uniparental) can be expected to affect the time for fixation of chloroplast genes in populations and hence influence the diversity of *rbcL* present within species.

b) Other aspects of chloroplasts (i.e. number and shape) were assessed as they characterise diatom groups and have been used to support taxonomic revisions (Sims et al., 2006 and reference therein). Because of their importance in the separation of diatoms, these aspects could be correlated with differences in the number of *rbcL* variants. It must be mentioned that the number of chloroplasts is greatly related to diatom symmetry as centric diatoms often have numerous chloroplasts per cell whereas pennate species often show 1, 2 or 4 chloroplasts (Mann 1996).

c) Finally, the ecological guilds of diatom species (i.e. low profile, high profile, euplanctonic and motile guilds) and their motility (motile vs non-motile species) were correlated with genetic diversity observed among species. In this regard,

Nakov et al. (2018) found that the rate of diversification has been higher in motile species than in non-motile diatoms likely because motility provides higher potential and capabilities for exploiting new habitats and reproducing (sexually) more efficiently which ultimately is reflected in a higher genetic diversity.

2. Material and Methods

2.1 Data collection

We conducted a comprehensive search during May and June of 2021 for diatom metabarcoding data based on the *rbcL* marker. This search was conducted in two public online repositories: 1) the Sequence Read Archive (SRA) repository of high-throughput sequencing data provided by NCBI and 2) the open-access research data repository Zenodo. In both repositories, the set of keywords used sought to cover any metabarcoding research that used the *rbcL* gene and diatoms were one of the target organisms. The search was not limited to any specific time period. We used the following different keywords, in different combinations, for the search: diatom, *rbcl*, metabarcoding, microeukaryotic and microalgae.

The results provided by this initial search were carefully and manually screened to find all possible studies that met our criteria. Following this examination, we were able to identify 9 *rbcL* metabarcoding datasets (7 deposited at SRA and 2 at Zenodo). These 9 datasets cover regions in North America (California, Ohio [Smucker et al., 2020] and Ontario [Maitland et al., 2020]), Europe (Fennoscandia [Bailet et al., 2020], France [Tardy et al., 2021] and Spain [Nistal-García et al., 2021]), Asia (Tibet [Kang et al., 2021]) and the Indian Ocean (Mayotte [Vasselon et al., 2017]) (Table 1). In addition to these 9 datasets, we include in our study the other datasets used in previous chapters (1, 3 and 4), which come from routine WFD biomonitoring programmes in rivers in the UK, France and Catalonia (NE Spain) (see detailed information in chapters 1, 3 and 4). Most of these datasets are derived from river communities, but a few of them include samples from lakes (Table 1). Finally, all of these datasets constituted the data analyzed in this study and were based on several types of DNA extraction kits, sequencing technologies and diatom *rbcL* markers (i.e. 331-bp or 263-bp) (Table 1).

2.2. Bioinformatic analyses and data merging

2.2.1 ASVs inference through DADA2 pipelines

Bioinformatics analyses were conducted on the forward (R1) and reverse (R2) reads from the different datasets to infer Amplicon Sequence Variants (ASVs), which constituted the fundamental units on which further examinations were carried out. ASVs were generated using the R package DADA2 (Callahan et al., 2016) and the different datasets were analyzed separately. When a dataset was formed by more than 1 Illumina run, each of the runs was also analysed individually. The first step conducted in the bioinformatics pipeline was to remove the primers from the raw R1 and R2 reads. For this, we used cutadapt (Martin, 2011) to specifically identify and remove the different sets of primers used to generate the metabarcoding sequences. Note that the 2 datasets derived from PGM Ion Torrent (see Table 1) were constituted by single-end reads and therefore cutadapt was applied in each file to remove both forward and reverse primers. Then, the resulting R1 and R2 reads (or single-end reads in the case of data derived from PGM Ion Torrent) were truncated to approximately 220–240 and 160–200 nucleotides respectively, based on their quality profiles (median quality score ≥ 30). After this truncation step, reads with ambiguities or showing an expected error (maxEE) higher than 2 were discarded. The DADA2 denoising algorithm was then applied to determine an error rates model in order to infer amplicon sequence variants (ASVs). Finally, ASVs detected as chimeras were identified and discarded using the function “removeBimeraDenovo”.

2.2.2 ASVs information merged according to the marker

Once the ASVs had been inferred from each dataset analysed, all the sequence tables based on the same marker (263-bp or 331-bp) were firstly truncated to the corresponding assumed marker length (i.e. 263-bp or 331-bp) using Mothur (Schloss et al., 2009) and then merged with the DADA2 function "mergeSequenceTables". This allowed us to homogenise the information (abundance and sequence identity of the ASVs) of the different sequence tables obtained, thus avoiding having two or more ASVs identical in their *rbcL* sequence over the shared region but labelled differently. By doing this, we obtained two merged datasets, one containing all the ASVs based on the 263-bp marker and another set with all those based on the 331-bp marker.

2.2.3 ASVs information merged according to the 263-bp shared region

The next step was to merge the ASVs from these two datasets (263-bp and 331-bp ASVs datasets), as the region of interest was the shared 263-bp region of both markers (intraspecific variation occurs in the 68-bp tail of the 331-bp marker, as shown in chapter 3 but obviously cannot be compared across datasets using the 263-bp marker). For this, the Basic Local Alignment Search Tool (BLAST) was used to compare all 263-bp ASVs with their 331-bp counterparts, in order to identify those ASVs from both datasets that were identical in the 263-bp region. In this analysis, we considered ASVs to be identical for the 263-bp region (i.e. synonymous) if, in this region, they showed a pairwise-alignment with 100% similarity, no gaps and mismatches, and a full cover of the query sequence (i.e. ASVs based on 263-bp). Next, a common identification name was provided for those ASVs identified as synonymous and abundance sequencing tables were merged accordingly. Thus, the resulting merged dataset contained all the unique 263-bp ASVs inferred from all the original datasets. Finally, a preliminary taxonomic classification for these ASVs was given by the naïve Bayesian classifier method (Wang et al., 2007) and the reference library Diat.barcode v10 (Rimet et al., 2019). In this regard, the naïve Bayesian classifier is based on sequence similarity and does not take into account phylogenetic relationships, which are necessary to successfully provide a reliable taxonomy for the different ASVs obtained (described in more detail in section 2.4).

2.3. Species selection

Since the genetic variants used in our study required a thorough pre-validation process (see section 2.4), it was impractical to use all the potentially detected species in our dataset (i.e. 504 firstly identified by the naïve bayesian classifier) and therefore we selected a subset of the total species. The selection of the subset of species was mainly aimed at achieving a sufficient representation of species with different diatom traits, since one of the objectives of this study was to correlate these different traits with the genetic diversity of the *rbcL* marker. To make such a selection, the number of centric and pennate species was first evaluated and, since the number of centric species was very limited (only 28), we decided to keep all centric species identified by the naïve Bayesian classifier. Among the pennate species, the selection was made to achieve a balance In which there was a sufficient number of representative species for each of the following traits: number of chloroplasts, chloroplast shape, biovolume and ecological guilds. In addition, we avoided selecting those rare species that were represented scarcely represented in our dataset (i.e. < 100 reads and < 4 samples). Based on these criteria,

we selected a total of 46 pennate species (table 2). Centric and pennate selected species comprised a total number of 74 species and are hereinafter referred to as target species.

2.4 Phylogenetic analyses for validation of ASVs

Phylogenetic analyses were performed in order to try to recover all possible genetic variants of the selected species. Such analyses were performed separately for each target species and each included 1) all available reference sequences of the target species deposited in Diat.barcode v10, 2) all inferred ASVs in our dataset that shared at least 95% nucleotide similarity with the reference sequences of the target species (step 1). [Note that the nucleotide similarity data were obtained by BLAST analyses that compared all inferred ASVs in our dataset with all reference sequences included in Diat.barcode v10.] 3) All reference sequences available in Diat.barcode v10 that shared at least 95% similarity with the ASVs previously selected in step 2 were also included. The latter was done to include species closely related to the target species and thus increase the robustness of the phylogenetic analysis.

All the phylogenetic trees evaluated were performed using raxmlGUI with the GRT-Gamma model (Silvestro & Michalak, 2012) and with 500 replicates for the bootstrap analyses. Note that though bootstrap values were useful for checking the robustness of the different clades, our validation of ASVs (detailed in following paragraph) was based on the topology of the tree regardless of the bootstrap support values the clades received. All the trees were previously aligned using the Muscle alignment algorithm (Edgar, 2004) and they were visualized using iTOL (https://itol.embl.de) (Letunic & Bork, 2019).

The trees generated were carefully examined to elucidate the phylogeny of the different ASVs. Validation of the ASVs was relatively straightforward for some species that were clearly monophyletic groups on the basis of reference sequences (e.g. *Aulacoseira granulata*, *Melosira nummuloides*, *Halamphora veneta*), whereas it was more complex and difficult for those species that were paraphyletic groups (mainly in *Achnanthidium minutissimum*, *Fistulifera saprophila* and *Amphora pediculus*). In the latter case, the criterion followed was to consider as reliable ASVs those variants of a given species that were located in subclades defined only by reference sequences of the target species. However, some ASVs that were distributed in different clades formed only by ASVs, i.e. without any reference sequence of any species, were also considered valid variants due to their close phylogenetic proximity to other subclades defined by reference sequences of the target species. Although the selection of ‘valid’ ASVs was performed consistently according to the criteria outlined, this process could introduce some subjectivity in our

analyses, especially in the case of paraphyletic species. Furthermore, because the validation process for ASVs from species complexes does not consider all the clades from the complex but just those related to reference sequences with the same name as the targeted morphospecies, certainly a small amount of genetic diversity from these complexes was artificially removed. *Achnanthidium minutissimum* is perhaps the most relevant case as the only ASVs we considered to be valid for analysis were those closely related to '*A. minutissimum*' reference sequences and we avoided considering those that appeared closely related (in our phylogenies) to reference sequences from *A. digitatum*, *A. saprophilum*, *A. jackii, A. eutrophilum* and A. *lineare*. However, the larger proportion of ASVs within this complex was related to *A. minutissimum* reference sequences and only a small number were related to the other species listed above. Finally, in order to remove likely HTS artefacts, we did not consider as valid ASVs those that showed stop codons and/or were recorded with less than 10 reads or in less than 2 samples.

2.5 Traits comparison and haplotype networks

The different traits evaluated were diatom stria pattern and symmetry (i.e. centric vs pennate), number of chloroplasts, chloroplast shape, size class, ecological guild, and motility (Table 2). For each of the 74 species analyzed, information about these traits was extracted from Diat.barcode v10.

Kruskal–Wallis (Hollander and Wolfe, 1973) tests with post hoc Dunn's test (Dunn, 1964) were performed to determine whether the number of valid ASVs per species differed statistically ($p < 0.05$) among the different traits. Note that since some of the chloroplast shapes were represented by only 1 or 2 cases, only those categories with more than 3 representatives were included in the statistical analyses (i.e. *elongate*, *plate, H-shaped* and *discoid*) (Table 2). For the same reason, the categories included in statistical analyses regarding the chloroplast number were 1, 2 and nb (i.e. numerous).

In order to evaluate the phylogeography of the different species across the regions surveyed, haplotype networks based on the TCS algorithm (Clement et al. 2002) were constructed individually for each species surveyed and using the corresponding validated ASVs and their occurrence (presence/absence data). Haplotype networks were performed and visualized using PopART software (Leigh and Bryant, 2015).

3.Results

3.1. Number of ASVs per species and relation with diatom traits.

Statistical analyses comparing ASV numbers with diatom traits indicated that the number of ASVs per species differs significantly between pennate and centric diatoms (p = 0.0018). The average and standard deviation of the number of ASVs per species were 2.25 and 3 for centric species and 6.40 and 10.97 for pennate diatoms (Table 3). Kruskal–Wallis test indicated statistically significant differences between life-form and the number of ASVs but the post hoc Dunn's test of multiple comparisons did not find statistical differences ($p < 0.05$), being found in the comparison between euplanctonic and low-profile the highest differences (p = 0.78). Both Kruskal–Wallis test and Dunn's test indicated statistically significant differences between the number of chloroplasts and the number of ASVs; in particular, Dunn's test indicated significant differences in ASVs between species with 2 and multiple chloroplasts (p = 0.036). The average and standard deviation of the number of ASVs per species were 7.21 and 7 for species with 2 chloroplasts and 2.71 and 3.21 for species with multiple chloroplasts (Table 3). Regarding the comparison between the ASV and chloroplast number, it must be noted that all the species analyzed in this study with 2 chloroplasts were pennate whereas all but 2 of the species analyzed with multiple chloroplasts constituted centric diatoms. The number of ASVs per species was close to being significantly different among chloroplast shapes ($p = 0.056$) and between mobile and non-mobile species ($p = 0.052$)

3.2 Phylogeographical patterns

Through the examination of the phylogeography of the different species shown by haplotype networks we could identify at least 4 types of phylogeographic patterns (Table 4; Annex 1):

Type I: This pattern was represented by species showing only 1 ASV. This pattern was most frequently observed in centric species (18 out of 28 species examined) but was also well represented in pennate species (14 out of 44 species assessed). Within this category, the single ASV observed for each species could be widely distributed in several regions (e.g. *Navicula tripunctata*) or restricted to a single region with little occurrence (i.e. rare ASVs) (e.g. *Gomphonema rosenstockianum*).

Type II: A second pattern was shown by those species with more than 2 ASVs in which 1 or 2 ASVs clearly predominated over the others both in occurrence and in the number of regions where they were distributed. An important feature of this type of pattern was the presence of rare ASVs. The number of rare ASVs could be only 1 (e.g. *Pinnularia*

neomajor) or several (e.g. *Melosira varians*). The exceptions were *Surirella solea* and *Aulacoseira granulata*. The former did not show rare ASVs restricted to only 1 region, but it was included in this pattern because of the clear predominance in terms of occurrence of 1 ASV over the other. By contrast, *A. granulata* shows 2 ASVs with very similar occurrences. However, 1 of them clearly dominated in terms of the number of regions where it was distributed.

Type III: A third pattern was characterized by the presence of 2 to 3 ASVs per species, with none of them apparently dominating over the others and the absence of rare ASVs. Although within some species the presence and distribution of certain ASVs may be greater than of other ASVs (e.g. Cyclotella cryptica), these did not constitute as clear patterns of dominance as those observed in type II species. Basically, the characteristics of this pattern can be regarded as the opposite of type 2.

Type IV: This pattern was represented by species consisting of a high number of ASVs (minimum observed in a species were 7 ASVs) of which a high proportion were recorded with a high occurrence and a wide distribution, making it difficult to draw strong conclusions about predominance patterns. In addition, species with this pattern have rare ASVs but in all the cases these constituted a small proportion of the total number identified per species.

Two exceptions were *Eunotia bilunaris* and *E. minor*, which did not fit into any of the categories previously defined. These two species were characterized by medium-high genetic diversity but no patterns of dominance were observed among their ASVs.

4.Discussion

The discussion of this chapter is included in section 2.1 of the general discussion of this thesis.

Table 1. Metadata of the different metabarcoding datasets used in this study

Repository id where data is available	Doi reference study	No. Samples	Sequencing platform	RbcL Marker	Location	Extraction method used
ERP124785	https://doi.org/10.1016/j.jhazmat.2021.125121	48	Illumina MiSeq	263	WWTP effluents, France.	FastDNA Spin Kit
SRP291163	https://doi.org/10.1371/journal.pone.0242143	45	Illumina MiSeq	263	Rivers, Ontario Canada	DNeasy PowerSoil kit
SRP234514	https://doi.org/10.1002/eap.2205	342	Illumina MiSeq	263	Rivers, Ohio, USA	DNeasy PowerLyzer PowerSoil Kit
SRP217406	https://doi.org/10.1016/j.ecolind.2020.107070	23	Illumina MiSeq	331	Lake Nam Co,Tibet.	DNeasy PowerMax Soil Kit (wet samples) / DNeasy PowerSoil Kit (sedsamples)
SRP290705	No study found - Data collected by southern california water research project	85	Illumina MiSeq	263	Small rivers, California, USA	No info
SRP255509	https://doi.org/10.1016/j.scitotenv.2021.147410.	22	Illumina MiSeq	263	Small ponds, Leon, Spain.	PowerSoil DNA Isolation Kit
https://zenodo.org/record/3885810#.YIbPJqHtaUk	https://doi.org/10.1016/j.scitotenv.2020.140948	48	Illumina MiSeq	263	Rivers and lakes, Fennoscandia	NucleoSpin Soil kit (MN-Soil)
https://zenodo.org/record/400160#.YK0ONaHtaUk	https://doi.org/10.1016/j.ecolind.2017.06.024	80	PGM Ion Torrent	263	Rivers, Mayotte	Own method
https://zenodo.org/record/1157865#.Yo9c51RByUk	https://doi.org/10.1007/s13127-018-0359-5	156	PGM Ion Torrent	263	Lakes, France	GenElute TM-LPA
Own data	https://doi.org/10.1016/j.scitotenv.2020.138445	307	Illumina MiSeq	263	Rivers, Catalonia, Spain	NucleoSpin Soil kit (MN-Soil)
Data available at https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/9EG5Z4&version=1.1	https://doi.org/10.1016/j.ecolind.2019.105775	447	Illumina MiSeq	263	Rivers, France.	GenElute TM-LPA / NucleoSpin Soil kit (MN-Soil)
Data supplied by Dr Kerry Walsh (UK Environment Agency)	https://doi.org/10.1016/j.ecolind.2020.106725	1714	Illumina MiSeq	331	Rivers, UK	DNeasy Blood and Tissue kit

Table 2 List of species used for this study including information on the number of ASVs, traits and inferred phylogeographic pattern type for each species. Information about different diatom traits was extracted from Diat.barcode v10. Note that * indicates those categories of specific traits that were not included in statistical analyses due to their low representation (i.e.N < 4).

Species	Taxonomic class	No. ASVs	Phylogeography type	Diatom symmetry	No. Chloroplast	Chloroplast shape	Ecological guild	Mobility
Achnanthidium minutissimum	Bacillariophyceae	70	IV	Pennate	1	plate	Low profile	Yes
Fistulifera saprophila	Bacillariophyceae	24	IV	Pennate	2	plate	Motile guild	Yes
Nitzschia palea	Bacillariophyceae	22	IV	Pennate	2	plate	Motile guild	Yes
Amphora pediculus	Bacillariophyceae	16	IV	Pennate	1	H-shape	Low profile	Yes
Ulnaria ulna	Fragilariophyceae	15	II	Pennate	2	ribbon*	High profile	No
Cyclotella meneghiniana	Mediophyceae	13	IV	Centric	Numerous	discoid	Euplanctonic	No
Rhoicosphenia abbreviata	Bacillariophyceae	13	II	Pennate	1	H-shape	Low profile	No
Melosira varians	Coscinodiscophyceae	12	II	Centric	Numerous	lobed, small plate-like*	High profile	No
Navicula lanceolata	Bacillariophyceae	12	II	Pennate	2	plate	Motile guild	Yes
Eunotia bilunaris	Bacillariophyceae	10	-	Pennate	2	elongate	High profile	Yes
Fragilaria gracilis	Fragilariophyceae	9	IV	Pennate	2	plate	High profile	No
Nitzschia inconspicua	Bacillariophyceae	9	IV	Pennate	2	plate	Motile guild	Yes
Navicula cryptocephala	Bacillariophyceae	8	IV	Pennate	2	plate	Motile guild	Yes
Encyonema minutum	Bacillariophyceae	8	II	Pennate	1	H-shape	High profile	Yes
Nitzschia fonticola	Bacillariophyceae	7	IV	Pennate	2	plate	Motile guild	Yes
Reimeria sinuata	Bacillariophyceae	7	II	Pennate	1	much-lobes*	Low profile	Yes

Sellaphora saugerresii	Bacillariophyceae	7	II	Pennate	1	H-shape	Motile guild	Yes
Eunotia minor	Bacillariophyceae	6	-	Pennate	2	elongate	High profile	Yes
Tabellaria flocculosa	Fragilariophyceae	5	II	Pennate	Numerous	strip-like*	High profile	No
Conticribra weissflogii	Mediophyceae	4	III	Centric	Numerous	discoid	Euplanctonic	No
Halamphora veneta	Bacillariophyceae	4	II	Pennate	1	H-shape	Low profile	Yes
Diatoma vulgaris	Fragilariophyceae	3	II	Pennate	Numerous	flat, divided into 2 lobes*	High profile	No
Aulacoseira subarctica	Coscinodiscophyceae	3	II	Centric	Numerous	discoid	Euplanctonic	No
Stephanodiscus hantzschii	Mediophyceae	3	II	Centric	Numerous	discoid	Euplanctonic	No
Denticula tenuis	Bacillariophyceae	3	III	Pennate	2	plate	Motile guild	Yes
Cymbella excisa	Bacillariophyceae	3	II	Pennate	1	H-shape	High profile	Yes
Parlibellus protracta	Bacillariophyceae	3	III	Pennate			High profile	Yes
Hydrosera sp.	Mediophyceae	2	III	Centric	Numerous	elliptical platelets*	No	
Aulacoseira granulata	Coscinodiscophyceae	2	II	Centric	Numerous	discoid	Euplanctonic	No
Cyclostephanos invisitatus	Mediophyceae	2	II	Centric	Numerous	discoid	Euplanctonic	No
Cyclotella cryptica	Mediophyceae	2	III	Centric	Numerous	discoid	Euplanctonic	No
Pleurosira laevis	Mediophyceae	2	II	Centric	Numerous	discoid	High profile	No
Eunotia arcus	Bacillariophyceae	2	II	Pennate	2	elongate	High profile	Yes
Craticula subminuscula	Bacillariophyceae	2	III	Pennate	2	plate	Motile guild	Yes
Pinnularia neomajor	Bacillariophyceae	2	II	Pennate	2	plate	Motile guild	Yes
Navicula radiosa	Bacillariophyceae	2	III	Pennate	2	plate	Motile guild	Yes
Surirella solea	Bacillariophyceae	2	II	Pennate	1	lamellar*	Motile guild	Yes
Epithemia turgida	Bacillariophyceae	2	II	Pennate	1	large, plate*	Motile guild	Yes

Gomphonella olivacea	Bacillariophyceae	2	III	Pennate	1	H-shape		No
Cymbella lanceolata	Bacillariophyceae	2	II	Pennate	1	H-shape	High profile	Yes
Cymbella aspera	Bacillariophyceae	2	III	Pennate	1	H-shape	High profile	Yes
Planothidium frequentissimum	Bacillariophyceae	2	III	Pennate	1	plate	Low profile	Yes
Ellerbeckia sp.	Coscinodiscophyceae	1	I	Centric	Numerous	small, discoid*	High profile	No
Melosira nummuloides	Coscinodiscophyceae	1	I	Centric	Numerous	lobed, small plate*	Euplanctonic	No
Aulacoseira ambigua	Coscinodiscophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Cyclostephanos dubius	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Cyclostephanos tholiformis	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Cyclotella atomus	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Discostella pseudostelligera	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Discostella sp.	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Discostella stelligera	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Pleurosira nanjiensis	Mediophyceae	1	I	Centric	Numerous	discoid		No
Thalassiosira gessneri	Mediophyceae	1	I	Centric	Numerous	discoid	Euplanctonic	No
Urosolenia eriensis	Coscinodiscophyceae	1	I	Centric	Numerous	discoid		No
Skeletonema potamos	Mediophyceae	1	I	Centric	few/cell*	disc-like or cup-shape*	Euplanctonic	No
Skeletonema subsalsum	Mediophyceae	1	I	Centric	few/cell*	disc-like or cup-shape*	Euplanctonic	No
Terpsinoe musica	Mediophyceae	1	I	Centric			High profile	No
Eunotia pectinalis	Bacillariophyceae	1	I	Pennate	2	elongate	High profile	Yes
Fragilaria perminuta	Fragilariophyceae	1	I	Pennate	2	plate	High profile	No
Navicula tripunctata	Bacillariophyceae	1	I	Pennate	2	plate	Motile guild	Yes

Nitzschia frustulum	Bacillariophyceae	1	I	Pennate	2	plate	Motile guild	Yes
Cocconeis pediculus	Bacillariophyceae	1	I	Pennate	1	C-shape*	Low profile	Yes
Epithemia gibba	Bacillariophyceae	1	I	Pennate	1	large, plate*	Motile guild	Yes
Luticola goeppertiana	Bacillariophyceae	1	I	Pennate	1	2 lobes*	Motile guild	Yes
Surirella ovalis	Bacillariophyceae	1	I	Pennate	1	lobed, plate*	Motile guild	Yes
Gomphonella olivaceolacuum	Bacillariophyceae	1	I	Pennate	1	H-shape		No
Encyonopsis minuta	Bacillariophyceae	1	I	Pennate	1	H-shape	Low profile	Yes
Gomphonema bourbonense	Bacillariophyceae	1	I	Pennate	1	H-shape	High profile	Yes
Gomphonema rosenstockianum	Bacillariophyceae	1	I	Pennate	1	H-shape	High profile	Yes
Gomphonema truncatum	Bacillariophyceae	1	I	Pennate	1	H-shape	High profile	Yes
Sellaphora capitata	Bacillariophyceae	1	I	Pennate	1	H-shape	Motile guild	Yes
Discostella nipponica	Mediophyceae	1	I	Centric				No
Discostella woltereckii	Mediophyceae	1	I	Centric			Euplanctonic	No

Table 3. Range, average and standard deviation of the number of ASVs per species in the different categories of diatom traits analyzed. * Indicate categories within a trait with statistically significant differences found in the number of ASVs per species.

Diatom trait	Category	No. species within category	Range No. ASVs per species within category	Average and Standard No. ASVs per species within category
Diatom symmetry	Centric*	28	1-13	2.25 ± 3
Diatom symmetry	Pennate*	46	1-70	6.48 ± 11.08
Ecological guild	Euplanctonic	19	1-13	2.16 ± 2.77
Ecological guild	High profile	21	1-15	4.24 ± 4.17
Ecological guild	Low profile	8	1-70	15.25 ± 23.21
Ecological guild	Motile	19	1-24	5.68 ± 6.94
Chloroplast shape	Discoid	18	1-13	2.28 ± 2.82
Chloroplast shape	Elongate	4	1-10	4.75 ± 4.11
Chloroplast shape	H-shape	15	1-16	4.2 ± 4.75
Chloroplast shape	Plate	16	1-70	10.94 ± 17.31
No. chloroplasts	1	24	1-70	6.25 ± 14.16
No. chloroplasts	2*	19	1-24	7.21 ± 7
No. chloroplasts	Numerous*	24	1-13	2.71 ± 3.21
Motility	No	35	1-15	3.17 ± 4.01
Motility	Yes	38	1-70	6.55 ± 11.98

Table 4. Summary of the different characteristics that defined the four phylogeographic patterns observed among the species analyzed.

Pattern	No. of species	No. ASVs per specie	Dominance of 1-2 ASVs	Presence of rare ASVs
Type I	32	**1**	**-**	**Yes**
Type II	20	**≥2**	**Yes**	**Yes**
Type III	10	**2-3**	**No**	**No**
Type IV	9	**≥7**	**No**	**Yes**

References

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

Burki, F., Sandin, M.M., Jamy, M., 2021. Diversity and ecology of protists revealed by metabarcoding. Curr. Biol. 31, R1267-R1280. https://doi.org/10.1016/j.cub.2021.07.066.

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

Casteleyn, G., Evans, K. M., Backeljau, T., D'hondt, S., Chepurnov, V. A., Sabbe, K., Vyverman, W., 2009. Lack of population genetic structuring in the marine planktonic diatom *Pseudo-nitzschia pungens* (Bacillariophyceae) in a heterogeneous area in the Southern Bight of the North Sea. Mar Biol 156, 1149-1158. https://doi.org/10.1007/s00227-009-1157-6

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

Dufresne, F., Stift, M., Vergilino, R., Mable, B.K., 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol. Ecol. 23, 40-69. https://doi.org/10.1111/mec.12581.

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

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

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, 261-273. https://doi.org/10.1078/1434-4610-00103.

Finlay, B.J., Fenchel, T., 2004. Cosmopolitan metapopulations of free-living microbial eukaryotes. Protist 155, 237-244. https://doi.org/10.1078/143446104774199619.

Godhe, A., Härnström, K., 2010. Linking the planktonic and benthic habitat: genetic structure of the marine diatom *Skeletonema marinoi*. Mol. Ecol. 19, 4478-4490. https://doi.org/10.1111/j.1365-294X.2010.04841.x.

Evans, K.M., Chepurnov, V.A., Sluiman, H.J., Thomas, S.J., Spears, B.M., Mann, D.G., 2009. Highly differentiated populations of the freshwater diatom *Sellaphora capitata* suggest limited dispersal and opportunities for allopatric speciation. Protist 160, 386-396. https://doi.org/10.1016/j.protis.2009.02.001.

Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods. 2nd ed. Wiley, New York, NY, USA.

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría Galindo, P., Vences, M., Wang, J., Schwalba, A., 2021. Diatom Metabarcoding and Microscopic Analyses from Sediment Samples at Lake Nam Co, Tibet: The Effect of Sample-Size and Bioinformatics on the Identified Communities. Ecol. Indic. 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070.

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

Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256-W259. https://doi.org/10.1093/nar/gkz239.

Macé, B., Hocdé, R., Marques, V., Guerin, P.E., Valentini, A., Arnal, V., Pellissier, L., Manel, S., 2022. Evaluating bioinformatics pipelines for population-level inference using environmental DNA. Environ. DNA 4. https://doi.org/10.1002/edn3.269.

Maitland, V.C., Robinson, C.V., Porter, T.M., Hajibabaei, M., 2020. Freshwater diatom biomonitoring through benthic kick-net metabarcoding. Plos one, 15, e0242143. https://doi.org/10.1371/journal.pone.0242143.

Mann, D.G., 1996. Chloroplast Morphology, Movements and Inheritance in Diatoms. In Chaudhary BR, Agrawal SB. (Eds.), Cytology, Genetics and Molecular Biology of Algae. SPB Academic Publishing, Amsterdam, pp 249—274

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

Nakov, T., Beaulieu, J.M., Alverson, A.J., 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol 219, 462-473. https://doi.org/10.1111/nph.15137.

Nistal-García, A., García-García, P., García-Girón, J., Borrego-Ramos, M., Blanco, S., Bécares, E., 2021. DNA metabarcoding and morphological methods show complementary patterns in the metacommunity organization of lentic epiphytic diatoms. Sci. Total Environ. 786, 147410. https://doi.org/10.1016/j.scitotenv.2021.147410

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-throughput sequencing technologies. Mol. Cell. 58, 586-597. https://doi.org/10.1016/j.molcel.2015.05.004.

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

Round, F.E., Crawford, R.M., Mann, D.G., 1990. The diatoms. Biology and morphology of the genera. Cambridge University Press, Cambridge.

Ruggiero, M.V., Kooistra, W.H., Piredda, R., Sarno, D., Zampicinini, G., Zingone, A., Montresor, M., 2022. Temporal changes of genetic structure and diversity in a marine diatom genus discovered via metabarcoding. Environ. DNA. https://doi.org/10.1002/edn3.288

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

Silvestro, D., Michalak, I., 2012. raxmlGUI: a graphical front-end for RAxML. Org. Divers. Evol. 12, 335-337. https://doi.org/10.1007/s13127-011-0056-0.

Sims, P.A., Mann, D.G., Medlin, L.K., 2006. Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45, 361-402. https://doi.org/10.2216/05-22.1.

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

Tardy, V., Bonnineau, C., Bouchez, A., Miège, C., Masson, M., Jeannin, P., Pesce, S., 2021. A pilot experiment to assess the efficiency of pharmaceutical plant wastewater treatment and the decreasing effluent toxicity to periphytic biofilms. J. Hazard. Mater. 411, 125121. https://doi.org/10.1016/j.jhazmat.2021.125121.

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., Yamanaka, H., 2019. Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. Environ. DNA 2, 42-52. https://doi.org/10.1002/edn3.44.

Turon, X., Antich, A., Palacín, C., Præbel, K., Wangensteen, O.S., 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol. Appl. 30, e02036. https://doi.org/10.1002/eap.2036.

Vanormelingen, P., Evans, K.M., Mann, D.G., Lance, S., Debeer, A.-E., D'Hondt, S., Verstraete, T., DeMeester, L., Vyverman,W., 2015. Genotypic diversity and differentiation among populations of two benthic freshwater diatoms as revealed by microsatellites. Mol. Ecol. 24, 4433–4448. https://doi.org/10.1111/mec.13336

Van den Wyngaert, S., Möst, M., Freimann, R., Ibelings, B.W., Spaak, P., 2015. Hidden diversity in the freshwater planktonic diatom *Asterionella formosa*. Mol. Ecol. 24, 2955-2972. https://doi.org/10.1111/mec.13218.

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

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. Appl. Environ. Microbiol. 73, 5261–5267. https://doi.org/10.1128/AEM.00062-07.

General discussion

Diatoms are a very diverse group of unicellular cells distributed worldwide in almost all types of aquatic systems where they play a key role in the biogeochemical cycles and food webs (Fry & Wainright, 1991; Smetacek et al., 1999). Benthic diatoms have been widely used as biological indicators in biomonitoring programs due to their broad biogeographical distribution, their sensitivity to environmental changes and their well-known ecological preferences (Stevenson 2014). However, the difficulties associated with diatom morphological identifications have led to the search for alternative methods that can facilitate the taxonomic identification of diatoms at the species level.

The arrival of DNA metabarcoding (i.e. the identification of multiple species based on high-throughput sequencing [HTS] of a particular marker) has emerged as an alternative to light microscopic identifications because it overcomes some of the inconveniences associated with morphological examinations (i.e. high time demands, the need for highly trained personnel and limitations in spatio-temporal scale). Moreover, the study of metabarcoding data can offer new insights into species diversity and ecology. However, numerous factors are known to bias the effectiveness of diatom DNA metabarcoding. Some of these can be overcome simply by adjusting the methodology used, but other limitations require technological advances to be resolved.

The work developed in this thesis aimed to identify the main limitations of the applicability of DNA metabarcoding to benthic diatom communities (with special attention put on Mediterranean areas) and to present solutions to some of these identified drawbacks. Furthermore, this work aimed to explore the possibilities achievable by the current position of metabarcoding for addressing ecological questions beyond biomonitoring and biodiversity assessments. Finally, we propose future lines of action that could improve the current state-of-the-art.

1. Main factors compromising the effectiveness of diatom DNA metabarcoding

1.1. Completeness of the reference library

One of the most important factors that determine the success of DNA metabarcoding is the availability of a complete and curated reference library since it largely determines the amount of HTS information that can be used. The current version of the diatom reference library (Diat.barcode v10; Rimet et al., 2019) covers 4783 *rbcL* entries, which represents a total of 1280 diatom taxa. This figure is certainly a minor proportion of the total diatom species estimated to exist, for instance, at least 30000-100000 species were guessed by Mann & Vanormelingen (2013). However, as discussed below, the degree of completeness of the reference library is relative to the purpose of the study and the studied environment.

Thus, if the purpose is to establish DNA metabarcoding for the study of benthic diatom communities in Mediterranean rivers, such as those of Catalan rivers (NE Spain), our analyses described in Chapter 1 indicated that the reference library can be considered quite complete since it includes most of the common species occurring in the rivers surveyed. This showed that DNA metabarcoding and LM offered a similar picture of diatom communities and a very good agreement between methods in WFD ecological status assessment, which is explained because the IPS index evaluated is strongly influenced by the indicator values of the most abundant species (i.e. pollution sensitivity values [IPSS] and pollution tolerance values [IPSV]). Indeed, to our knowledge, the agreement between approaches in WFD ecological assessments of Catalan rivers is higher than those obtained in any other similar study (e.g. Mortágua et al., 2019; Kelly et al., 2020; Vasselon et al., 2017) which strongly encourages the use of DNA metabarcoding for WFD biomonitoring of these systems. In addition, these results indicate the potential of metabarcoding not only for biomonitoring purposes but also for aspects of biodiversity, such as species distribution and ecology.

More in detail, we found that out of the 20 most abundant species recorded by morphological analysis in Catalan rivers, only 3 species (*Cocconeis euglypta*, *C. placentula* var. *lineata* and *Gomphonema lateripunctatum*) lacked representative

sequences in the version of the reference library used at the time the study was conducted (Diat.barcode v7). In contrast, the current version of the reference library (Diat.barcode v10) includes *rbcL* reference sequences for *C. euglypta*. This is clearly an important addition since our sensitivity analysis applied to morphological data (Chapter 1) showed that this species was among the 10 species that most contributed to determining the IPS scores of Catalan rivers. This inclusion is also meaningful for other Mediterranean rivers where this species is abundant and has importantly contributed to the differences between methods (Kulaš et al., 2022; Mortágua et al., 2019; Pissaridou et al., 2021).

However, our analyses detected that some of the genetic variants of *C. euglypta* widely distributed in European freshwater rivers are at risk of being discarded during bioinformatic analyses. The reason behind this is that some of these *rbcL* variants slightly differ in nucleotide length regarding the length typically assumed for the short diatom *rbcL* markers developed for metabarcoding (i.e. 331-bp or 263-bp without considering primers). In addition, the short *rbcL* sequences of *C. euglypta* show a very high nucleotide similarity with some *C. placentula* reference sequences. Therefore, it is advisable not to strictly trim the sequences using the expected marker length of 331-bp (but to allow for variations of a few nucleotides) and to assess the taxonomy of variants from this species by phylogenetic analysis rather than by classifiers based on sequence similarity (issue further developed in section 1.4).

By contrast, the analyses detailed in chapter 2 indicated that the current state of the reference library for studying benthic diatom communities in coastal Mediterranean environments is far from the level achieved for western Mediterranean rivers. The incompleteness of the reference library for these systems is explained because freshwater diatoms have traditionally been more studied than their marine counterparts, despite the enormous diversity of benthic diatoms known to occur in these environments (Witkowski et al. 2000). These gaps in the reference dataset are reflected in the fact that out of the 20 most abundant species identified morphologically, 12 lack representative *rbcL* sequences, while the equivalent figure for Catalan rivers is 2 (detailed above). Some of these abundant taxa in Ebro Delta Bays, such as *Cocconeis scutellum*, *Navicula normaloides* and *N. normalis,* are also often reported as very abundant

species in coastal systems from other Mediterranean areas which suggests that they should be considered as priorities for being barcoded.

However, it is unrealistic to expect that all the abundant and moderately abundant species of marine benthic diatoms that lack representative sequences will be covered soon by the reference library. So the question arises as to how much of the total *rbcL* data generated by HTS can be successfully translated into a molecular taxonomic inventory? There are two main ways of addressing this question. Firstly, if the criterion for considering reliable taxonomic assignments at the species level is to find an exact match with reference sequences, then less than 7% of the total inferred genetic variants in the Ebro delta bays can be classified, which contrasts sharply with the 17-22% of the genetic diversity from freshwater European systems currently represented by reference sequences (i.e. data from Catalan, UK and French rivers). Secondly, If the strategy is rather the use of commonly applied automated identification methods for metabarcoding data (e.g., the Naïve Bayesian classifier; Wang et al., 2007), the proportion of discards is reduced, as these algorithms can successfully classify query sequences that are not yet barcoded but are phylogenetically related to other sequences included in the reference library. However, even assuming that a reliable taxonomic assignment could be provided for *rbcL* variants that share ≥ 99% similarity with reference sequences, our data indicate that only 32% of the total diatom reads captured by HTS platforms in Ebro delta bays could be assigned at the species level. The same calculation for Catalan rivers shows that 80% of the total diatom reads can potentially be classified with the current state of the reference library.

Finally, these results clearly show that only a small portion of the entire benthic diatom community in coastal environments can be provided by DNA metabarcoding, which clearly undermines its applicability in these systems. It is also true, as our study and others have shown (e.g. Car et al., 2019; Kanjer et al.,2019; Hafner et al., 2018), that some of the most common marine taxa in Mediterranean coastal areas remain morphologically undescribed, indicating that both approaches are incomplete for these environments and therefore the best current strategy for assessing diatom diversity is to use both methods in parallel exploiting the strengths of each.

1.2. Variations in *rbcL* copy number per cell

Apart from the gaps in the reference library, DNA metabarcoding protocols involve a variety of technical and biological factors that can bias the results obtained (Santoferrara et al., 2019) leading to the molecular inventory produced not accurately reflecting the diatom community inhabiting a particular environment.

Among them, Chapters 1 and 2 indicated that interspecific variation in *rbcL* copy number per cell probably contributes importantly to the observed discrepancies between methods (i.e. DNA metabarcoding versus LM) in the relative abundance of species in both freshwater and marine systems. This variation depends on the gene copy number per chloroplast and chloroplast number per cell. Thus, high biovolume species with multiple chloroplasts tend to be represented by DNA metabarcoding with higher relative abundance than low biovolume species with low numbers of chloroplasts (Vasselon et al., 2018). In our study, species likely to have been underrepresented by DNA metabarcoding due to this reason were *Nitzscha inconspicua* and *Achnanthidium minutissimum* (cell biovolume 89 μm^3 and 76 μm^3 respectively; Diat.barcode v10) and overrepresented were *Ulnaria ulna, Pleurosira laevis, Achnanthes longipes* (cell biovolume 4724 μm^3, 133916 μm^3 and 8450 μm^3 respectively; Diat.barcode v10) and *Pleurosigma* sp. In freshwater systems, this bias led to important discrepancies in IPS scores between methods. In marine systems, it caused different patterns of dominance in ecological guilds depending on whether the method used was morphology or metabarcoding. It is therefore important to identify how the relative abundance of species is affected by this bias since different conclusions could be drawn about both the suitability of metabarcoding for WFD biomonitoring and the response of diatom communities to environmental pressures.

A correction factor (CF) has been developed to reduce the effect of this bias on the relative abundance of species (Vasselon et al., 2018). Chapter 1 showed that the application of CFs in our study area improved the agreement in species relative abundance between DNA metabarcoding and morphological examinations, which was translated into a higher correlation in IPS scores between methods. However, some authors have advocated avoiding the use of CFs in diatom metabarcoding data, as variations in *rbcL* copies in turn reflect

differences in cell size that may ultimately provide information on the contribution of different species to diatom productivity (Kelly et al., 2020). Indeed, the recent study by Teittinen et al. (2022) has shown that benthic diatom communities comprising larger taxa are more productive compared to those comprising smaller species, and furthermore, body size was a more important factor in explaining ecosystem productivity than species richness or evenness.

1.3. Choice of short *rbcL* markers for diatom metabarcoding

The two main markers used for diatom metabarcoding studies are the V4 region of the nuclear 18S rRNA gene and a region within the plastid *rbcL* gene, both regions being circa 300-400 bp long (including primers). It must be noted that studies aiming to evaluate potential diatom barcodes, examining sequences longer than 300-400 bp and closer to the full lengths of the genes, have shown that *rbcL* and 18S markers show a lower nucleotide divergence than COI and ITS markers which can compromise the resolution of phylogenetic relationships in certain groups (Evans et al., 2007; Guo et al., 2015). However, intragenomic variation in the ITS region (Behnke et al., 2004) and the reduced amplification success of COI for some species (Trobajo et al., 2010) undermine the potential of both regions as diatom barcodes. Thus, 18S rRNA and *rbcL* are currently the preferred markers for DNA metabarcoding because they contain enough variability to discriminate between most of the species currently recognized and are easily amplifiable. In particular, for freshwater benthic diatom metabarcoding, *rbcL* is often preferred because it is better covered in the reference library than 18S rRNA. Some exceptions are the planktonic marine genus *Chaetoceros* for which most reference sequences belong to 18S rRNA (Gaonkar et al., 2018) and, therefore, this marker is preferred for metabarcoding studies involving marine taxa of this genus (e.g. De Luca et al., 2021; Gaonkar et al., 2020).

There are currently two *rbcL* markers (263-bp and 331-bp in length) widely used for diatom metabarcoding that share a common 263-bp region but differ in the presence or absence of a 68-bp tail located at the 5’ end. Both markers have been used for biomonitoring and diversity analyses with successful results (e.g. Kang et al., 2021; Kelly et al., 2020; Rimet et al., 2018; Rivera et al., 2020) but the effects of including or discarding the variability of the 68-bp tail had not been tested. Chapter 3 indicated that the choice between these two similar diatom *rbcL*

markers has little effect on both biomonitoring purposes and biodiversity analyses. However, the shorter marker (i.e. 263-bp) shows limitations in discriminating *rbcL* variants from some common freshwater species. The most striking example is a widely distributed and abundant variant in UK rivers which is unambiguously classified as *Surirella brebissonii* by the 331-bp marker but shares the 263-bp region with a total of 10 different *Surirella* taxa. As shown in our results, this may have implications for biomonitoring as some of these identical species for the 263-bp region differ in the indicator values. Indeed, this was the reason that explained why some sites altered their ecological status from acceptable classes (i.e. "Good"/"High") to unacceptable status ("Poor"/"Poor"/"Moderate") after switching from the 331-bp marker to the 263-bp one respectively. Additionally to the WFD implications, these cases also reflect limitations of the 263-bp marker for studying aspects related to species distribution, ecology and intraspecific diversity.

On the other hand, chapter 3 indicated that the 331-bp marker reflects more information about the amino acid sequence of the *rbcL* gene than the 263-bp region since the extra 68-bp tail includes important amino acid variability. Importantly, this extra aminoacid information could provide additional insights into the ecology of the species once it is understood how the amino acid sequence of the *rbcL* gene affects the efficiency of the Rubisco. Thus, the *rbcL* gene contains the catalytic sites of Rubisco and therefore the amino acid structure of this gene is expected to determine the efficiency of the enzyme (Liu et al., 2010). The key point here, as indicated by Valegård et al. (2018), is that interspecific differences in Rubisco efficiency, which have been reported for some marine species (Young et al., 2016), could be ecologically relevant if they affect the competitive ability and environmental adaptation of diatom species, similar to what has been suggested for some gymnosperm species for which it has been observed that changes in *rbcL* amino acids were correlated with distribution differences along the altitudinal gradient (Liu et al., 2010). However, very little is known about the structure of Rubisco *rbcL* in diatoms (the only report of crystal structure in diatoms has recently been given by Valegård et al., 2018), making it difficult to infer to what extent amino acid changes across the 331-bp region may affect Rubisco efficiency and, consequently, how these changes might influence the

environmental adaptability of diatom species. To conclude, the study of large amino acid datasets, easily facilitated by DNA metabarcoding, may shed light on these aspects once the understanding of diatom *rbcL* structure is improved.

1.4. Taxonomic classification of *rbcL* diatom genetic variants: Biases and recommendations

A common finding in our studies was that an important number of diatom genetic variants captured by HTS, and whose taxonomy at the species level can be resolved because there are closely related reference sequences, were nevertheless at risk of being discarded after the application of the bioinformatic pipeline due to one of the following reasons: 1) the classifier does not assign the *rbcL* variant to any species, 2) *rbcL* variants receive a taxonomic assignation but not the correct one (i.e. false positives) or 3) correctly classified variants are rejected because of the poor bootstrapping support received (i.e. false negatives). Our analyses in chapter 3 found that the above cases were often explained by a decrease in Bayesian classifier efficiency caused either by low coverage of *rbcL* species diversity in the reference library or by high nucleotide similarity in the *rbcL* marker between separate species.

Phylogenetic analyses, performed on a large number of species during our studies, proved to be a very efficient procedure to accurately classify at the species level a significant number of genetic variants that were at risk of being discarded for the reasons mentioned above. The main limitation of this procedure is its complexity since performing phylogenetics trees for many taxa is a laborious task that requires much more time than classifying sequences using similarity-based classifiers. In addition, reconstructing the phylogeny of query variants also requires taxonomic knowledge because the choice of the reference taxa to be included is crucial to properly resolve phylogenetic relationships between clades. Despite these disadvantages, the procedure has proven to be very efficient, as demonstrated by the highly diverse species complexes *Nitzschia inconspicua, Fistuilfera saprophila* and *Achanthidium minutissimum* (Chapter 4). For each of these 3 species, the taxonomy of more than 50% of the total variants detected by our analyses in Catalan and French rivers were not represented in the reference library. Preliminary identification was made by the Bayesian classifier, but in addition, phylogenetic analyses were made to check their assignment.

Importantly, the higher adequacy of phylogenetic-based classification for diatom *rbcL* variants was also reflected in a lower number of false negatives compared to the number obtained by naïve Bayesian approach.

Therefore, these results encourage the use of phylogenetic (or evolutionary) placement algorithms such as EPA-ng or PPLACER (Barbera et al., 2018; Matsen et al., 2010) that classify short sequences on the basis of a phylogenetic reference tree. In addition, recently developed tools (e.g. Genesis and Gappa; Czech et al., 2020) have optimised the time and computational resources required by phylogenetic placement methods (main limitations associated with these methods), making these algorithms even more appealing for studies that involve large spatial and temporal scales and where the taxonomy of thousands of sequences is to be elucidated.

On the other hand, our results (Chapters 3, 4 and 5) reflected that the divergence of the short *rbcL* markers evaluated (i.e. 263-bp and 331-bp) is not very similar among species but it varies greatly. Therefore, it is highly recommendable to avoid the use of OTUs approaches based on fixed and arbitrary similarity thresholds such as the traditionally assumed 97% threshold, which was initially established for the 16S rRNA in bacteria (Stackebrandt & Goebel, 1994). Though it has been argued that this cutoff may also work for these short diatom *rbcL* markers (Kelly et al., 2020), our results cast doubt on the effectiveness of this cutoff, as *rbcL* variants from different morphospecies were often observed to have nucleotide similarities above 97% and therefore would be allocated to the same OTU. Some of the examples detected in our data are, among others, *Diatoma moniliformis* and *D. tenuis*; *Encyonema ventricosum* and *E. minutum*; *Nitzschia perminuta* and *N. acidoclinata*. Importantly, these pairs of species differ in their indicator values (i.e. IPSS and IPSV), which means that the decision about the threshold to be used can have a significant impact on the final ecological status assessment, as demonstrated by Tapolczai et al. (2019). It also artificially undermines the potential of DNA metabarcoding for other ecological studies where achieving the lowest possible taxonomy level is crucial. For instance, metabarcoding procedures avoiding taxonomical classification of genetic entities (e.g. Apothéloz-Perret-Gentil et al., 2017; Feio et al., 2020; Smucker et al., 2020) are not able to allow analysis of diatom traits, such as ecological guilds, which

provide meaningful information about the response of diatom communities to environmental conditions and stressors (Passy, 2007; Tapolczai et al., 2016 & 2017).

2. Possibilities brought by DNA metabarcoding in the current state-of-the-art

Although the factors mentioned above may diminish the effectiveness of DNA metabarcoding for characterising benthic diatom communities, our analyses indicated that DNA metabarcoding is in its current state a promising tool for assessing aspects of diatom diversity and ecology that are difficult or impossible to address through morphological analyses. Thus, in the following sections, we discuss the opportunities of metabarcoding to study the phylogeography of diatoms, the significance of genetic diversity within species complexes and the potential of the method to reflect the diversity that is overlooked by LM.

2.1. Phylogeographical patterns and meaning of intraspecific variation in freshwater diatoms

The use of markers with sufficient phylogenetic signal at the intraspecific level, together with methods capable of detecting and separating sequencing artefacts, make DNA metabarcoding a technique capable of providing significant information on the genetic diversity of species and how that diversity is geographically structured (Turon et al., 2022). Our results from the analysis of a large metabarcoding dataset spanning different biogeographic regions (chapter 5) indicated important differences between species with respect to the intraspecific diversity of the 263-bp marker. Thus, after studying the intraspecific diversity of a total of 74 freshwater species, we could define four common phylogeographic patterns among species that were characterised by a) the number of variants per species, b) the presence of rare and/or dominant variants and, c) the apparent geographic dispersal ability. Although we were able to identify four different phylogeographic patterns between species for the 263-bp region, we could not reach definitive conclusions about what causes these

patterns, though there we found some interesting correlations that we speculate may be related to the intraspecific heterogeneity.

Thus, our data clearly indicated that centric species showed significantly fewer *rbcL* variants than pennate species. Except for *Cyclotella meneghiniana* and *Melosira varians*, there were no centric species with more than 4 *rbcL* variants and most of them were represented by only 1 or 2 variants. The cause of these differences between pennate and centric species is difficult to disentangle but could be related to several aspects. On the one hand, the higher *rbcL* diversity of pennate species could be associated with the fact that different patterns of plastid inheritance have been reported for pennate and centric diatom species. Thus, plastids are inherited uniparentally in centric diatom species whereas they are inherited biparentally in most pennate species in which this topic has been studied (Jensen et al. 2003; Round et al. 1990). Given that biparental inheritance may produce a greater number of *rbcL* haplotypes in the F1 generation, it is to be expected, perhaps, that pennate species will maintain higher *rbcL* diversity than centric species. Additionally, the rate of diversification in diatoms has been reported to be higher in lineages showing an isogamous reproductive mode (i.e. mainly observed in pennate species) than in lineages with an oogamous mode (i.e. only reported in centric species) (Nakov et al., 2018).

Although both reproduction mode (i.e. during auxosporulation) and plastid inheritance patterns (segregation during mitotic cell division) could be behind the higher intraspecific diversity found for the pennate species, it should be noted that such differences could be partly because of the number of centric species analysed was much lower compared to the pennate species. This is explained because the samples analysed in this study correspond to the benthic habitat where most species are pennate. A future study increasing the sampling effort for centric species, and covering both benthic and planktonic habitats, could shed light on the differences in intraspecific diversity between groups as well as in the phylogeographic patterns reported among species.

On the other hand, as demonstrated in Chapter 4, phylogeographic studies based on metabarcoding of short markers are particularly useful for improving our understanding of the significance of intraspecific variation in diatom species complexes. In this regard, our analyses in Catalan and French rivers showed that

rbcL variants of *Achnanthium minutissimum, Fistulifera saprophila* and *Nitzschia inconspicua* species were widely but not uniformly distributed. Thus, a high proportion of variants were detected in both Catalan and French rivers and in addition, a comparison with the UK dataset used in Chapter 3 revealed that many of these variants were also found in UK rivers. Moreover, some of these variants showed 100% identity with clonal *rbcL* sequences isolated from elsewhere in Europe, North America and Hawaii. Overall, these results agree well with the ubiquitous dispersal hypothesis of Finlay (2002) and some other studies, based on Sanger sequence data, have reported that similar or identical diatom *rbcL* variants are distributed in different geographical regions (Vanormelingen et al., 2015).

Despite the presence of many variants in separate regions, the distribution of genetic variants varied greatly at more local scales, which could indicate that, although individuals are able to disperse over long distances, their biogeography at the local level is shaped by a combination of local barriers to dispersal, fine-scale environmental conditions and stochastic processes. These factors have been suggested to explain the biogeography of diatoms in other regions (Keck et al., 2018) but also the distribution of a wide range of protist groups (Logares et al., 2018, Singer et al., 2021).

In addition to the uneven geographical distribution, chapter 4 indicated that *rbcL* variants within *A. minutissimum* and *F. saprophila* differed in their ecological preferences. This may indicate that the broad ecological tolerance assumed for these complexes may be the result of a continuum of overlapping preferences between variants, which has clear implications for biomonitoring programmes, as not all variants within a species complex should be assigned the same indicator values. Based on these results, it is strongly recommended to carry out a similar approach that could shed light on the significance of the high intraspecific diversity detected in some other common species such as *Nitzschia palea, Amphora pediculus, Ulnaria ulna, Cyclotella meneghiniana, Eunotia bilunaris* or *Cocconeis placentula*.

2.2. DNA metabarcoding is able to identify weakly-silicified, rare, small and recently described species easily overlooked by LM

Several reasons explain why DNA metabarcoding is an effective tool for the study of species that are often neglected or misidentified by LM. Firstly, samples analysed by DNA metabarcoding do not undergo the specific chemical treatment used in morphological analysis to facilitate the visualisation of diatom valves under LM. Chapters 1 and 2 evidenced that the absence of this preparation process was an important factor explaining some of the significant discrepancies observed between methods in both freshwater and marine environments. In the case of Catalan rivers, *Fistulifera saprophila* was the most remarkable example. This species was better represented by metabarcoding probably because its weakly silicified frustules (Zgrundo et al., 2013) were dissolved after the chemical treatment used in the LM. Importantly, the higher representation of this species with DNA metabarcoding may have important economic consequences if the transition from LM to DNA metabarcoding for routine biomonitoring of WFD becomes effective. Similarly, in marine environments, *Thalassiosira profunda* was the predominant species according to the molecular method while it was hardly recorded by LM. We hypothesise that this species is widely present in this environment as an endosymbiont in foraminifera or dinoflagellates and, consequently, the species and/or host organism might have been lost after chemical treatment of the samples.

Secondly, our results also indicated that DNA metabarcoding is more sensitive than LM for detecting rare, small and recently described species. Thus, some species are too rare to be found by the common 300-400 valve counts performed in routine LM examinations, whereas the generation of thousands of reads per sample by metabarcoding allows the detection of these very rare specimens. Similarly, DNA metabarcoding is most effective in identifying small and newly described species that are easily overlooked during LM counts as a consequence of their small size and the fact that they may not be included in the taxonomic keys used during routine LM counts. Examples of species identified by metabarcoding, but missed by LM due to the above reasons, were *Pseudo-nitzschia delicatissima*, *Planothidium victorii* and *Gedaniella panicellus*.

2.3. Non-diatom taxa amplified by diatom designed *rbcL* primers

Our analyses demonstrate that *rbcL* primers designed for amplifying a short 263-bp region of freshwater diatoms (Vasselon et al., 2017) can amplify non-diatom species in both marine and freshwater environments. In marine environments, results from chapter 2 show that a total of 41 ASVs were classified across 10 different classes of the phylum Ochrophyta. This reflects that the primer binding region is highly conserved among the different classes of this phylum which could be due to the fact that all classes of Ochrophyta share the same Rubisco ID Form (Tabita et al., 1999; Íñiguez et al., 2020). Interestingly, this binding region seems to be also conserved among other rubisco forms, since we observed that Chlorophyta species (Rubisco IB form) were detected with these primers in Catalan and French rivers.

The possibility to identify other groups by metabarcoding should be interpreted as valuable additional information provided by the method, as some of these non-diatom taxa identified in our study area are difficult to identify by LM, due to their small size and similar morphologies, and some others are relevant taxa from an economic and ecological perspective. An example is *Chattonella subsalsa*, which was detected in the Ebro delta bays and is associated with red tides and mass fish kill events (Lewitus et al., 2008). Overall, these results highlight the importance of further exploring the range of taxa that can be detected by primers designed for diatoms. However, despite these interesting findings, non-diatom ASVs constituted only a minor proportion of the total reads generated (e.g. ~ 1% in coastal environments) and therefore these primers cannot be expected to provide a global view of the microphytobenthos communities occurring in the environment, but can only recover a few examples for some particular clades. Thus, if the purpose is to study other specific groups of non-diatom microphytobenthos using short *rbcL* markers, primers designed for targeting these groups should be used as recently demonstrated by the high diversity of Eustigmatophyceae *rbcL* variants obtained via metabarcoding (Fawley et al., 2021). If the objective is rather to obtain a broader picture of the microbenthic communities, different and more variable markers should be used such as the V4 of the 18S rRNA (further developed in section 3.2)

3. Future perspectives

As discussed in the previous sections, the current position of DNA metabarcoding may provide an effective method for WFD biomonitoring programmes. Furthermore, it may offer new opportunities to study aspects of species ecology that were previously unknown and difficult to infer through traditional methods. Despite these possibilities, several technological and methodological advances could extend the currently achievable dimension with metabarcoding.

3.1. Third-generation sequencing technologies

Our studies identified that one important factor undermining the potential of diatom DNA metabarcoding was the existence of high sequence homology among *rbcL* variants from separated species, which difficulted or precluded the unambiguous identification of some variants at the species level (see section 1.4). This limitation can be overcome with the arrival of long-read sequencing platforms (e.g. Pacific Bioscience or Oxford Nanopore Technologies) capable of providing reliable sequencing lengths well above 1200 - 1600 bp (Tedersoo et al., 2020), which is the common length of the full *rbcL* region in diatoms. Moreover, phylogeography analyses using the full *rbcL* genotypes could well characterize the genetic diversity structure of diatoms species, which could confirm the differences observed within species complexes in ecological preferences and phylogeographic patterns.

Another benefit derived from this technology is that long sequencing reads extracted from the environment can be used to create more robust and complete reference phylogenies, filling the gaps generated by the lack of reference sequences. At the same time, these robust phylogenies lead to increased confidence in the taxonomic placement of short reads (Jamy et al., 2019). This could be especially useful for assessing the taxonomy of genetic variants from poorly studied environments, such as coastal environments, where most inferred diatom variants remained unclassified after applying bioinformatics analyses. Despite the advantages, there are some drawbacks associated with long-read sequencing technologies. The most important limitations are perhaps the fact that both Pacific Bioscience (PacBio) and Oxford Nanopore Technologies show lower sequencing depth and higher error rates than short-read technologies (Sanding

et al., 2021; Tedersoo et al., 2020), which could particularly compromise the study of species biodiversity and genetic diversity, as both factors may lead to an underestimation of the number of genetic variants occurring in the environment. However, the accuracy of third-generation sequencing technologies has greatly increased in the past years. In the case of PacBio technologies, it has been reported the capability of providing highly accurate long reads (≥99.8%) through the use of the circular consensus sequencing method (CSS) (Wenger et al., 2019). On the other hand, it should be noted that the order of genes in the chloroplast genome in diatoms is not conserved (Hamsher et al., 2019). This makes it difficult to find universal primers capable of amplifying chloroplast regions containing genes other than the *rbcL* gene, which limits the potential of these technologies.

3.2. Broaden the view of the microeukaryotic community

As indicated in section 2.3, short *rbcL* primers cannot provide a representative view of the whole microeukaryotic community occurring in a particular environment but only specific groups can be examined. For covering a wider range of groups, other regions are needed and in this regard, markers from the 18S rRNA region have been the most widely used in protist metabarcoding studies and particularly, the V4 hypervariable regions have been often applied for exploring the diversity of protists (e.g. de Vargas et al., 2015; Massana et al., 2015; Yeh et al., 2020). The attraction of this marker for metabarcoding studies is explained because it contains sufficient phylogenetic resolution in many groups, there are universal primers (e.g. Reuk454FWD1 and ReukREV3), and importantly, it is covered by several curated references libraries such as Silva and PR^2 (Guillou et al. 2013; Quast et al. 2012). Importantly, the wider taxonomic coverage achievable using these markers extends the applicability of DNA metabarcoding for other objectives such as, for example, the study of species' role in bioremediation processes (Annex 1).

Despite these advantages, the use of the rRNA region also has some drawbacks. For instance, the V4 marker is not optimal for separating certain taxa within several groups such as Haptophyte, Streptophyta or Chlorophyta (Lopes dos Santos et al., 2017; Pawlowski et al., 2012), although this limitation will be likely bypassed by integrating further rRNA genes or even the full rRNA operon using

long sequencing technologies (Heeger et al., 2018; Jamy et al., 2019). Furthermore, it has been shown that environmental changes can drive intraspecific variation in rRNA gene copy number (Lavrinienko et al., 2021), making it difficult to extract robust statements about species response to the environment and population dynamics via metabarcoding of rRNA markers, as exemplified in Ruggiero et al. (2022).

Finally, in the particular case of diatoms, switching to another marker is probably suboptimal because *rbcL* is better covered by the diatom reference library. For this reason, most of the diatom metabarcoding work done in previous years has been based on *rbcL*, so switching to another marker would increase the number of protists groups covered but at the same time, it would compromise the comparability with previous studies clearly undermining potential of DNA metabarcoding. This trade-off, therefore, implies that the choice of the marker should be made according to the research question being addressed and the community wanted to be studied.

3.3. Enhancing the compatibility of data and developing new metrics and ecological understanding

DNA metabarcoding involves a variety of different steps, from sampling to bioinformatics analysis, which can be adjusted according to the research question addressed, thus considerably improving the efficiency of the method. However, it is this flexibility in methodology that in turn can compromise the comparability of results and the transferability of the methods used (e.g. Bailet et al., 2020, Vasselon et al., 2018). Currently, in order to increase the reproducibility and transferability of routine biomonitoring assessment on a European scale, major efforts are being undertaken to standardise some of the critical steps of DNA metabarcoding, such as DNA extraction and PCR amplification (Vasselon et al., 2021). However, universalising a single strict protocol and method on a European scale may be a mistake. Instead it may be more advisable to allow some flexibility, since, as outlined below, the potentially most effective strategy depends on the benthic community inhabiting the particular region to be examined.

Thus, Vasselon et al. (2018) reported that different DNA extraction kits did not perform equally among all the taxa, for instance, some kits were apparently better in extracting DNA from *Nitzschia* and *Amphora,* whereas other kits were more efficient for taxa from *Encyonema, Gomphonema* and *Navicula*. In relation to bioinformatics and the marker used, this thesis has shown some cases where standardisation towards a fixed protocol could again be sub-optimal. Examples are those *rbcL* variants that cannot be discriminated using the 263-bp marker (e.g. variants of *Surirella brebissonii*) or those prone to be misclassified by commonly applied automatic classifiers based on sequence similarity (e.g. variants of *Achnanthidium minutissimum* complex, *Nitzschia perminuta*, *Encyonema ventricosum*). Another example was variants of *Cocconeis placentula* where the ASVs have different lengths to those assumed for either 263-bp and 331-bp diatom markers and are thus at risk of being deleted during bioinformatics analyses if a strict or inappropriate length filter is applied. We conclude that, although standardisation can be positive to increase transferability between laboratories, the establishment of rigid protocols can potentially compromise the effectiveness of the method for both biomonitoring and ecological studies, so expert judgement should prevail when deciding which protocols and methods to apply in each specific case.

On the other hand, comparability and future use of data would be enhanced if researchers made the inferred ASVs or OTUs publicly available, and in a comparable format, along with physical and chemical data of the samples analyzed, if these are available. Such matched datasets are not common and their further analysis could lead to significant advances. For example, it would be possible to increase knowledge of the occurrence and preferences of some species whose ecology is little known (e.g. *Nitzschia dissipata var. media, Planothidium victorii, Fragilaria agnesiae*). The better characterisation of the ecological profiles of diatom species could then be used to establish indicator values for species that currently lack them. However, it would be a mistake to restrict the use of metabarcoding data to computing metrics that emulate indices designed for use with morphological data. Instead, it will be important to take advantage of aspects that are exclusively provided by DNA metabarcoding. Examples, as discussed in the previous sections, could be the use of information

about the contribution of species to diatom productivity inferred from species differences in *rbcL* copy number per cell; the integration of information about the non-diatom taxa co-amplified with diatoms, which could provide a wealth of information about them that is often missing because of difficulties in identifying them (e.g. many eustigmatophytes and unicellular green algae have few morphological characteristics that can be used in diagnosis and rapidly decay after sampling); and the inclusion of entities whose taxonomy cannot be determined using the current reference library but whose ecological profiles can be assessed via a taxonomic free approach (e.g. Tapolczai et al., 2021). All of these factors, together with the application of third-generation (long-read) sequencing technologies and the technical recommendations presented in this thesis, would enable the development of more informative biological indices of ecosystem health and would also enhance our current capacity to study the ecology and diversity of microeukaryotes.

References

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., Pawlowski, J., 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Mol. Ecol. Resour. 17, 1231-1242. https://doi.org/10.1111/1755-0998.12668.

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

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., Stamatakis, A., 2019. EPA-ng: massively parallel evolutionary placement of genetic sequences. Syst. Biol. 68, 365-369. https://doi.org/10.1093/sysbio/syy054.

Behnke, A., Friedl, T., Chepurnov, V.A., Mann, D.G., 2004. Reproductive compatibility and Rdna sequence analyses in the *Sellaphora Pupula* species complex (Bacillariophyta). J. Phycol. 40, 193-208. https://doi.org/10.1046/j.1529-8817.2004.03037.x

Car, A., Witkowski, A., Dobosz, S., Jasprica, N., Ljubimir, S., Zgłobicka, I., 2019. Epiphytic diatom assemblages on invasive *Caulerpa taxifolia* and autochthonous *Halimeda tuna* and *Padina* sp. seaweeds in the Adriatic Sea – summer/autumn aspect. Oceanol Hidrobiol Stud. 48, 209-226. https://doi.org/10.2478/ohs-2019-0019.

Czech, L., Barbera, P., Stamatakis, A., 2020. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics, 36, 3263-3265. https://doi.org/10.1093/bioinformatics/btaa070.

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

de Vargas, C., Audic, S., Henry, N., Decelle, J., MahØ, F., Logares, R., Lara, E., Berney, C., Bescot, N.L., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Luke˘s, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stem-mann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E.G., Sardet, C., Sullivan, M.B., Velayoudon, D., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605, http://dx.doi.org/10.1126/science.1261605

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

Fawley, M.W., Fawley, K.P., Cahoon, A.B., 2021. Finding needles in a haystack—extensive diversity in the Eustigmatophyceae revealed by community metabarcode analysis targeting the *rbcL* gene using lineage-directed primers. J. Phycol. 57, 1636-1647. https://doi.org/10.1111/jpy.13196

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

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, 261-273. https://doi.org/10.1078/1434-4610-00103.

Fry, B., Wainright, S.C., 1991. Diatom sources of 13 C-rich carbon in marine food webs. Mar. Ecol. Prog. Ser. 76, 149-157.

Gaonkar, C.C., Piredda, R., Minucci, C., Mann, D.G., Montresor, M., Sarno, D., Kooistra, W.H., 2018. Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. PloS one 13, e0208929. https://doi.org/10.1371/journal.pone.0208929.

Gaonkar, C.C., Piredda, R., Sarno, D., Zingone, A., Montresor, M., Kooistra, W.H., 2020. Species detection and delineation in the marine planktonic diatoms Chaetoceros and Bacteriastrum through metabarcoding: making biological sense of haplotype diversity. Environ Microbiol. 22, 1917-1929. https://doi.org/10.1111/1462-2920.14984.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R., 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597-D604. https://doi.org/10.1093/nar/gks1160.

Guo, L., Sui, Z., Zhang, S., Ren, Y., Liu, Y., 2015. Comparison of potential diatom 'barcode'genes (the 18S rRNA gene and ITS, COI, *rbcL*) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. Int. J. Syst. Evol. Microbiol. 65, 1369-1380. https://doi.org/10.1099/ijs.0.000076.

Hafner, D., Car, A., Jasprica, N., Kapetanović, T., Dupčić Radić, I., 2018. Relationship between marine epilithic diatoms and environmental variables in oligotrophic bay, NE Mediterranean. Mediterr. Mar. Sci. 19, 223–239. https://doi.org/10.12681/mms.14151.

Hamsher, S.E., Keepers, K.G., Pogoda, C.S., Stepanek, J.G., Kane, N.C., Kociolek, J.P., 2019. Extensive chloroplast genome rearrangement amongst three closely related *Halamphora* spp. (Bacillariophyceae), and evidence for rapid evolution as compared to land plants. PloS one, 14, e0217824. https://doi.org/10.1371/journal.pone.0217824.

Heeger, F., Bourne, E.C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., Overmann, J., Mazzoni, C.J., Monaghan, M.T., 2018. Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Mol. Ecol. Resour. 18, 1500-1514. https://doi.org/10.1111/1755-0998.12937.

Iñiguez, C., Capó-Bauçà, S., Niinemets, Ü., Stoll, H., Aguiló-Nicolau, P., Galmes, J., 2020. Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms. Plant J. 101, 897-918. https://doi.org/10.1111/tpj.14643

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., Burki, F., 2019. Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. Mol. Ecol. Resour. 20, 429-443. https://doi.org/10.1111/1755-0998.13117.

Jensen, K.G., Moestrup, Ø., Schmid, A.M.M., 2003. Ultrastructure of the male gametes from two centric diatoms, *Chaetoceros laciniosus* and *Coscinodiscus wailesii* (Bacillariophyceae). Phycologia 42, 98-105. https://doi.org/10.2216/i0031-8884-42-1-98.1.

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría Galindo, P., Vences, M., Wang, J., Schwalba, A., 2021. Diatom Metabarcoding and Microscopic Analyses from Sediment Samples at Lake Nam Co, Tibet: The Effect of Sample-Size and Bioinformatics on the Identified Communities. Ecol. Indic. 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070.

Kanjer, L., Mucko, M., Car, A., Bosak, S., 2019. Epiphytic diatoms on *Posidonia oceanica* (L.) Delile leaves from eastern Adriatic Sea. Nat. Croat. 28, 1-20. https://doi.org/10.20302/NC.2019.28.1.

Keck, F., Franc, A., Kahlert, M., 2018. Disentangling the processes driving the biogeography of freshwater diatoms: A multiscale approach. J. Biogeogr. 45, 1582-1592. https://doi.org/10.1111/jbi.13239.

Keck, F., Blackman, R.C., Bossart, R., Brantschen, J., Couton, M., Hürlemann, S., Kirschner, D., Locher, N., Zhang, H., Altermatt, F., 2022. Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. Mol. Ecol. 31, 1820-1835. https://doi.org/10.1111/mec.16364

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

Kulaš, A., Udovič, M.G., Tapolczai, K., Žutinić, P., Orlić, S., & Levkov, Z., 2022. Diatom eDNA metabarcoding and morphological methods for bioassessment of karstic river. Sci. Total Environ. 829, 154536. https://doi.org/10.1016/j.scitotenv.2022.154536

Lavrinienko, A., Jernfors, T., Koskimäki, J.J., Pirttilä, A.M., Watts, P.C., 2021. Does intraspecific variation in rDNA copy number affect analysis of microbial communities?. Trends Microbiol. *29*, 19-27. https://doi.org/10.1016/j.tim.2020.05.019.

Lewitus, A.J., Brock, L.M., Burke, M.K., DeMattio, K.A., Wilde, S.B., 2008. Lagoonal stormwater detention ponds as promoters of harmful algal blooms and eutrophication along the South Carolina coast. Harmful Algae, 8, 60-65. https://doi.org/10.1016/j.hal.2008.08.012.

Liu, N., Wang, Q., Chen, J., Zhu, Y., Tashi, T., Hu, Y.Y., Chen, F., Zhong, Y., 2010. Adaptive evolution and structure modeling of *rbcL* gene in Ephedra. Chinese Sci. Bull. 55, 2341-2346. https://doi.org/10.1007/s11434-010-3023-9.

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noël, M.H., Decelle, J., Romac, S., Vaulot, D., 2017. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. ISME J 11, 512-528. https://doi.org/10.1038/ismej.2016.120.

Logares, R., Tesson, S.V., Canbäck, B., Pontarp, M., Hedlund, K., Rengefors, K., 2018. Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. Environ. Microbiol. 20, 2231-2240. https://doi.org/10.1111/1462-2920.14265.

Mann, D.G., Vanormelingen, P., 2013. An inordinate fondness? The number, distributions, and origins of diatom species. J. Eukaryot. Microbiol. 60, 414–420. https://doi.org/10.1111/jeu.12047

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W.H.C.F., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., de Vargas, C., 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ. Microbiol. 17, 4035e4049. https://doi.org/10.1111/1462-2920.12955.

Matsen, F.A., Kodner, R.B., Armbrust, E.V., 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC bioinform. 11, 1-16. https://doi.org/10.1186/1471-2105-11-538.

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

Nakov, T., Beaulieu, J.M., Alverson, A.J., 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol 219, 462-473. https://doi.org/10.1111/nph.15137.

Passy, S.I., 2007. Diatom ecological guilds display distinct and predictable behaviour along nutrient and disturbance gradients in running waters. Aquat. Bot. 86, 171-178. https://doi.org/10.1016/j.aquabot.2006.09.018.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S.S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D.G., Mitchell, E.A.D., Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G.B., Smirnov, A.V., Spouge, J.L., Stern, R.F., Stoeck, T., Zimmermann, J., Schindel, D., de Vargas, C., 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLoS Biol. 10, e1001419. http://dx.doi.org/10.1371/journal.pbio.1001419.

Pissaridou, P., Vasselon, V., Christou, A., Chonova, T., Papatheodoulou, A., Drakou, K., Tziortzis, I., Dörflinger, G., Rimet, F., Bouchez, A., Vasquez, M.I., 2021. Cyprus' diatom diversity and the association of environmental and anthropogenic influences for ecological assessment of rivers using DNA metabarcoding. Chemosphere 272, 129814. https://doi.org/10.1016/j.chemosphere.2021.129814.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-D596. https://doi.org/10.1093/nar/gks1219.

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. Org. Divers. Evol. 18, 51-62. https://doi.org/10.1007/s13127-018-0359-5.

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

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

Round, F.E., Crawford, R.M., Mann, D.G., 1990. The diatoms. Biology and morphology of the genera. Cambridge University Press, Cambridge.

Ruggiero, M.V., Kooistra, W.H., Piredda, R., Sarno, D., Zampicinini, G., Zingone, A., Montresor, M., 2022. Temporal changes of genetic structure and diversity in a marine diatom genus discovered via metabarcoding. Environ. DNA. https://doi.org/10.1002/edn3.288

Sandin, M.M., Romac, S., Not, F., 2021. Intra-genomic rDNA gene variability of Nassellaria and Spumellaria (Rhizaria, Radiolaria) assessed by Sanger, MinION and Illumina sequencing. Environ. Microbiol. https://doi.org/10.1111/1462-2920.16081

Santoferrara, L.F., 2019. Current practice in plankton metabarcoding: optimization and error management. J. Plankton Res. 41, 571-582. https://doi.org/10.1093/plankt/fbz041.

Stackebrandt, E., Boebel, B.M., 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 44, 846-849. https://doi.org/10.1099/00207713-44-4-846.

Singer, D., Seppey, C.V.W., Lentendu, G., Dunthorn, M., Bass, D., Belbahri, L., Blandenier, Q., Debroas, D., de Groot, G.A., de Vargas, C., Domaizon, I., Duckert, C., Izaguirre, I., Koenig, I., Mataloni, G., Schiaffino, M.R., Mitchell, E.A.D., Geisen, S., Lara, E., 2021. Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. Environ. Int. 146. https://doi.org/10.1016/j.envint.2020.106262.

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

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

Stevenson, J., 2014. Ecological assessments with algae: a review and synthesis. J. Phycol. 50, 437–461. https://doi.org/10.1111/jpy.12189

Tabita, F.R., 1999. Microbial ribulose 1, 5-bisphosphate carboxylase/oxygenase: a different perspective. Photosynth. Res. 60, 1-28. https://doi.org/10.1023/A:1006211417981.

Tapolczai, K., Bouchez, A., Stenger-Kovács, C., Padisák, J., Rimet, F., 2016. Trait-based ecological classifications for benthic algae: review and perspectives. Hydrobiologia 776, 1-17. https://doi.org/10.1007/s10750-016-2736-4.

Tapolczai, K., Bouchez, A., Stenger-Kovács, C., Padisák, J., Rimet, F., 2017. Taxonomy-or trait-based ecological assessment for tropical rivers? Case study on benthic diatoms in Mayotte island (France, Indian Ocean). Sci. Total Environ. 607, 1293-1303. https://doi.org/10.1016/j.scitotenv.2017.07.093.

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., Vasselon, V., 2019. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. Front. Ecol. Evol. 409. https://doi.org/10.3389/fevo.2019.00409.

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

Tedersoo, L., Albertsen, M., Anslan, S., Callahan, B., 2021. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. Appl. Environ. Microbiol. 87, e00626-21. https://doi.org/10.1128/AEM.00626-21

Teittinen, A., Soininen, J., Virta, L., 2022. Studying biodiversity–ecosystem function relationships in experimental microcosms among islands. Ecology 103, e3664. https://doi.org/10.1002/ecy.3664.

Trobajo, R., Mann, D.G., Clavero, E., Evans, K.M., Vanormelingen, P., McGregor, R. C., 2010. The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). Eur. J. Phycol. 45, 413-425. https://doi.org/10.1080/09670262.2010.498586.

Turon, X., Antich, A., Palacín, C., Præbel, K., Wangensteen, O.S., 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol. Appl. 30, e02036. https://doi.org/10.1002/eap.2036.

Yeh, H.D., Questel, J.M., Maas, K.R., Bucklin, A., 2020. Metabarcoding analysis of regional variation in gut contents of the copepod *Calanus finmarchicus* in the North Atlantic Ocean. Deep Sea Research Part II: Topical Studies in Oceanography, 180, 104738. https://doi.org/10.1016/j.dsr2.2020.104738.

Valegård, K., Andralojc, P.J., Haslam, R.P., Pearce, F.G., Eriksen, G.K., Madgwick, P.J., Kristoffersen, A.K., van Lun, M., Klein, U., Eilertsen, H.C., Parry, M., Andersson, I., 2018. Structural and functional analyses of Rubisco from arctic diatom species reveal unusual posttranslational modifications. J. Biol. Chem. 293, 13033-13043. https://doi.org/10.1074/jbc.RA118.003518.

Vanormelingen, P., Evans, K.M.,Mann, D.G., Lance, S., Debeer, A.-E., D'Hondt, S., Verstraete, T., DeMeester, L., Vyverman,W., 2015. Genotypic diversity and differentiation among populations of two benthic freshwater diatoms as revealed by microsatellites. Mol. Ecol. 24, 4433-4448. https://doi.org/10.1111/mec.13336

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

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

Vasselon, V., Ács, É., Almeida, S., Andree, K., Apothéloz-Perret-Gentil, L., Bailet, B., Baricevic, A., Beentjes, K., Bettig, J., Bouchez, A., Capelli, C., Chardon, C., Duleba, M., Elersek, T., Genthon, C., Hurtz, M., Jacas, L., Kahlert, M., Kelly, M., Lewis, M., Macher, JN., Mauri, F., Moletta-Denat, M., Mortágua, A., Pawlowski, J., Pérez-Burillo, J., Pfannkuchen, M., Pilgrim, E., Pissaridou, P., Porter, J., Rimet, F., Stanic, K., Tapolczai, K., Theroux, S., Trobajo, R., van der Hoorn, B., Vasquez Hadjilyra, MI., Walsh, K., Wanless, D., Warren, J., Zimmermann, J., Zupančič, M., 2021. The Fellowship of the Ring Test: DNAqua-Net WG2 initiative to compare diatom metabarcoding protocols used in routine freshwater biomonitoring for standardisation. ARPHA Conference Abstracts 4: e65142. https://doi.org/10.3897/aca.4.e65142.

Young, J.N., Heureux, A.M., Sharwood, R.E., Rickaby, R.E., Morel, F.M., Whitney, S.M., 2016. Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms. J. Exp. Bot. 67, 3445-3456. https://doi.org/10.1093/jxb/erw163.

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. Appl. Environ. Microbiol. 73, 5261-5267. https://doi.org/10.1128/AEM.00062-07.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y.,

Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

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

Conclusions

1. This thesis is the first to evaluate the applicability of DNA metabarcoding of benthic diatoms (using chloroplast-encoded *rbcL*) to assess the ecological status of Catalan rivers for the EU WFD; in fact, it was the first such study in Spain. Comparisons of the biotic index (IPS values) and the ecological status classes derived from them between the traditional method (LM-based morphology) and metabarcoding (based on high throughput sequencing [HTS] of DNA) gave very good correspondence between the two methods. Thus, DNA metabarcoding of diatoms, even in its current state, constitutes an efficient and reliable alternative to traditional morphology-based analyses for WFD biomonitoring of Mediterranean rivers of the north-eastern Iberian Peninsula.

2. The sensitivity analysis developed in Chapter 1 showed that one reason for such a good correspondence was that many of the species that had most influence on the IPS values in Catalan rivers are present in the diatom reference sequence database.

3. A complete sequence reference database, though desirable, is unlikely to be realistic for many study areas. Our study shows that for biomonitoring purposes the crucial requirement is to have the sequences of the species that have most impact on the IPS and our sensitivity analyses can be considered a simple and effective tool to identify these (from LM count data).

4. In spite of the encouragingly good correspondence between LM and HTS approaches, some discrepancies were analysed in detail because of their possible consequences for river management. Some of the discrepancies were found to be due to misidentifications and overlooking in LM of a few species, which were better recovered by HTS. This was particularly the case with the weakly silicified diatom *Fistulifera saprophila*. Some other discrepancies were probably due to differences in *rbcL* copy number per cell, as has been suggested previously in other similar studies.

5. Applying a combined morphological-metabarcoding (*rbcL*) approach to the benthic diatom communities of Ebro bays revealed very high diversity and many undescribed species.

6. DNA metabarcoding in these shallow coastal habitats is still far from ready to be applied as an effective alternative to microscopy, since the low sequence coverage of coastal

benthic diatom species in the reference database means that many DNA reads cannot be assigned to species.

7. We found strong circumstantial evidence that one very abundant diatom of the biofilms of Ebro bays (*Thalassiosira profunda*) was present without frustules, very likely as an endosymbiont. This, together with the capacity to detect small or/and delicate diatom species that are often missed by LM, illustrates the complementarity of LM and metabarcoding approaches.

8. *In silico* analyses on a large benthic diatom metabarcoding dataset indicated that the choice between two short and similar diatom *rbcL* barcodes, overlapping in a common 263-bp region and differing in the presence or absence of a 68-bp tail at the 5' end, have very few implications for WFD ecological status assessments.

9. Despite the irrelevance of the barcode choice for WFD purposes, our analyses indicated that the longer *rbcL* marker is preferable for ecological and biogeographical studies, as the additional nucleotide variability provided by the 68-bp tail was shown to reduce the number of false negatives and false positives and, in some particular cases, allowed species-level classification of some genetic variants that could not be unambiguously identified on the basis of the shared 263 bp region. This was particularly the case for genetic variants of *Surirella brebisonni*, *Halamphora montana* and *Fragilaria agnesiae*.

10. Primers designed to amplify the short 263-bp region of freshwater diatoms can also amplify taxa of the phyla Ochrophyta and Chlorophyta, some of which are rarely recorded groups and species with economic and ecological relevance. However, the non-diatom sequencing reads generated are a minor proportion of the total, reflecting that specific primers or different markers should be used to study non-diatom groups via metabarcoding If they are the principal targets of study, rather than diatoms.

11. 263-bp *rbcL* variants within the *Achnanthidium minutissimum* and *Fistulifera saprophila* species complexes differed in their ecological preferences, illustrating the important extra potential of being able to analyse diatom communities at the haplotype level. Our data suggest that the broad ecological tolerances assumed for these complexes are the result of summing the specific ecological preferences of each variant and the impossibility of discriminating between them when using non-molecular approaches. These findings also have implications from a biomonitoring perspective since they reflect that assigning the same indicator value to all the variants within a species complex is suboptimal.

12. We found that 263 bp *rbcL* variants of many species spread across regions of Europe, North America, the Indian Ocean and/or Asia, suggesting ubiquitous dispersal of individuals. At the same time, the distribution of other *rbcL* variants was geographically restricted to specific regions and, in addition, it was observed that at local scales *rbcL* variants varied greatly in their distribution. It seems that, although individuals may disperse over large geographical distances, stochastic events of colonisation and extinction, combined with environmental variation at local scales, are shaping the distribution of species and individual *rbcL* variants.

13. Our studies on a large dataset of benthic diatom samples evidenced a very high intraspecific heterogeneity of the 263-bp marker among freshwater diatom species and the existence of 4 main phylogeographic patterns (defined on the basis of the number, dominance and spatial structure of 263-bp *rbcL* variants) that were common among species. Furthermore, our results showed that centric species showed significantly fewer *rbcL* variants than pennate species, which may be related to differences between centric and pennate species in reproductive mode (oogamous vs isogamous) and chloroplast inheritance patterns (biparental vs uniparental). However, to reach definitive conclusions on the causes of different phylogeographic patterns and high intraspecific heterogeneity observed among species, studies covering a larger number of centric species are needed.

Annexes

Annex 1

Haplotype networks

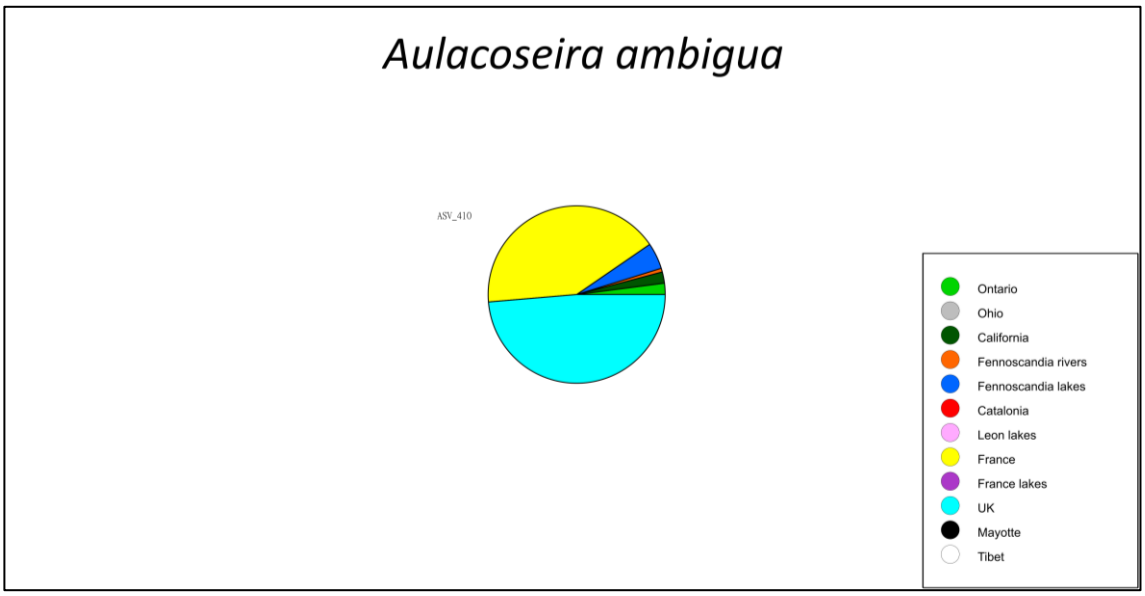
This annex includes the TCS haplotype networks that were used in chapter 5 to define the 4 main types of phylogeographic patterns found in freshwater diatoms. Each circle represents a unique ASV (the ASV label is given next to the circle), and its size is proportional to the number of samples in which the ASV was identified. Colour codes represent the geographical locations where the ASVs were found. The haplotype networks have a comparable scale (only minor adjustments have been made in some cases) to facilitate distinguishing differences in ASV occurrence across the regions studied. Small black circles represent hypothetical variants automatically inferred and black crosshatches indicate the number of nucleotide differences between ASVs.

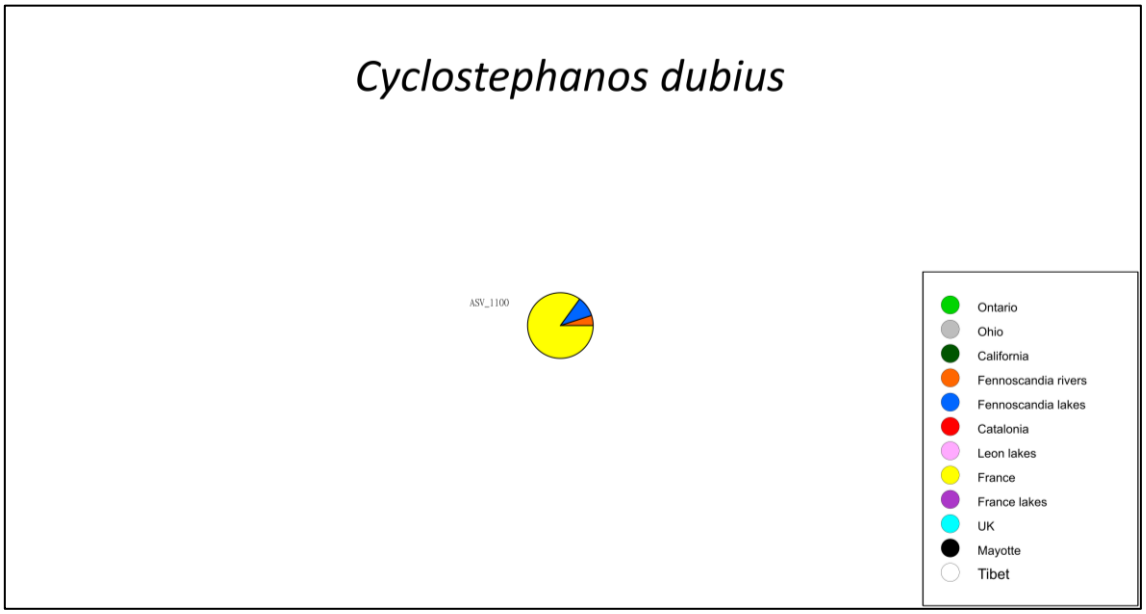
Note that the header of the page contains information on the species represented (centric, pennate or araphid), its corresponding phylogeographic pattern and the main characteristics of this pattern.

The haplotype networks are shown in the following order: Firstly, according to the type of phylogeographic pattern in which they were classified. Secondly, according to the type of diatom (i.e. centric, pennate and araphid, respectively). And lastly, according to alphabetical order (see Index below).

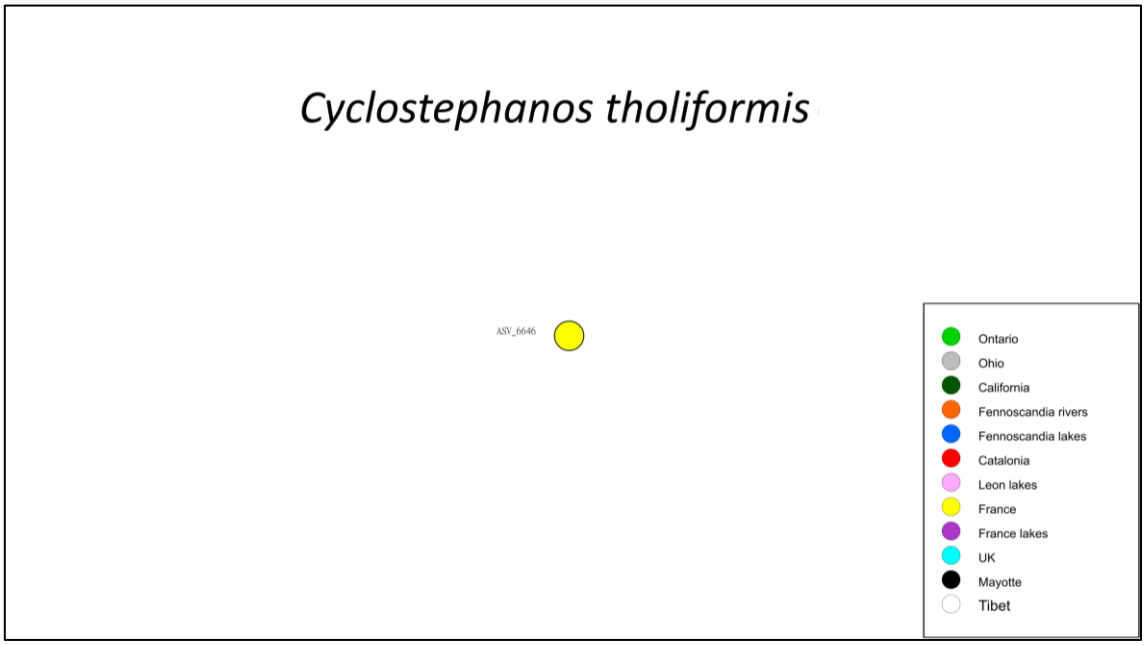
Index

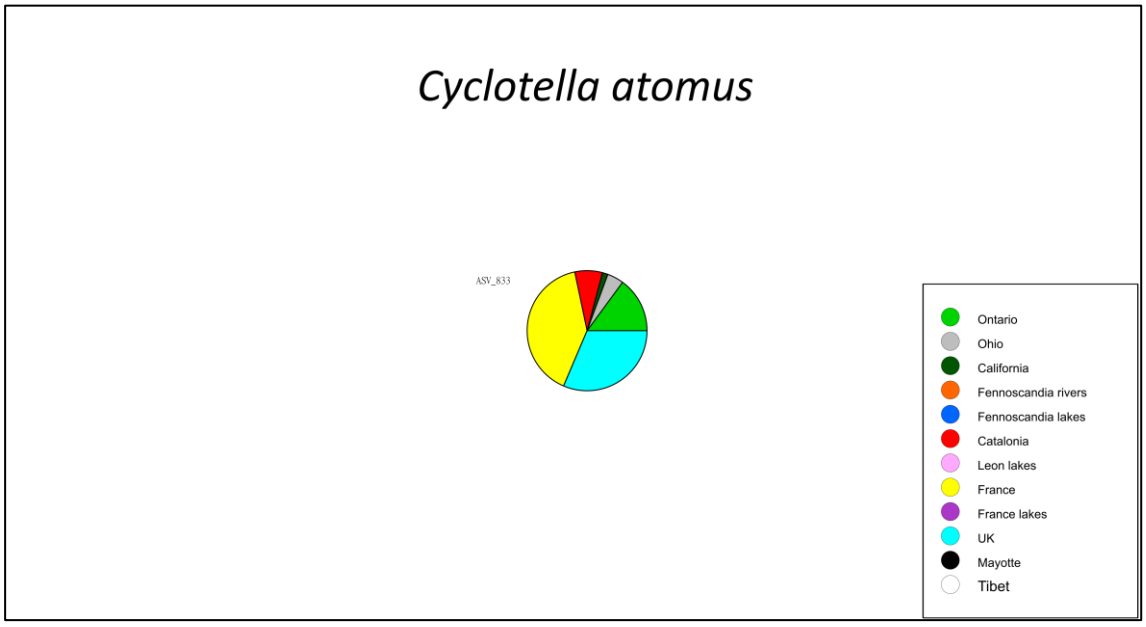
Endemic species Pattern I 1 ASV per species



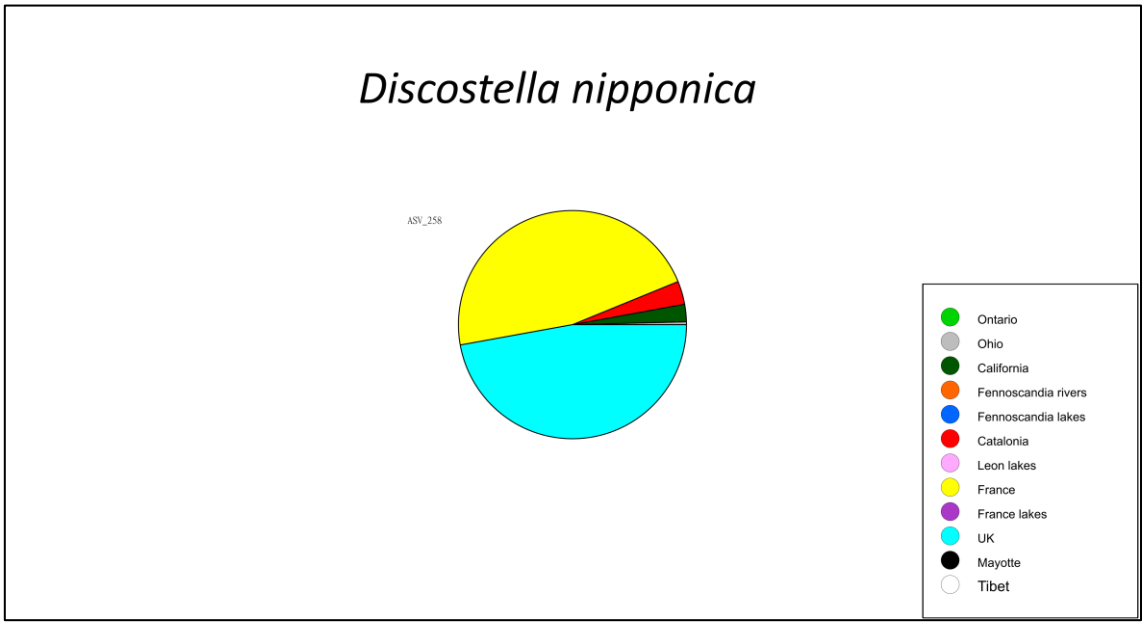


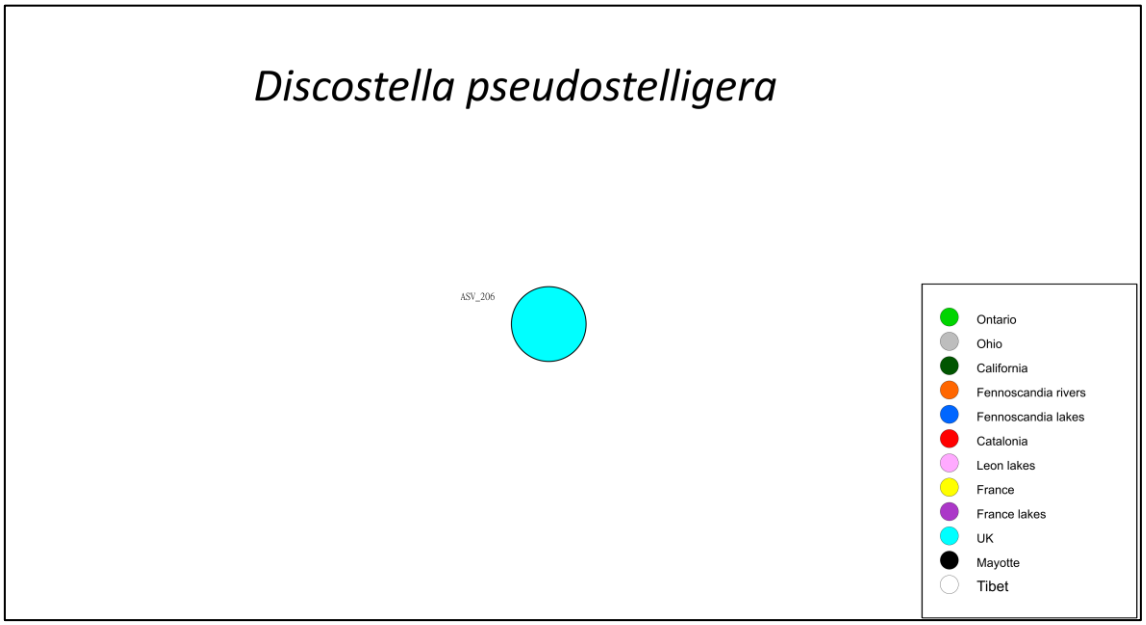
Centric species Pattern 1 1 ASV per species



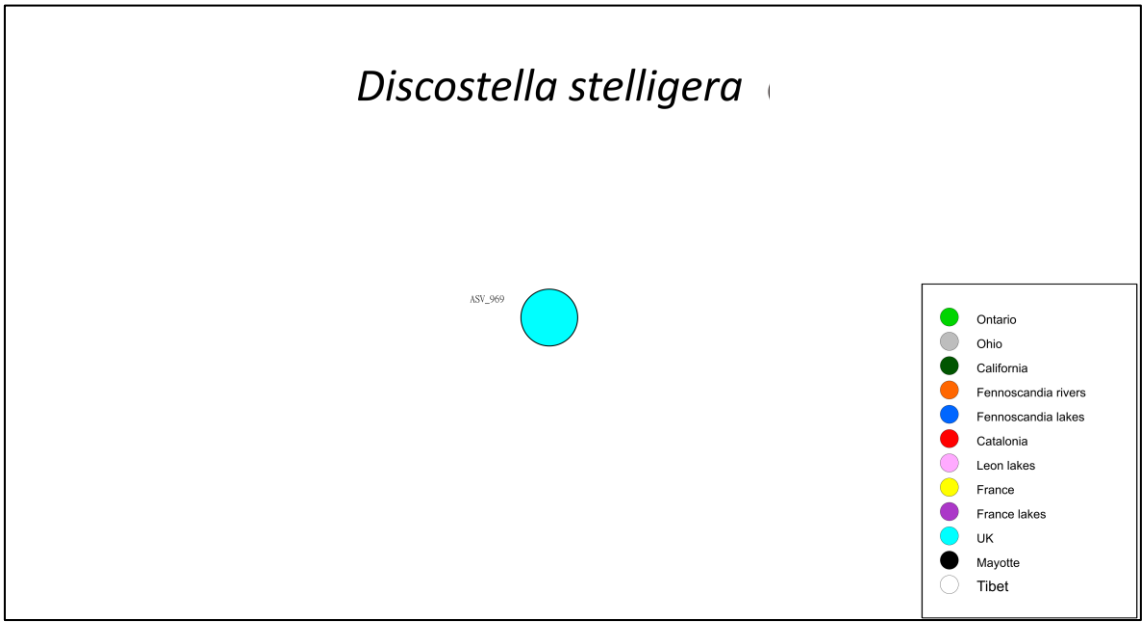


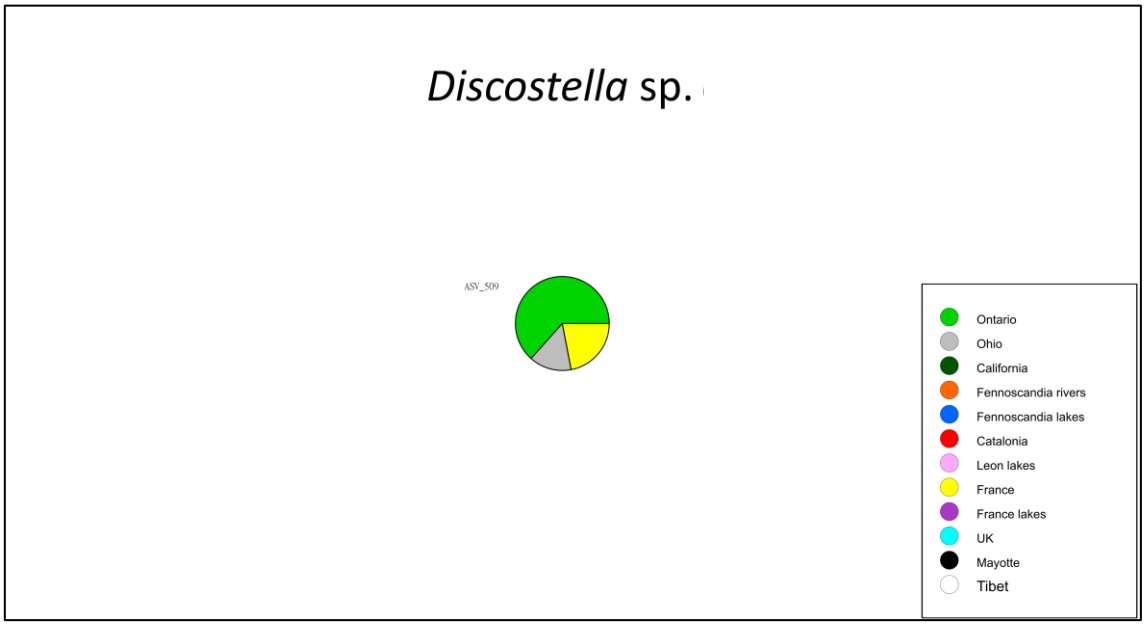
Centric species Pattern 1 1 ASV per species





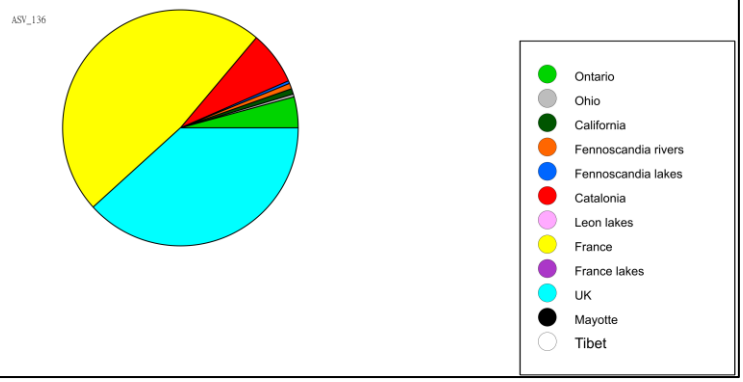
Centric species Pattern I 1 ASV per species



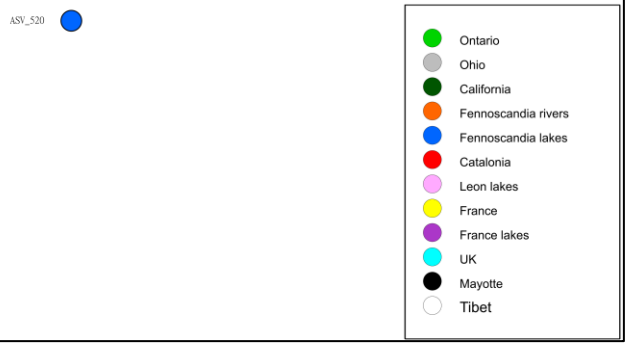


Centric species Pattern I 1 ASV per species

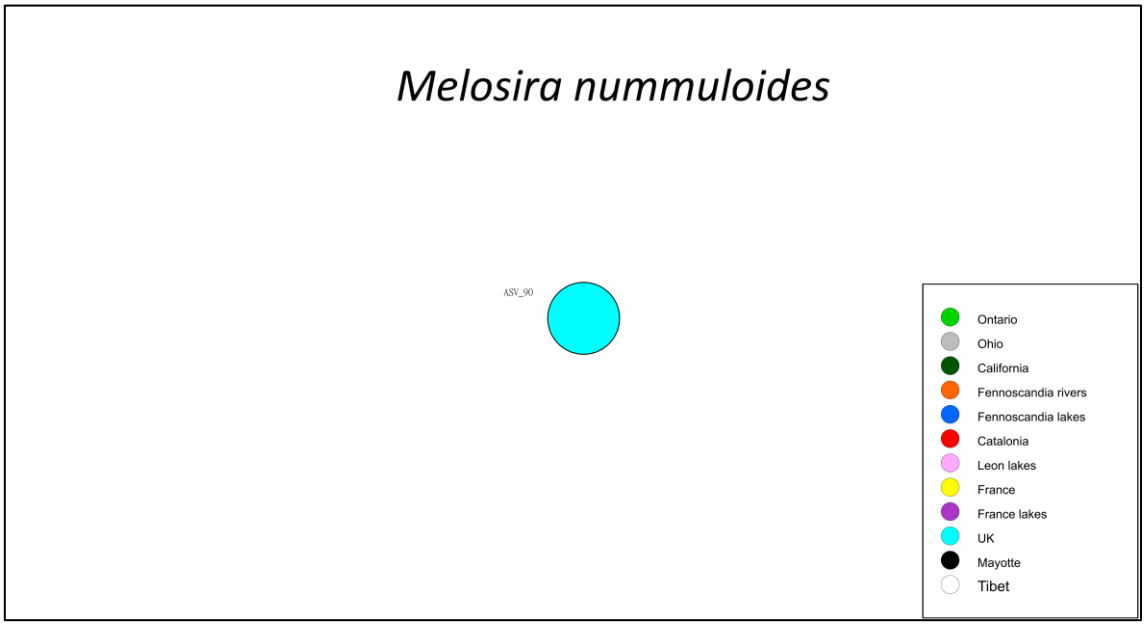
Discostella woltereckii

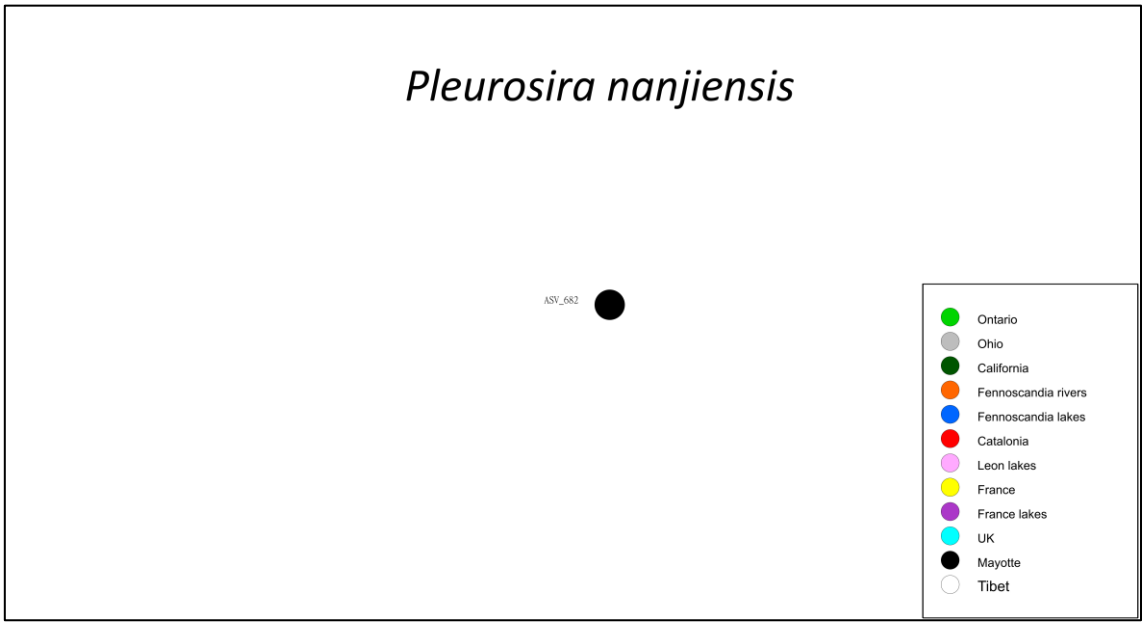


Ellerbeckia sp.



Centric species Pattern 1 1 ASV per species



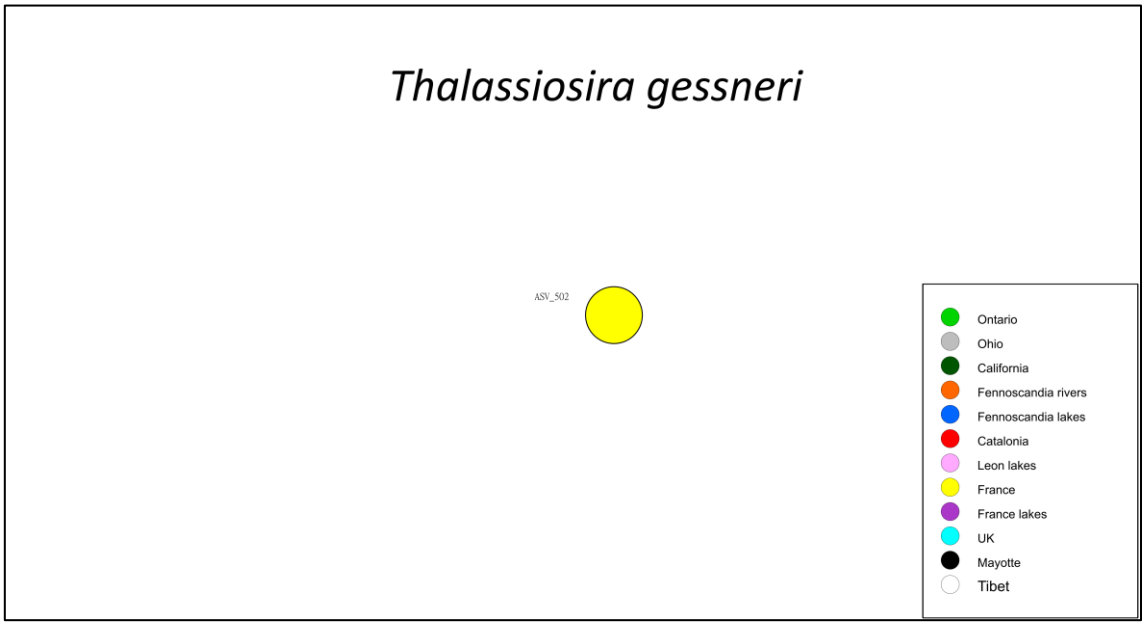


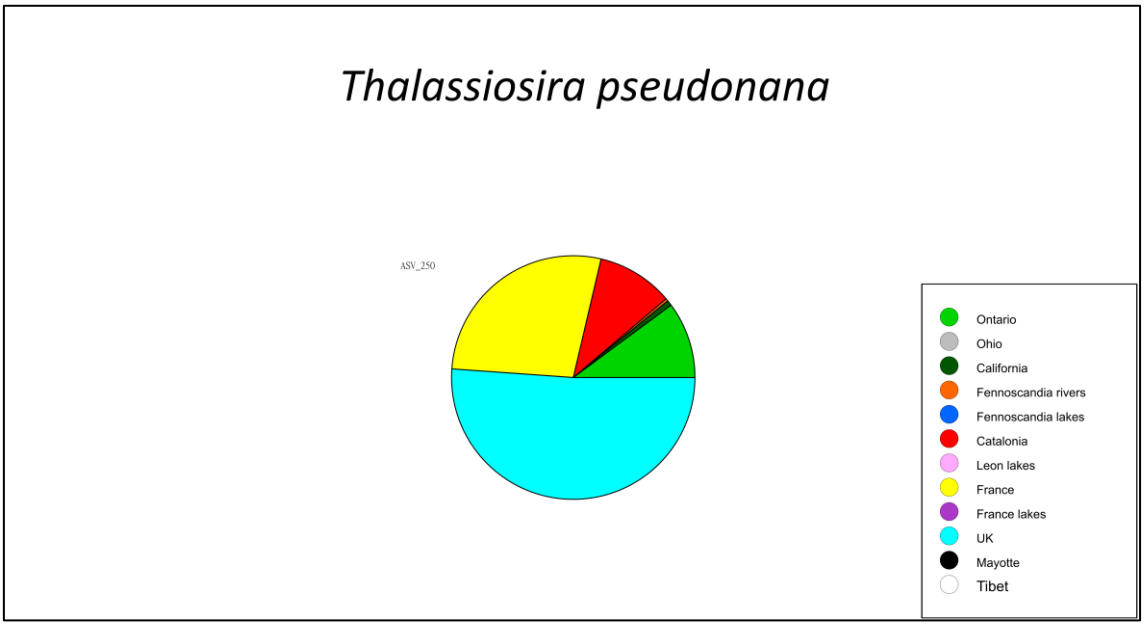
Centric species Pattern 1 1 ASV per species

Skeletonema potamos

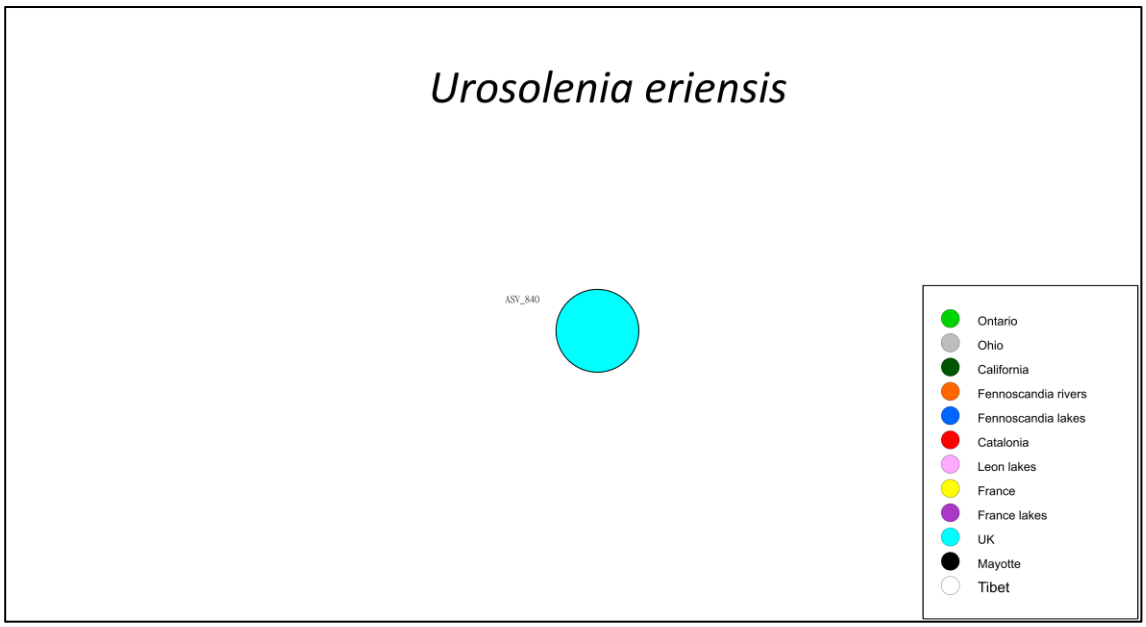
Skeletonema subsalsum

Centric species Pattern 1 1 ASV per species

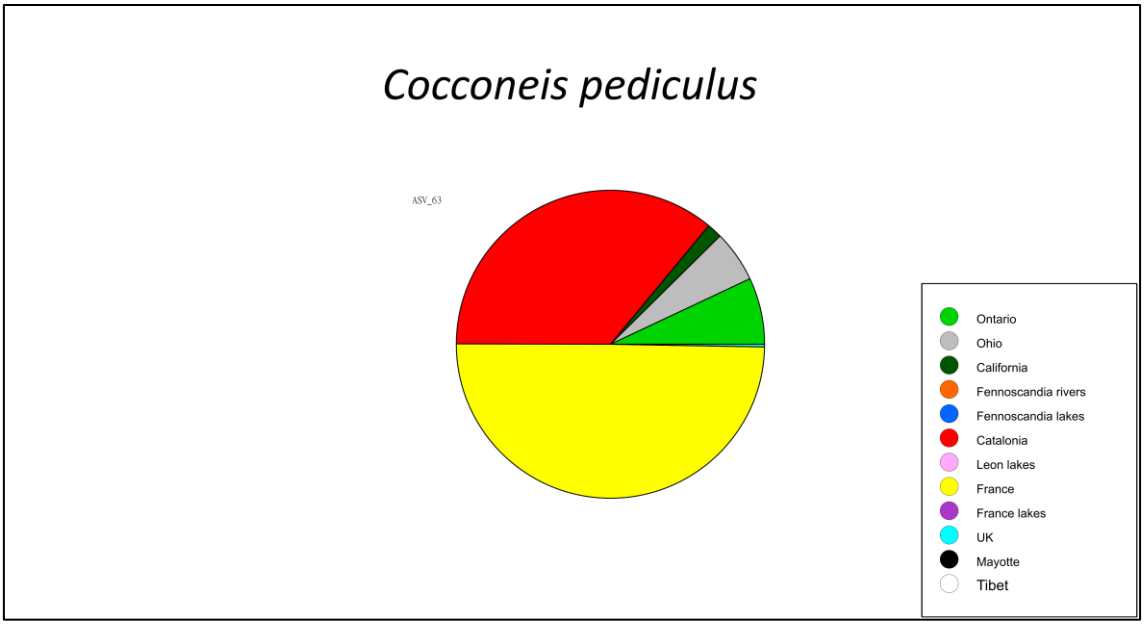


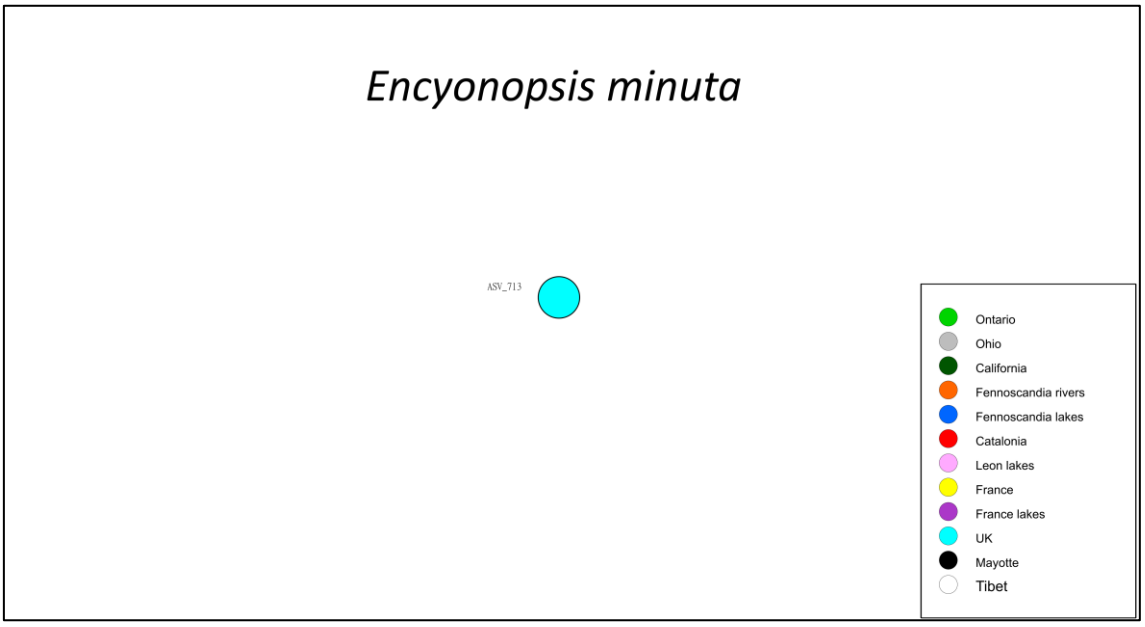


Centric species Pattern 1 1 ASV per species

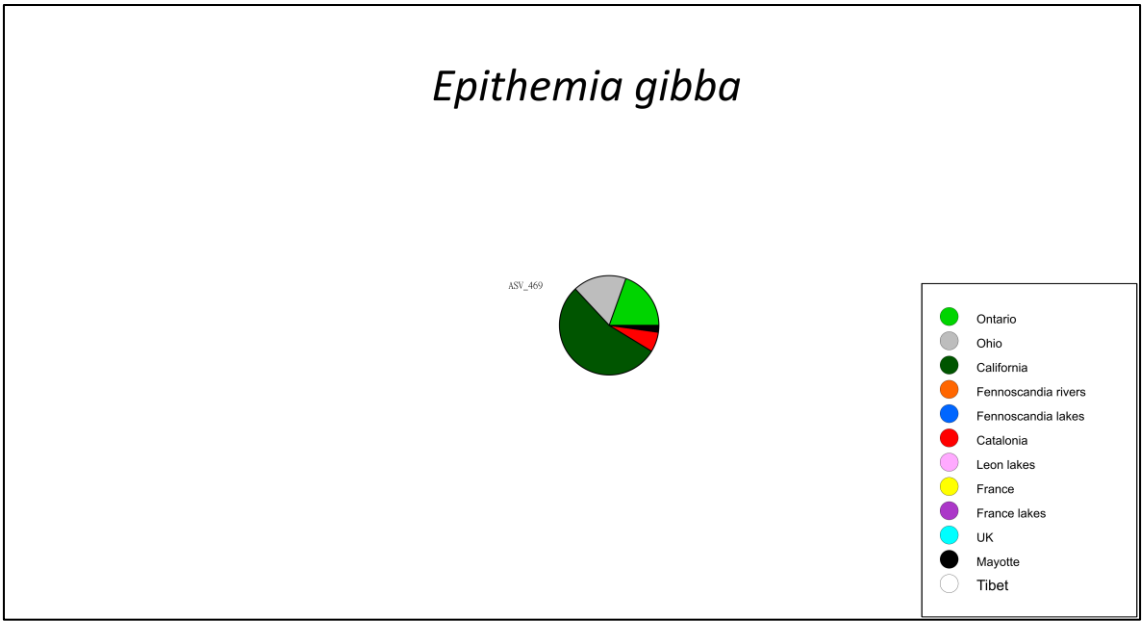


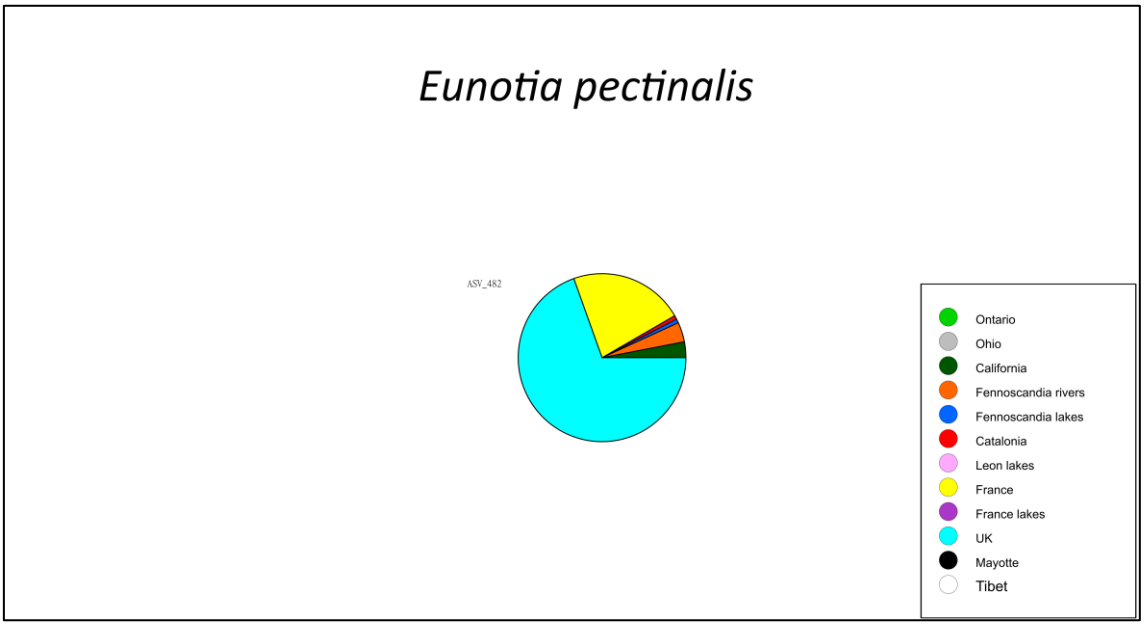
Rapid pennate species Pattern I 1 ASV per species



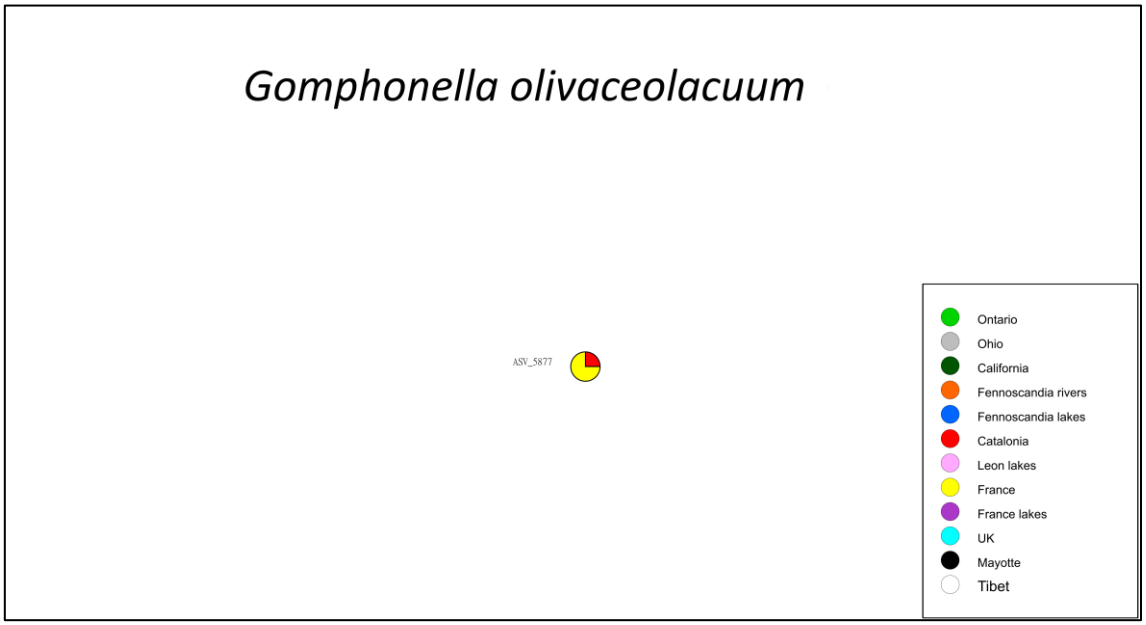


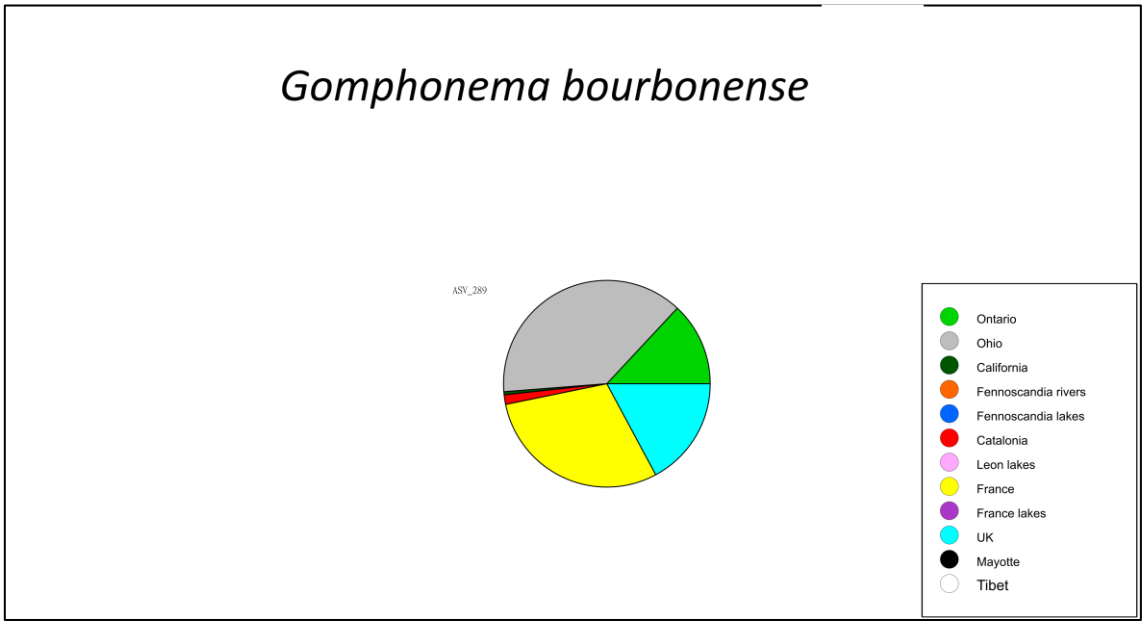
Rapid pennate species Pattern I 1 ASV per species

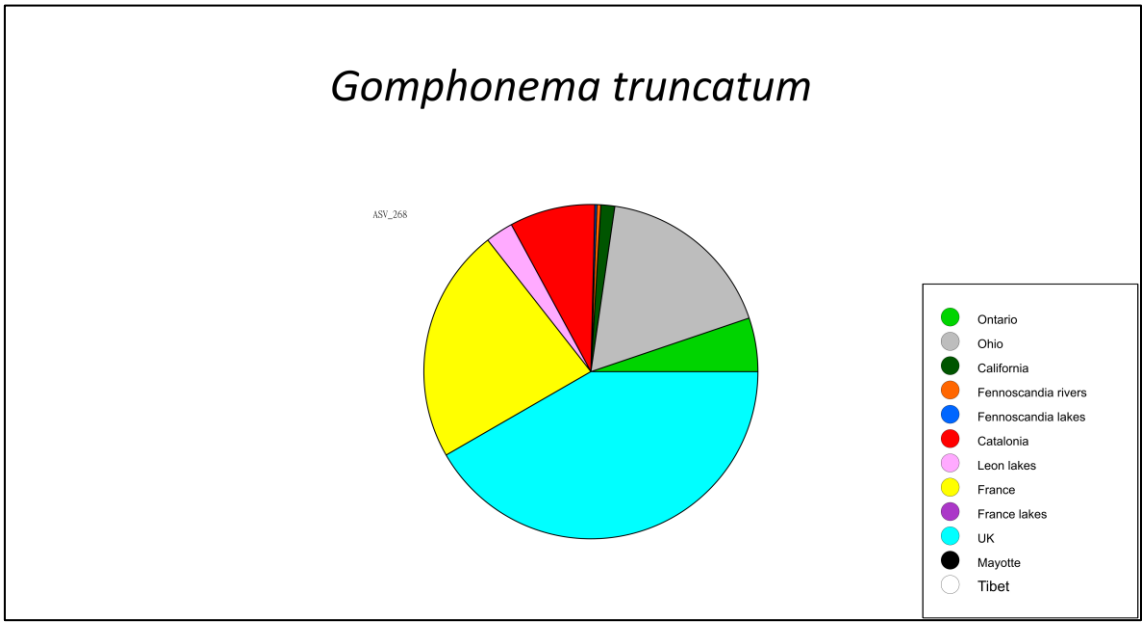


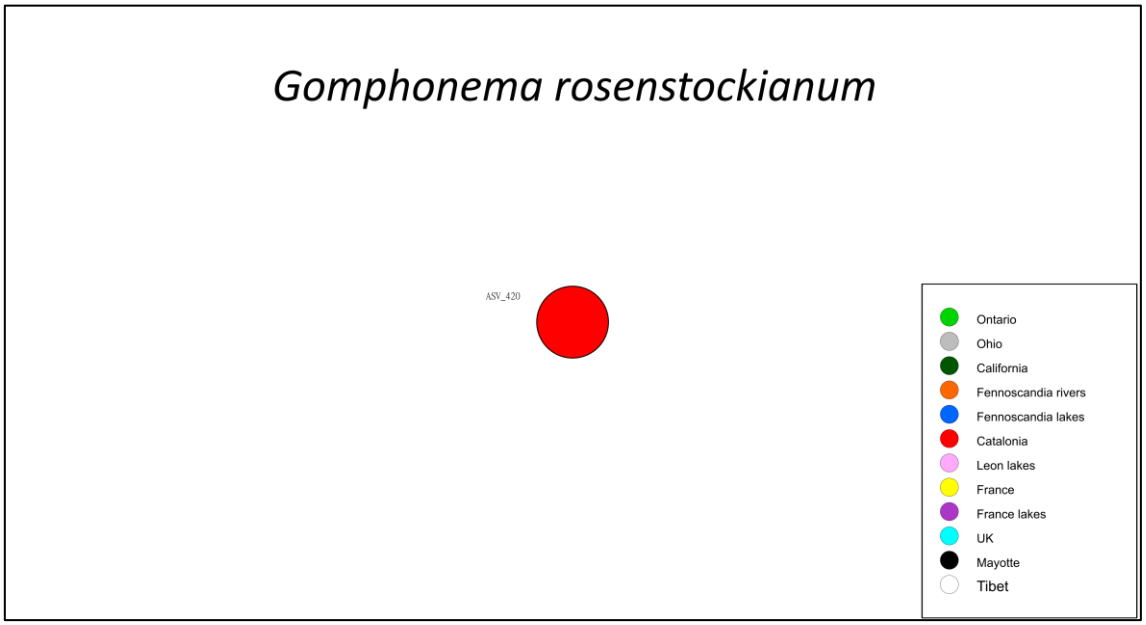


Raphid pennate species Pattern I 1 ASV per species

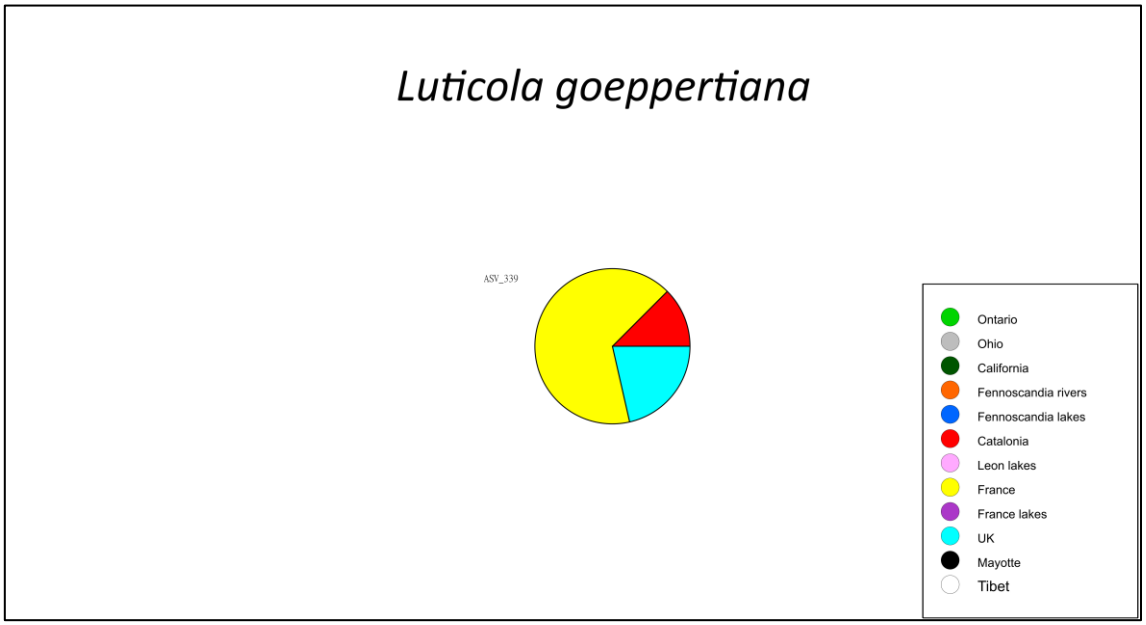


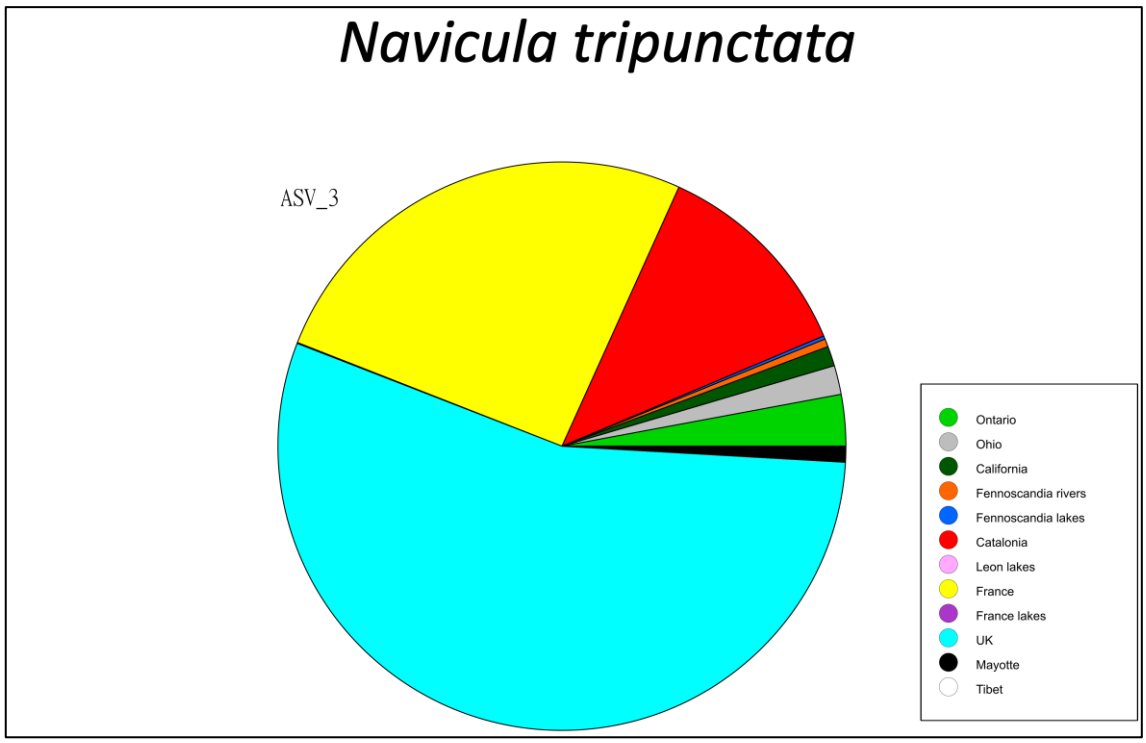



Gomphonema truncatum
ASV_268
Ontario
Ohio
California
Fennoscandia rivers
Fennoscandia lakes
Catalonia
Leon lakes
France
France lakes
UK
Mayotte
Tibet

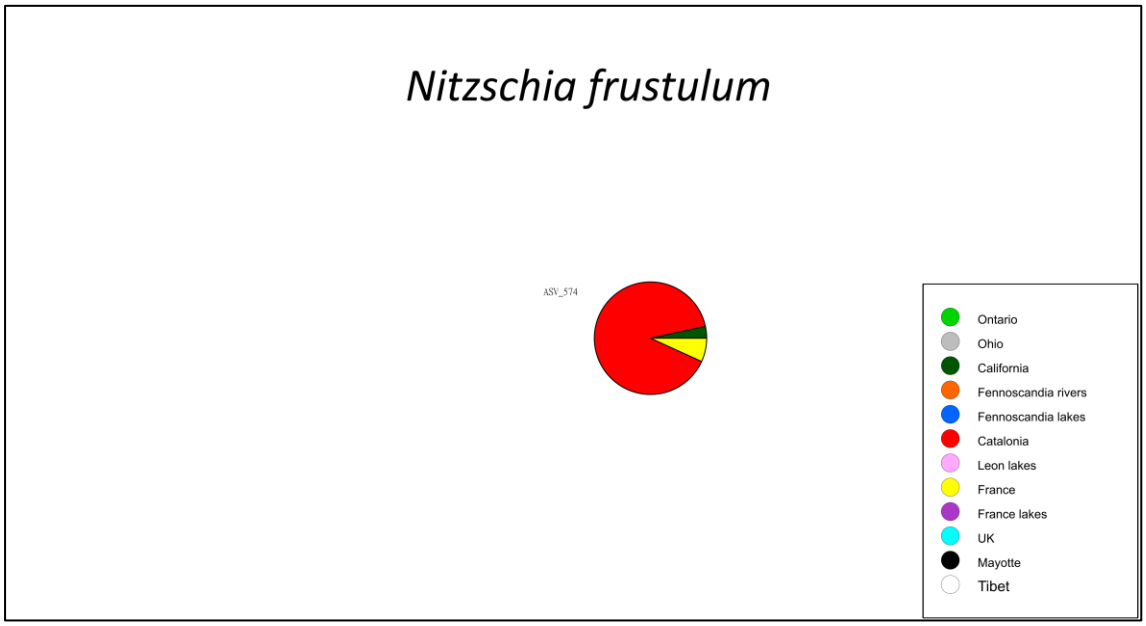

Gomphonema rosenstockianum
ASV_420
Ontario
Ohio
California
Fennoscandia rivers
Fennoscandia lakes
Catalonia
Leon lakes
France
France lakes
UK
Mayotte
Tibet

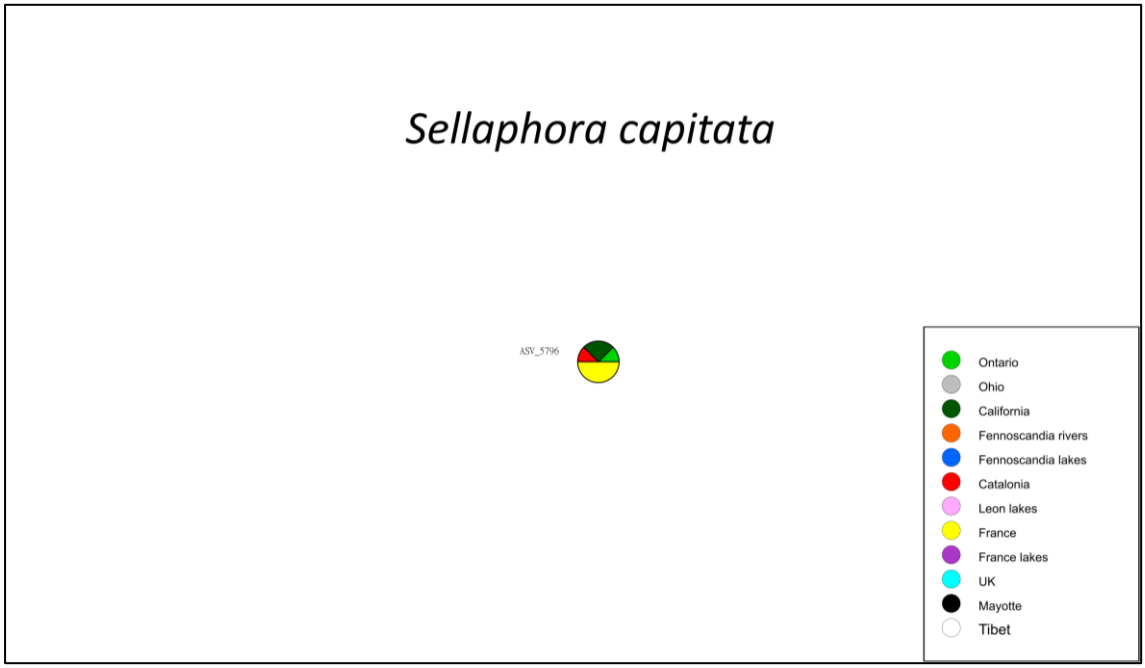
Raphid pennate species Pattern I 1 ASV per species



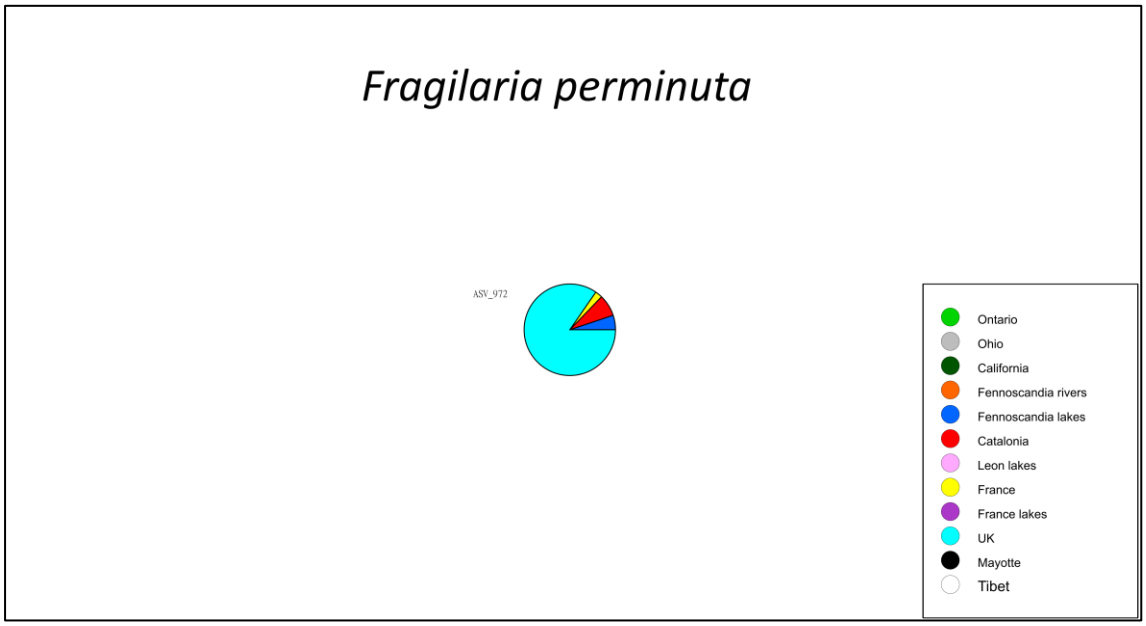


Raphid pennate species — Pattern I — 1 ASV per species





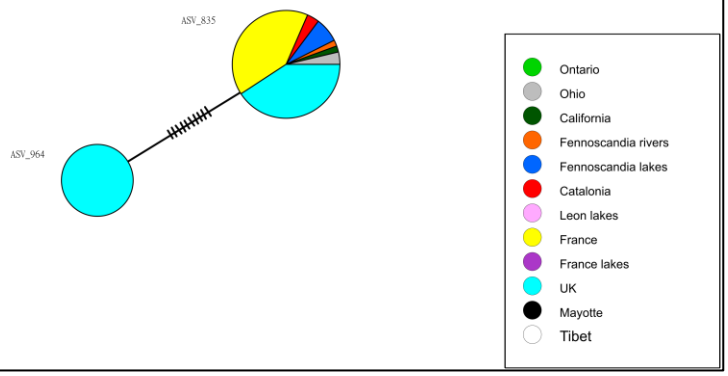
Araphid pennate species — Pattern 1 — 1 ASV per species



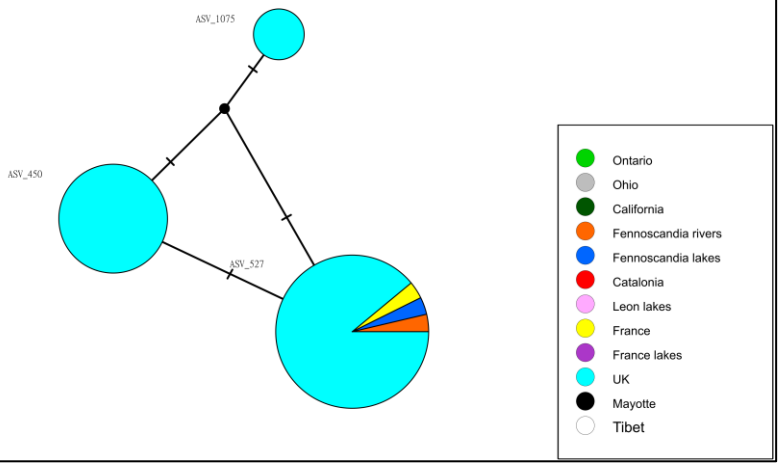
Centric species — Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs

Aulacoseira granulata



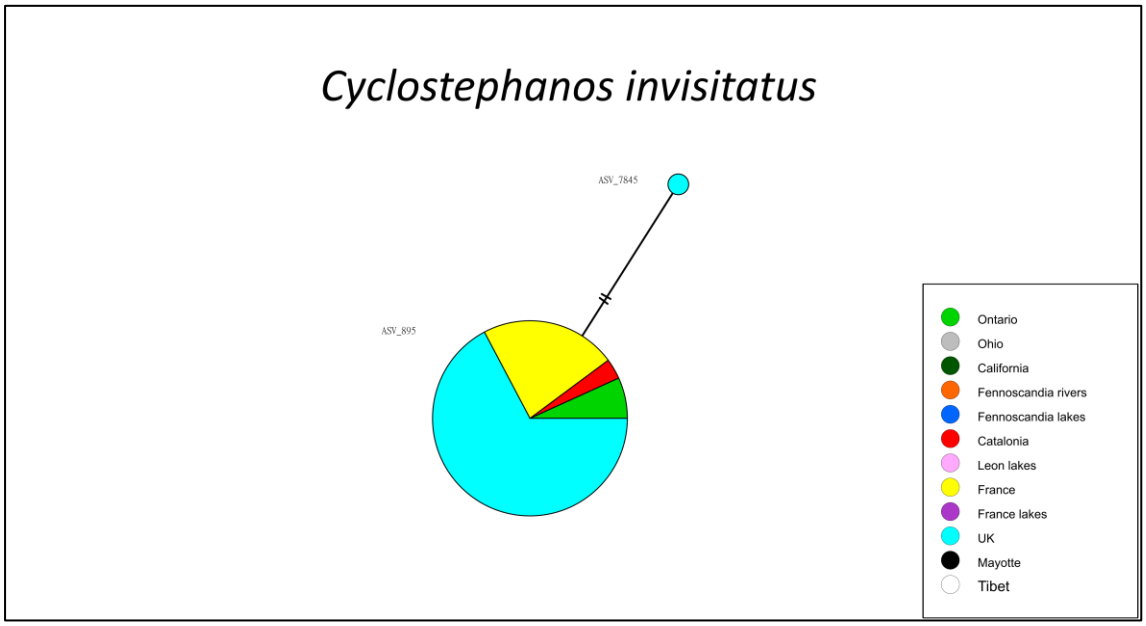
Aulacoseira subarctica

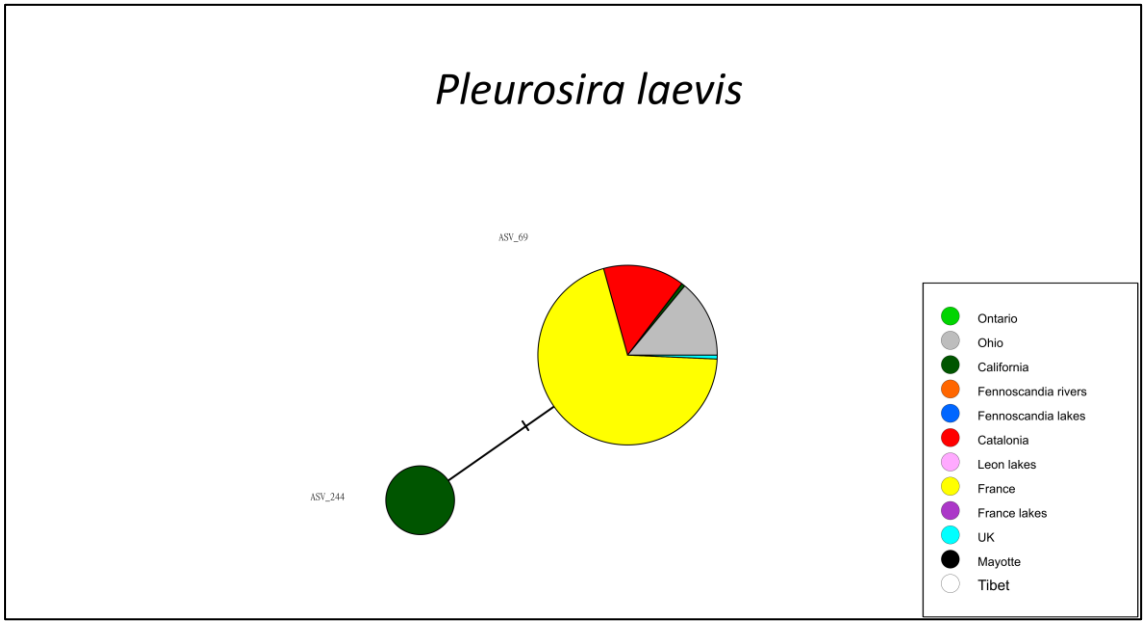


Centric species

Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs

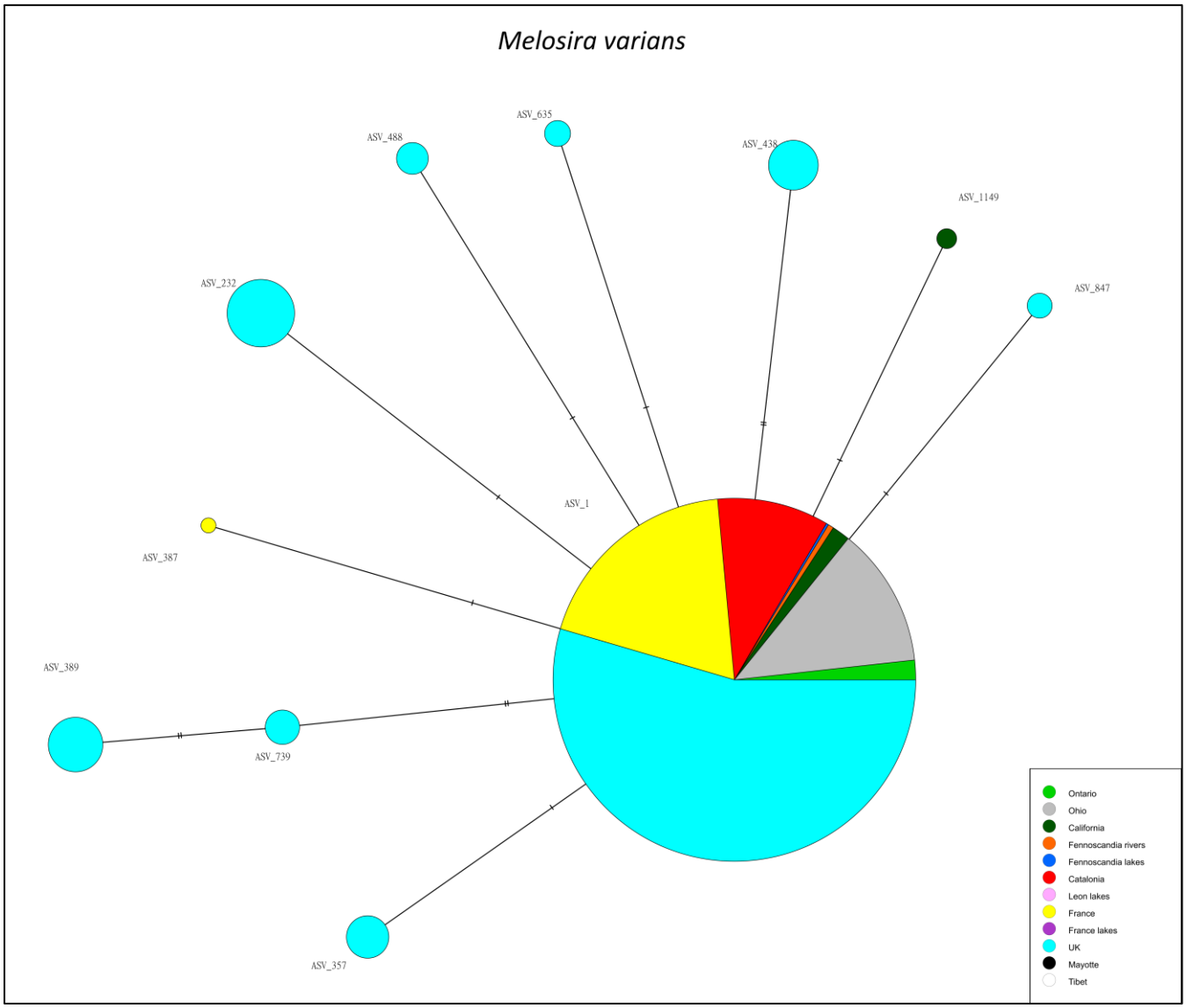




Centric species

Pattern II

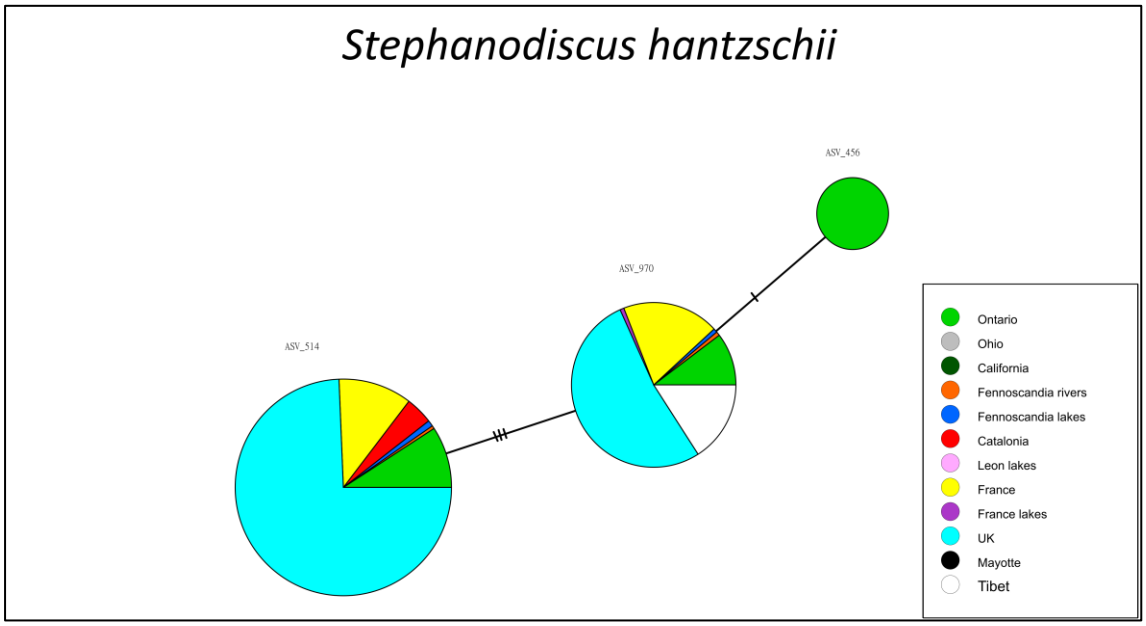
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



Centric species

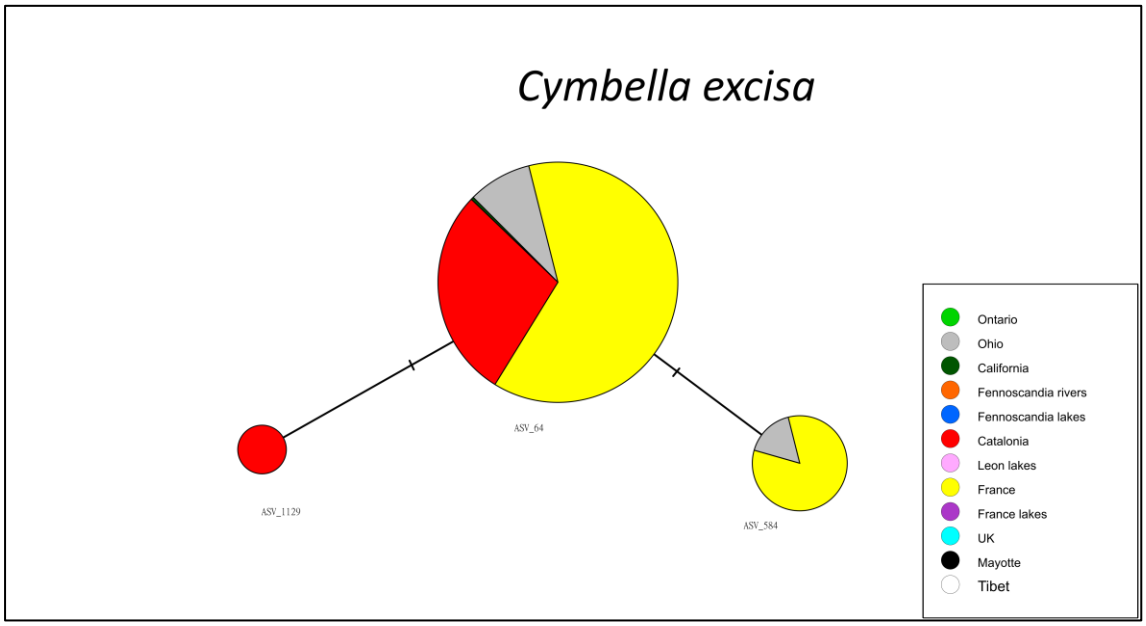
Pattern II

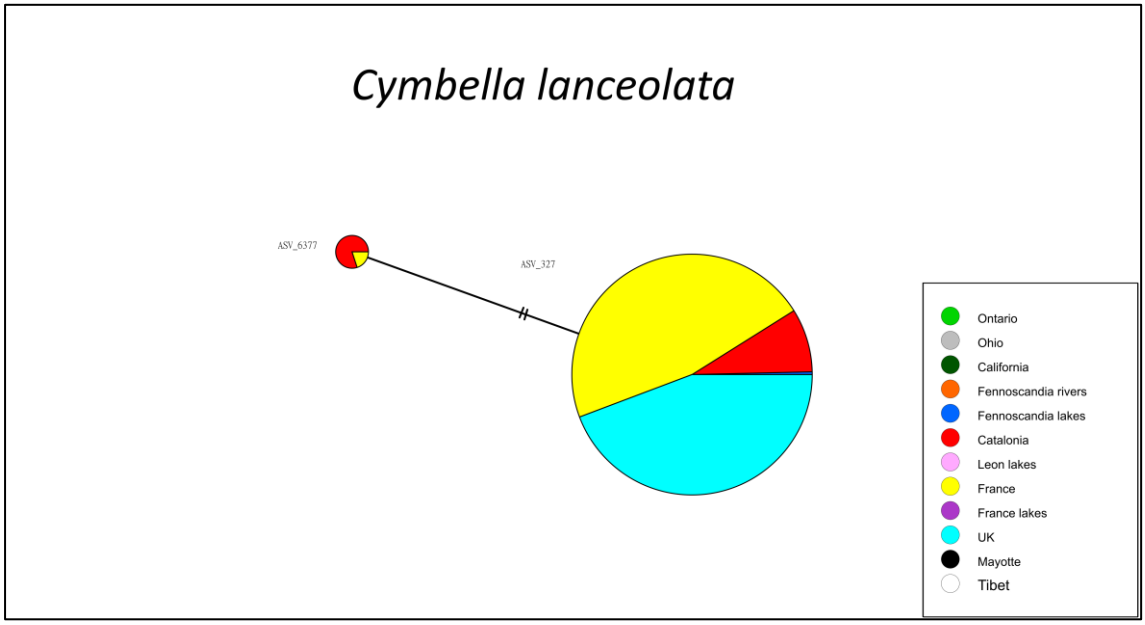
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs

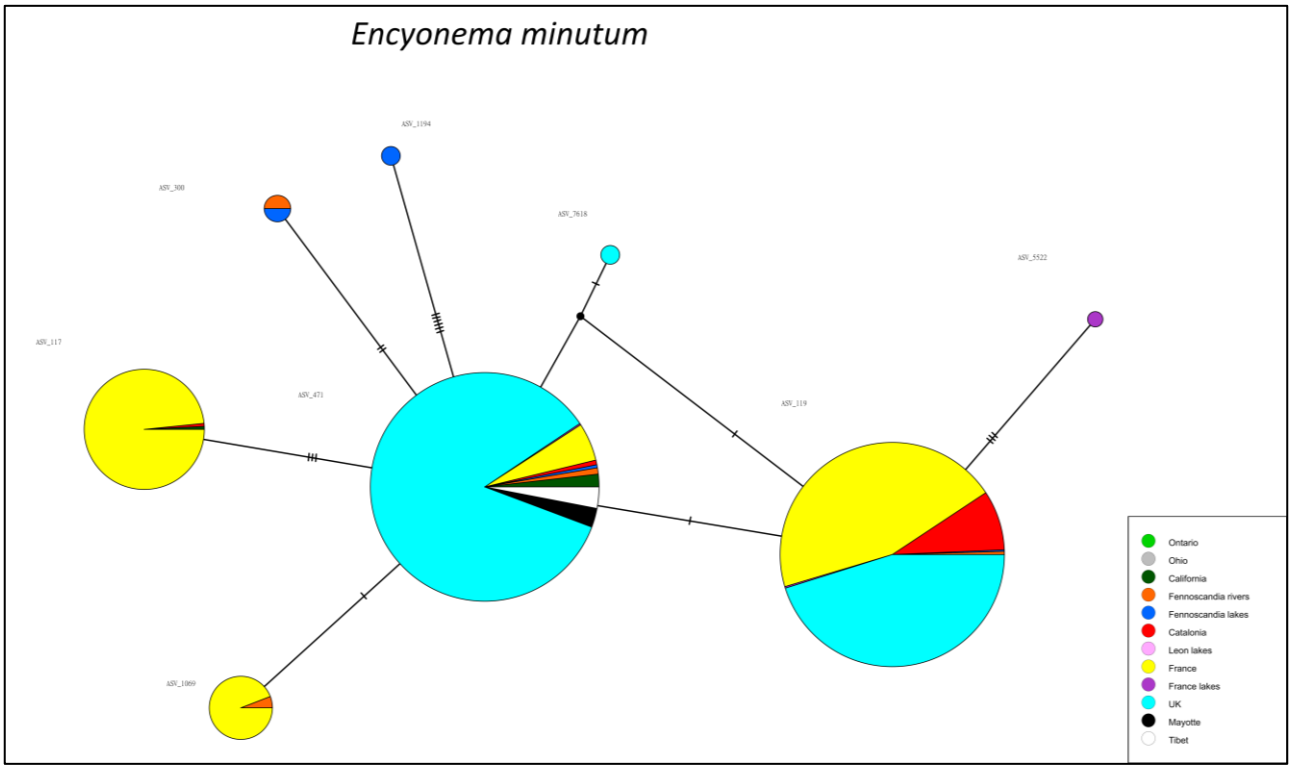


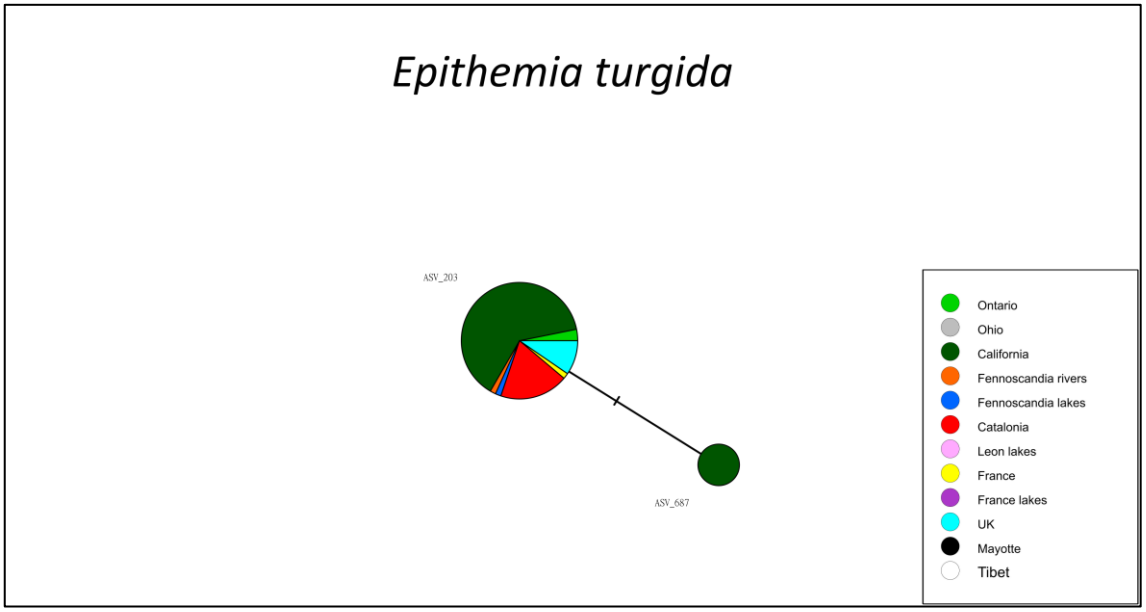


Rapid pennate species

Pattern II

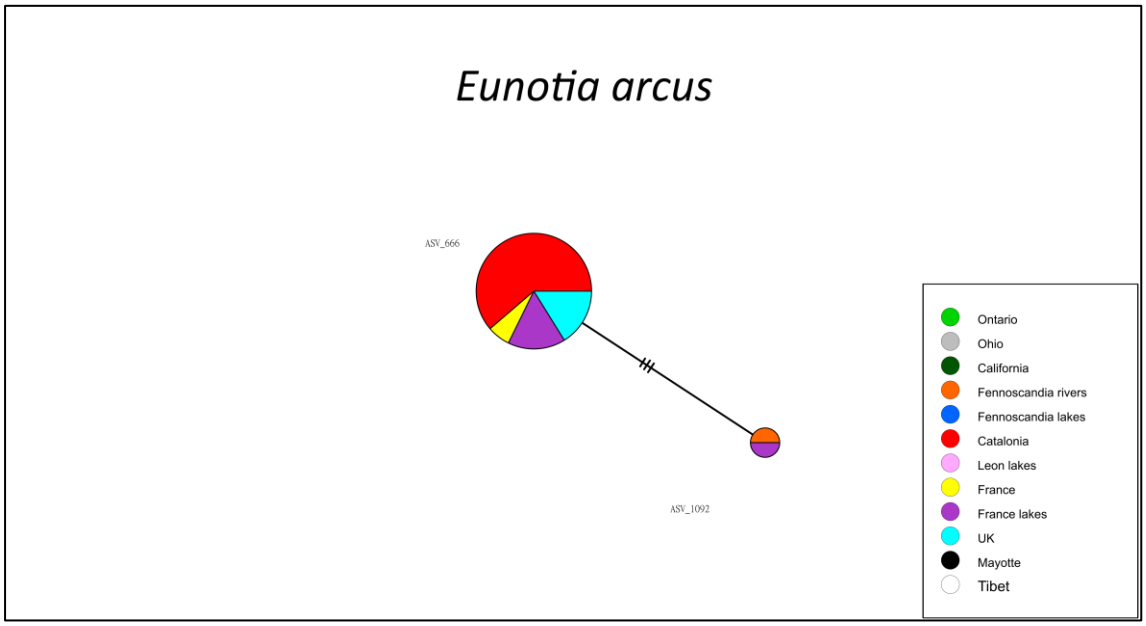
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs

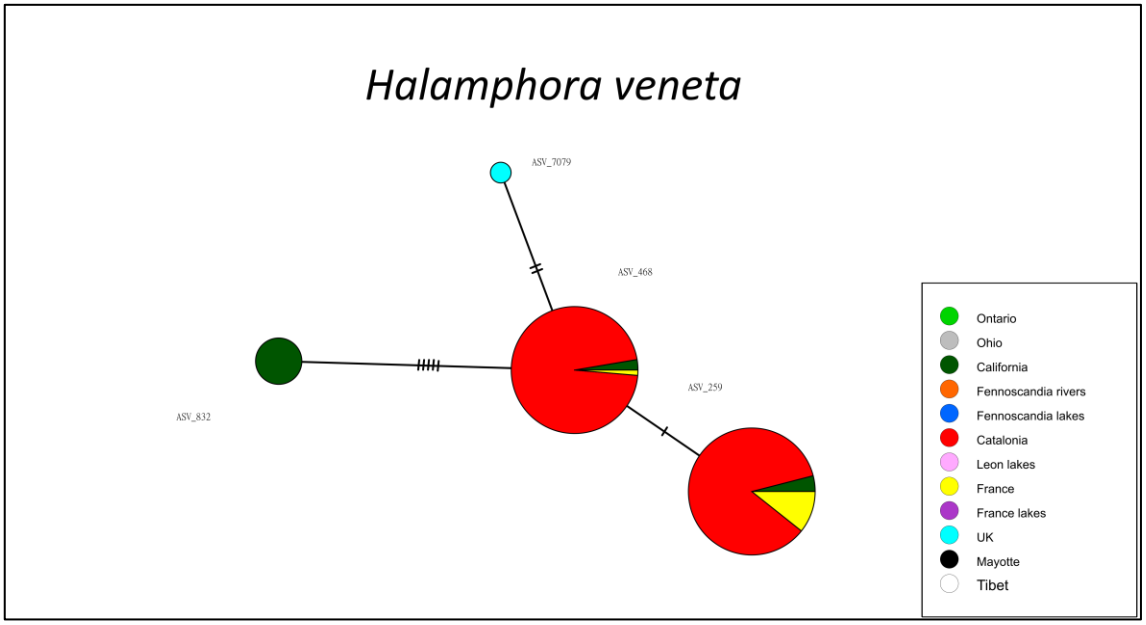




Rapid pennate species Pattern II

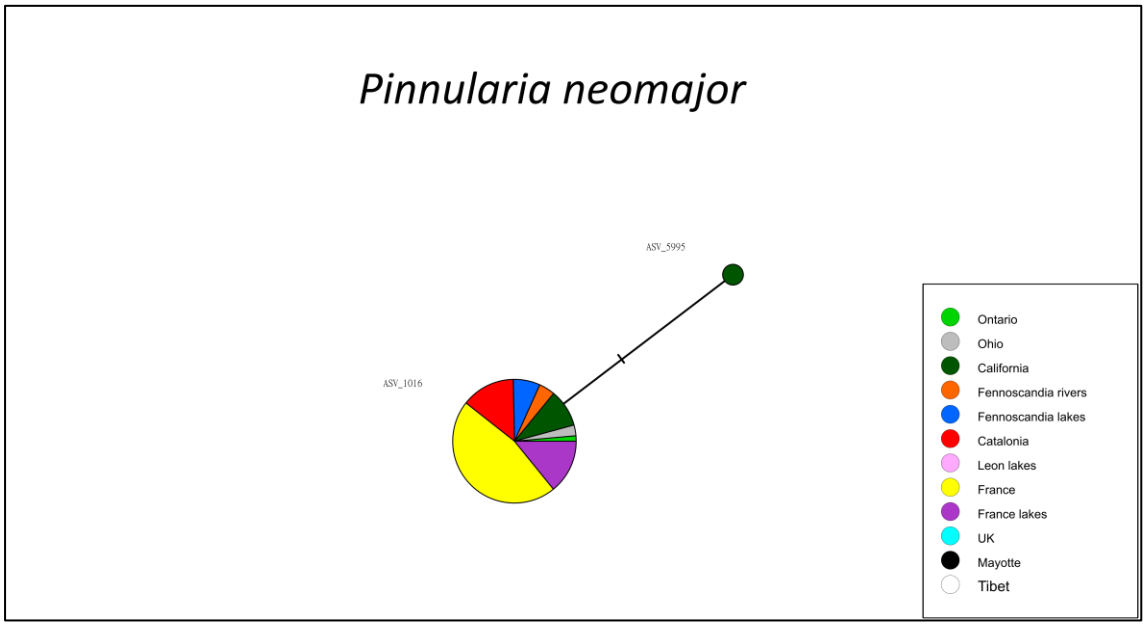
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs

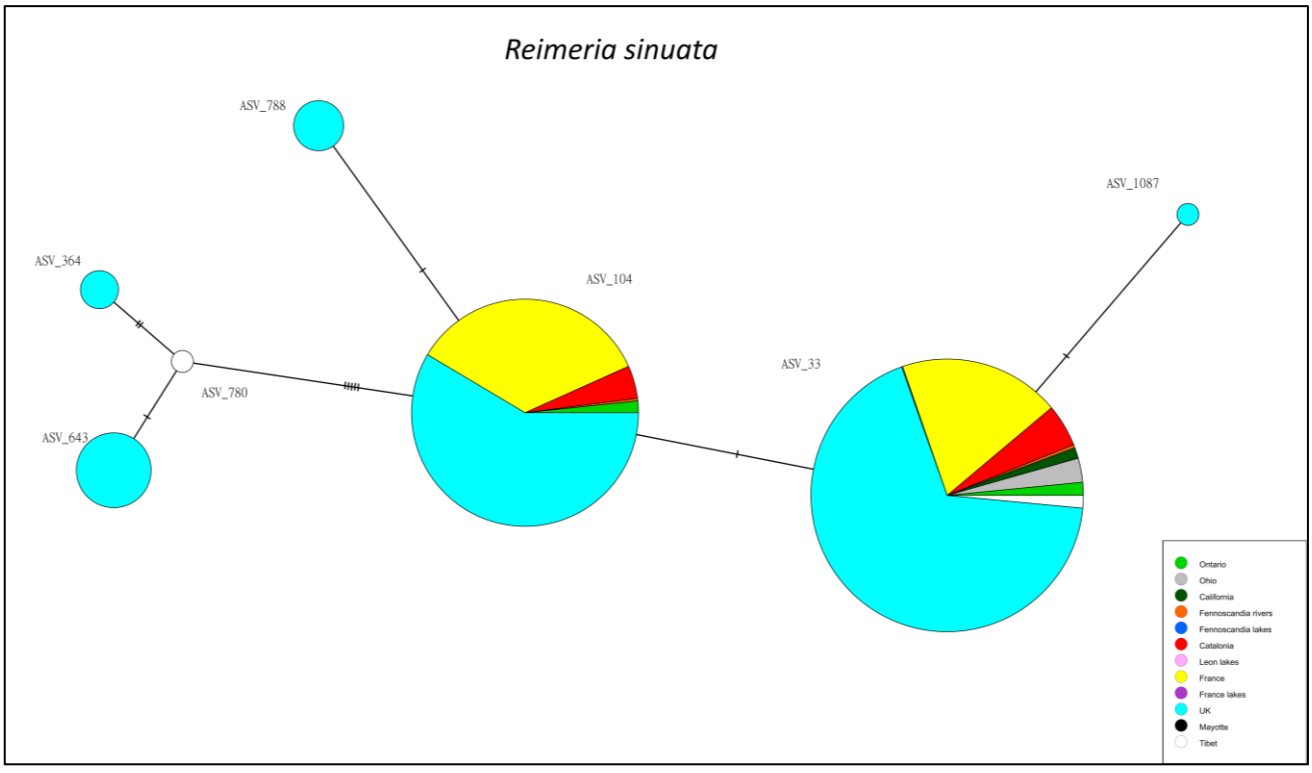




Rapid pennate species Pattern II

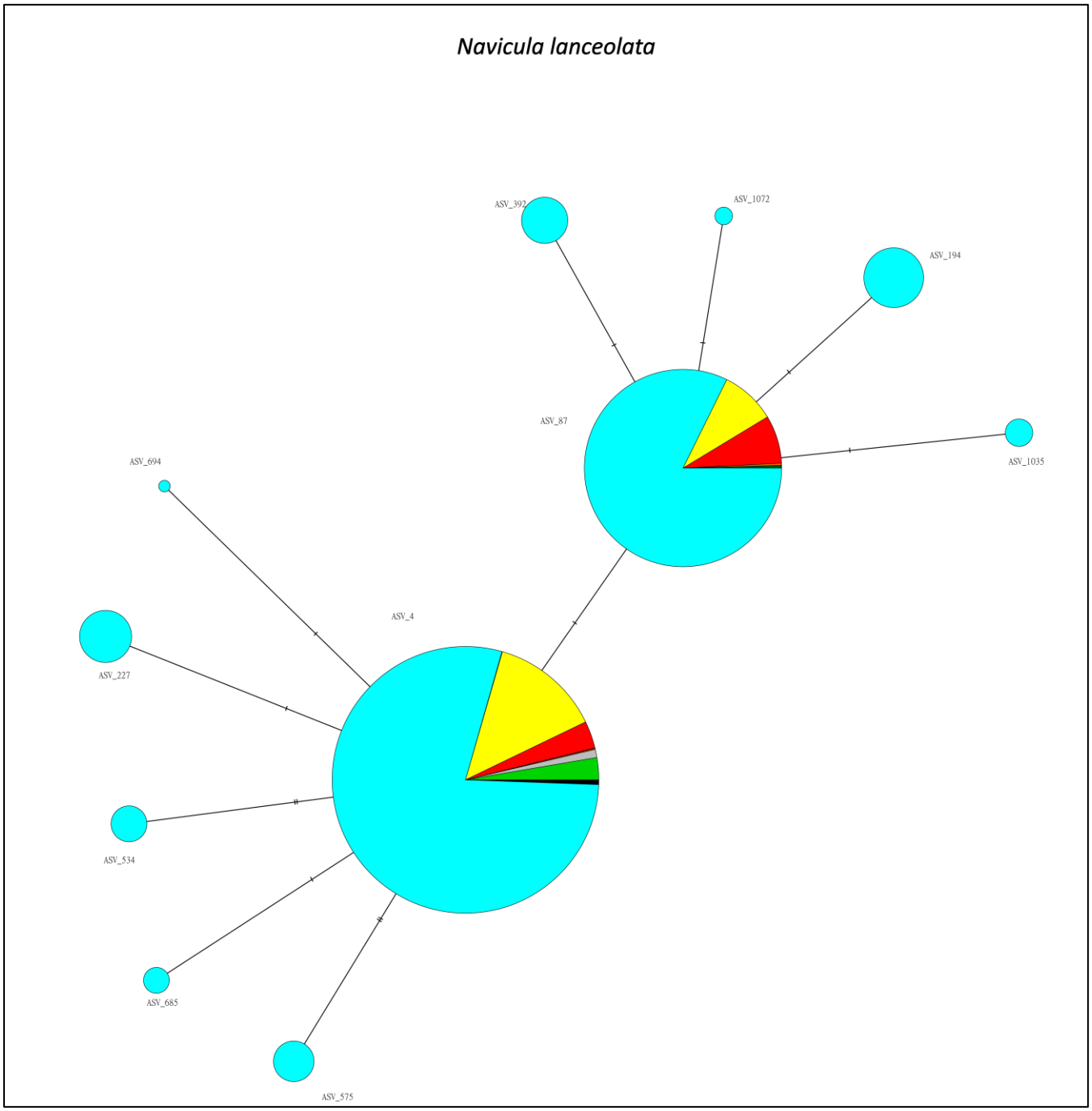
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs





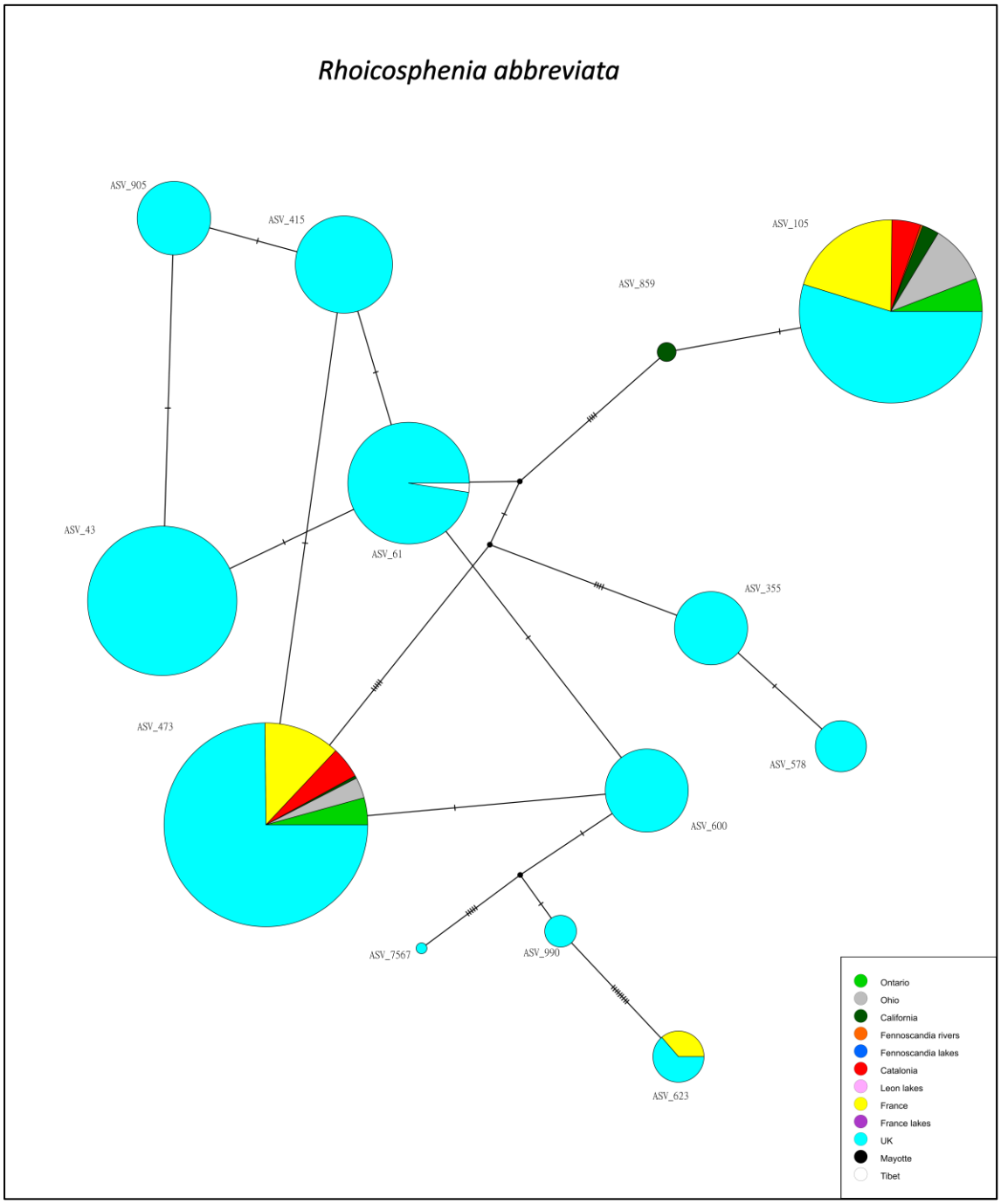
Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



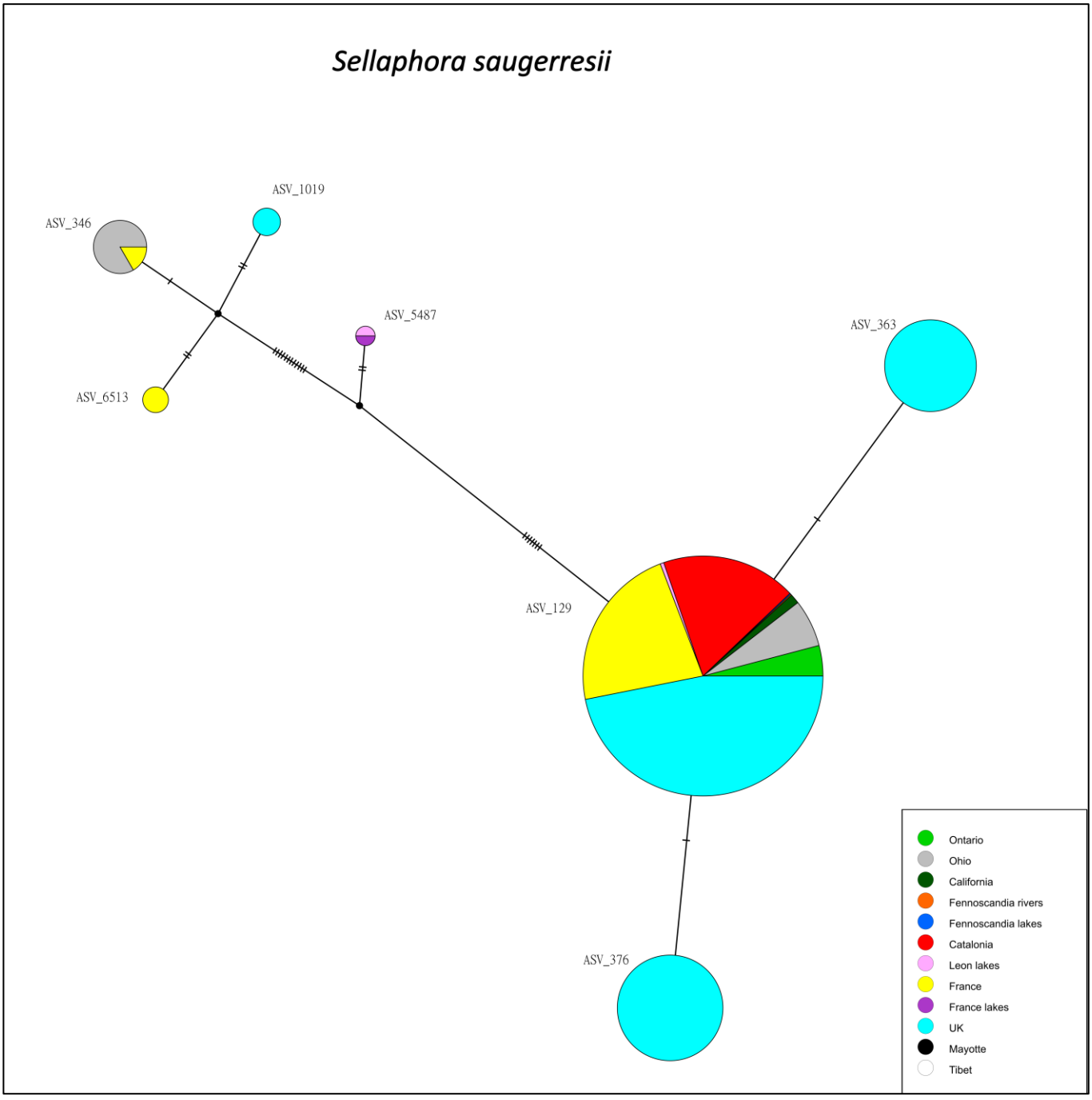
Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



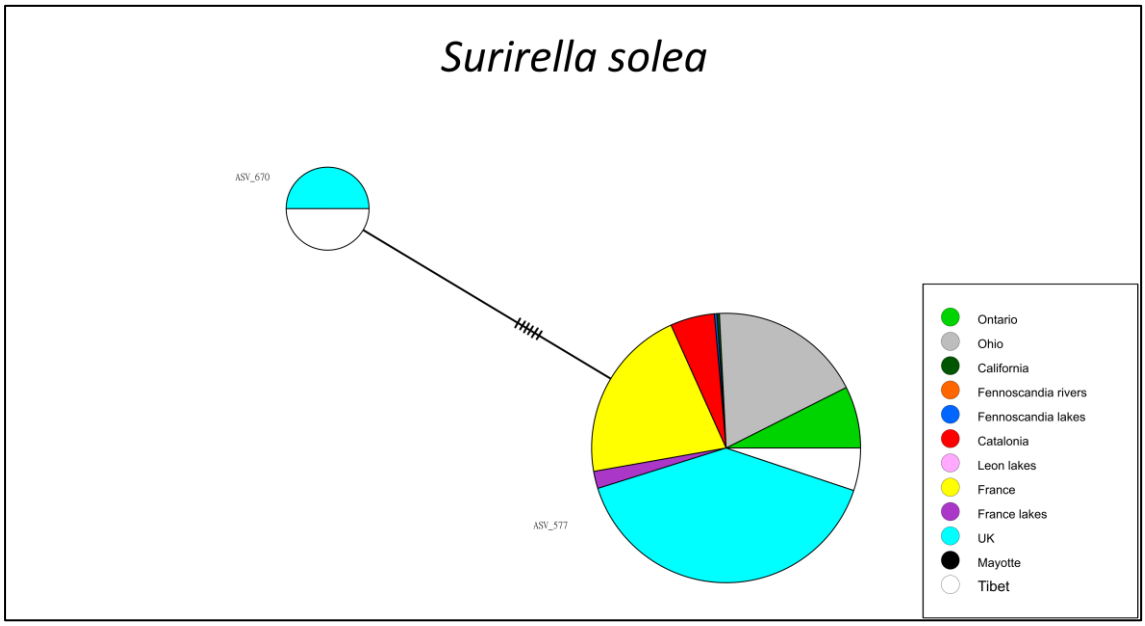
Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



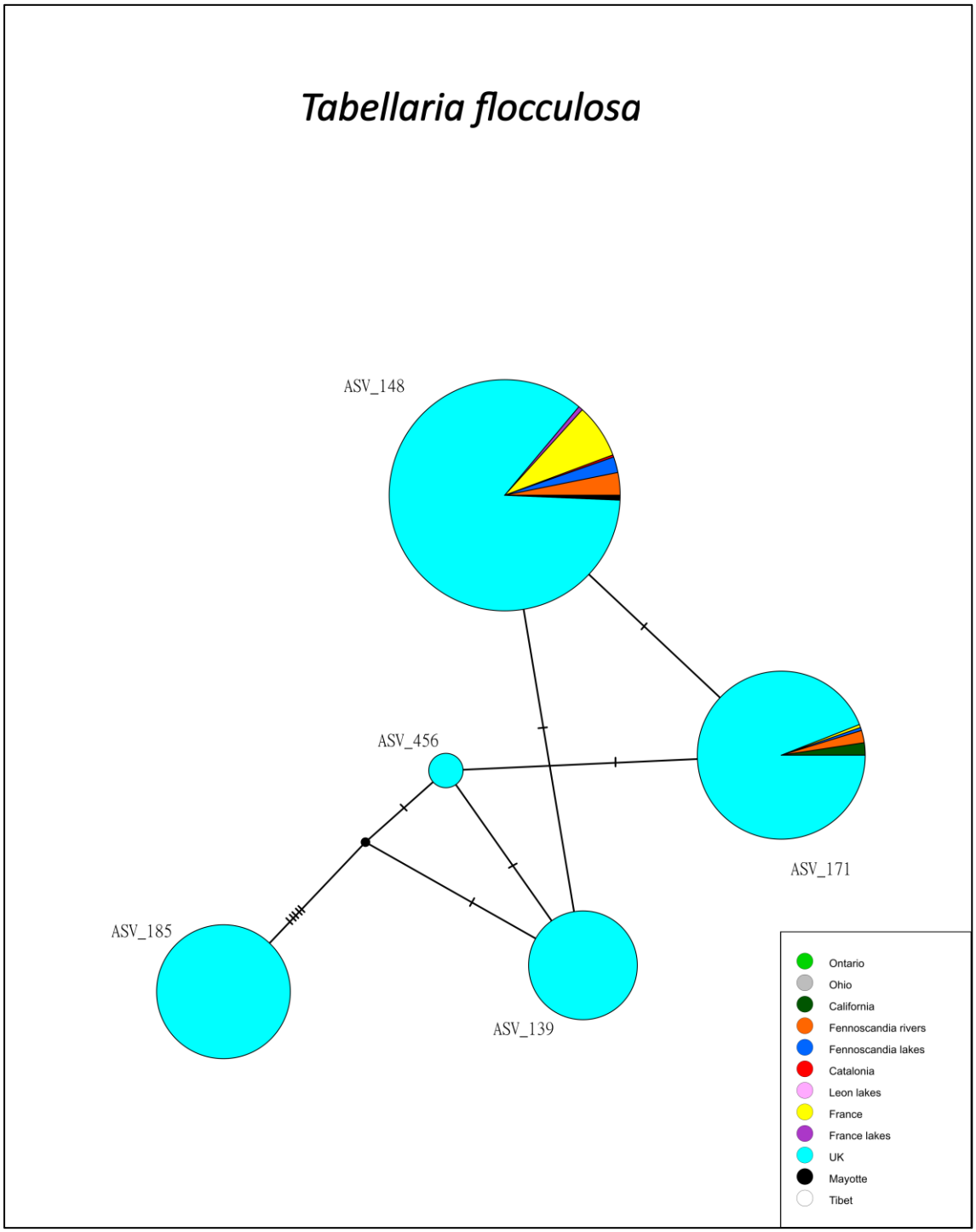
Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



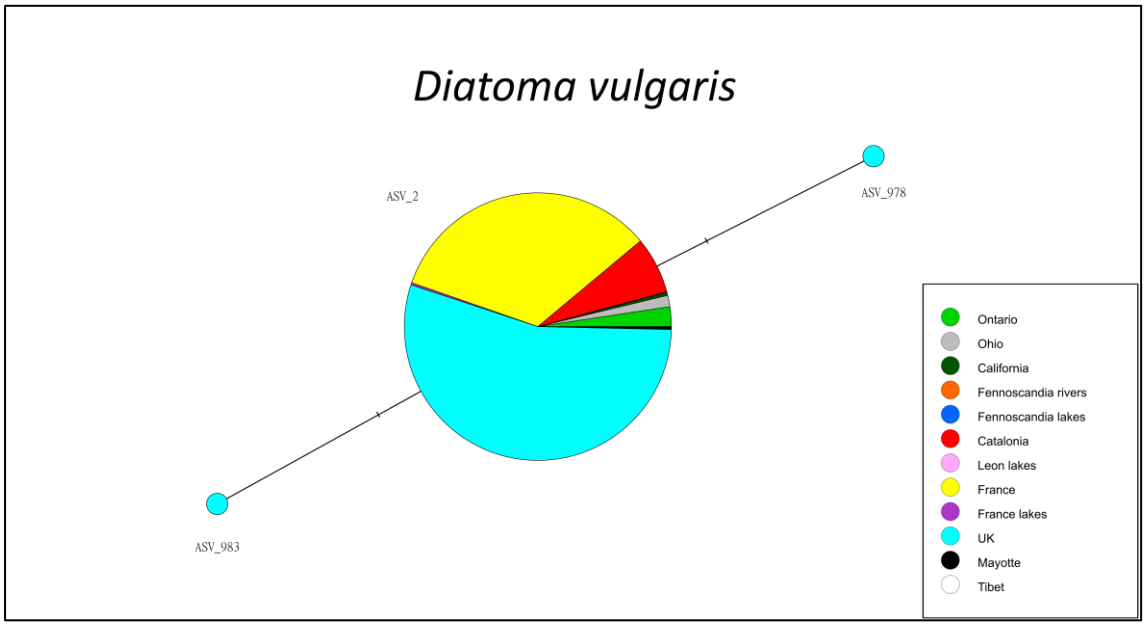
Rapid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



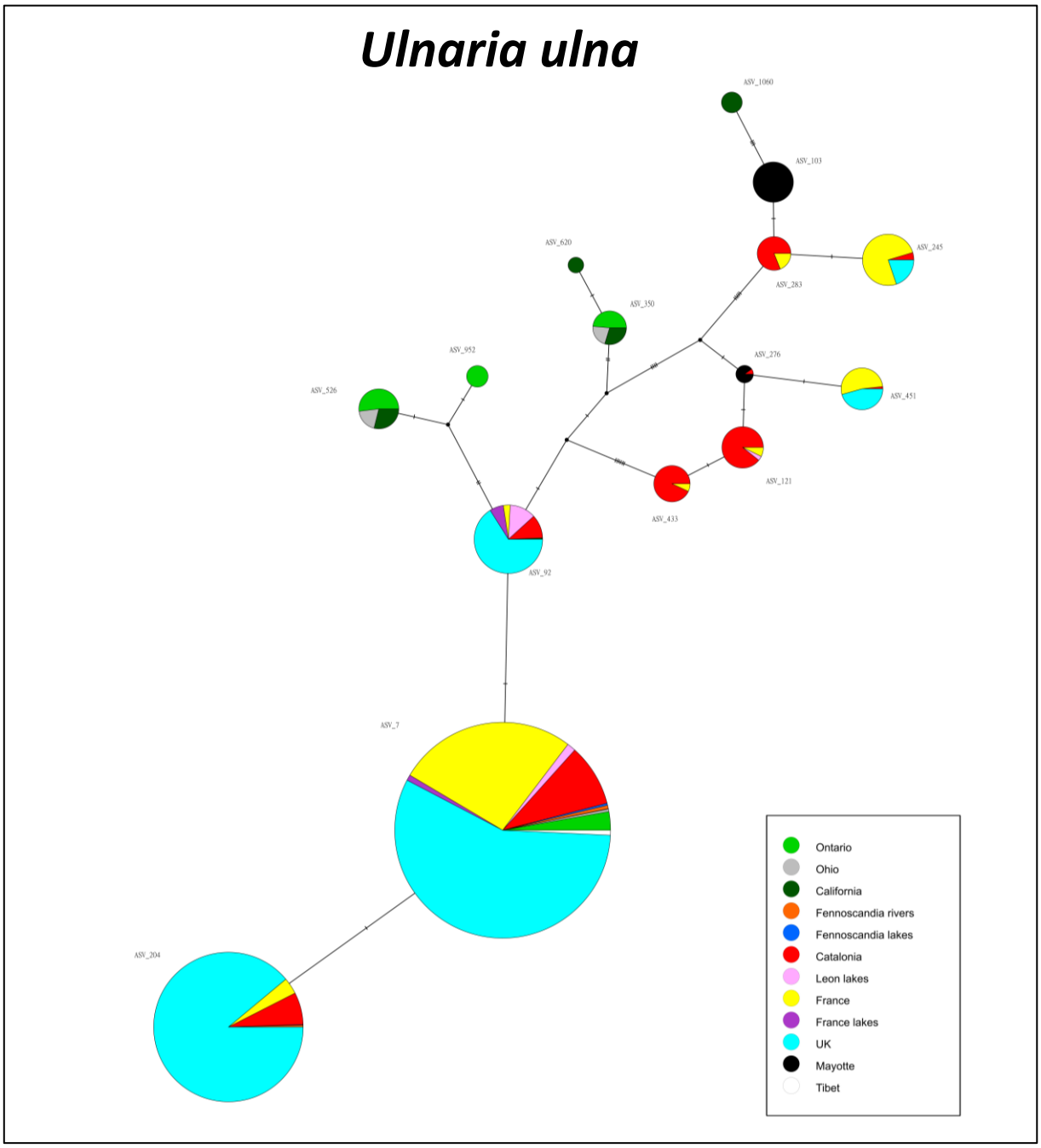
Araphid pennate species Pattern II

- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



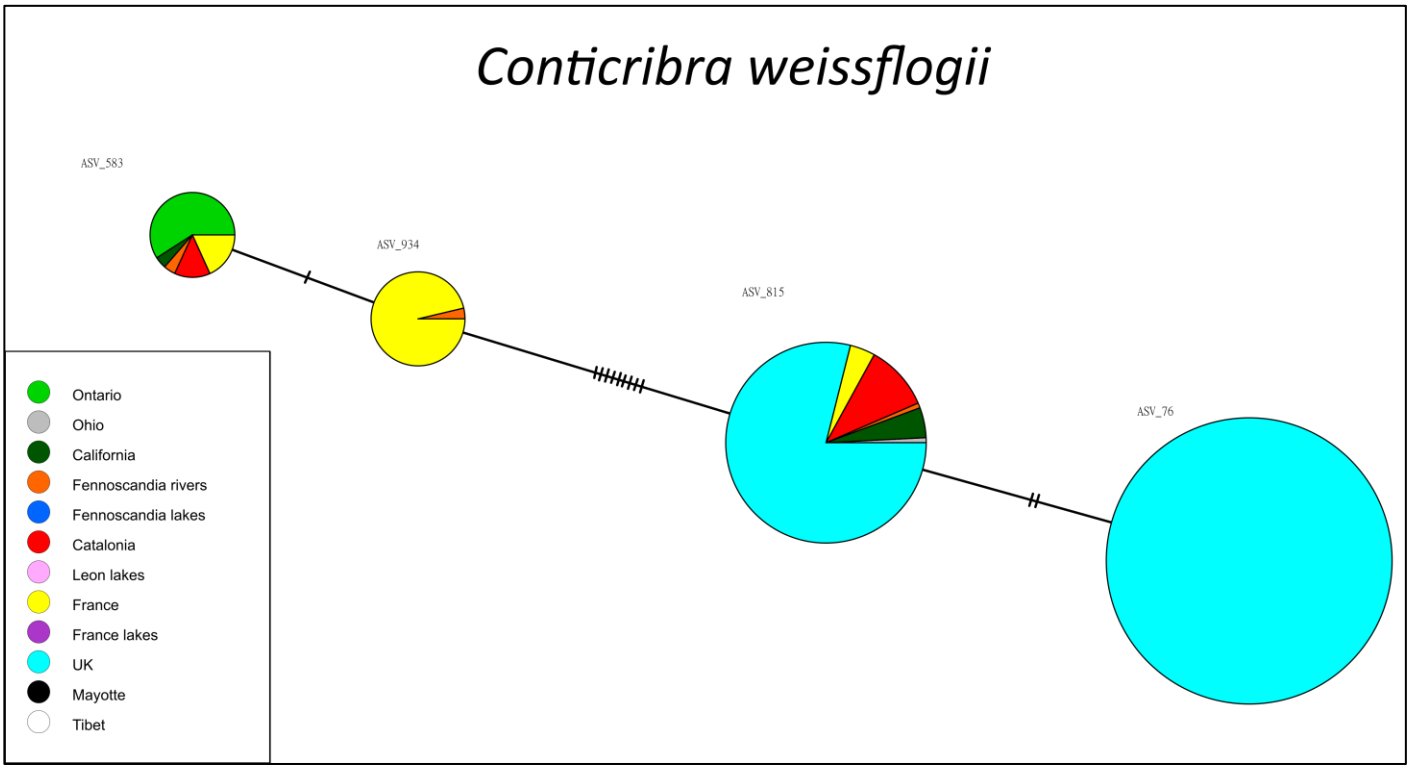
Araphid pennate species — Pattern II

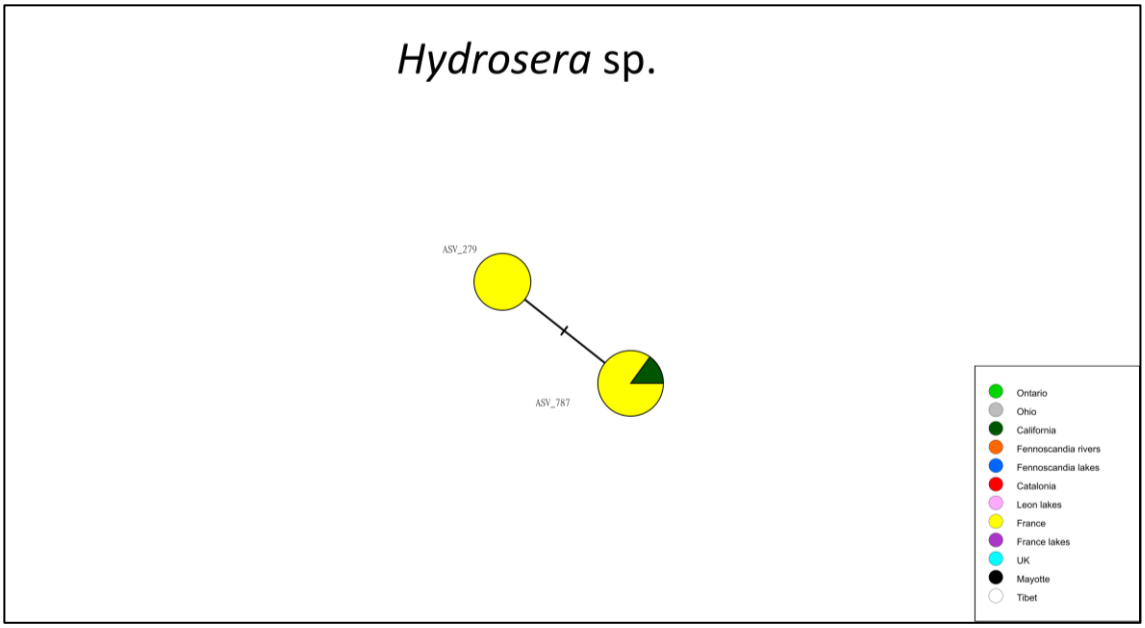
- ≥2 ASV per species
- Dominance of 1 or 2 ASVs
- Presence of rare ASVs



Centric species Pattern III

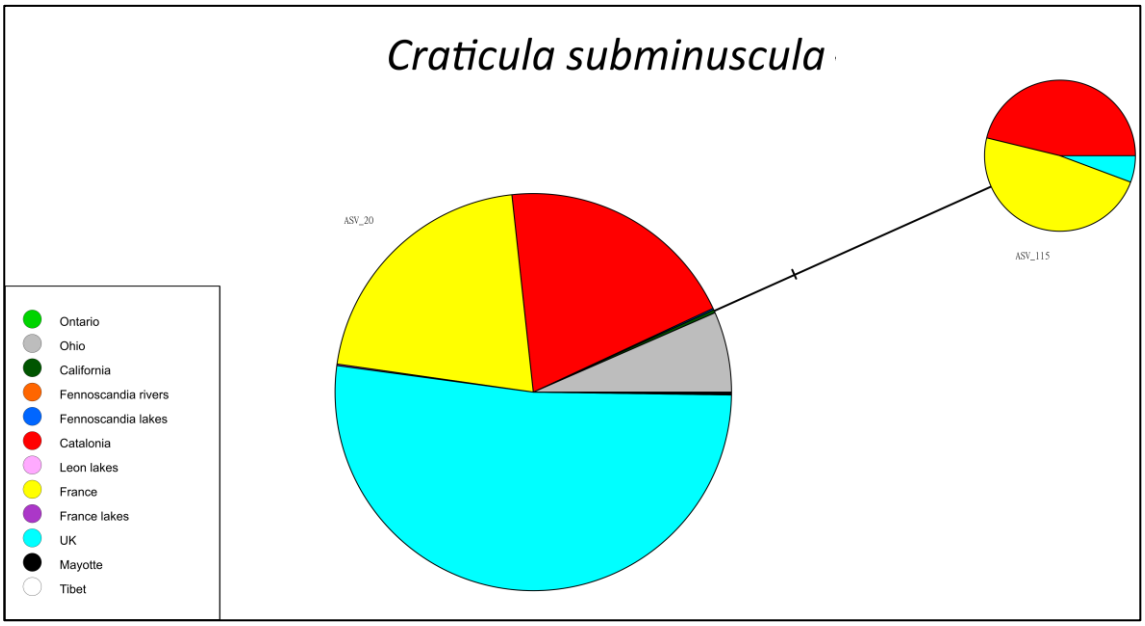
- 2-3 ASV per species
- No dominance of 1 or 2 ASVs
- No presence of rare ASVs

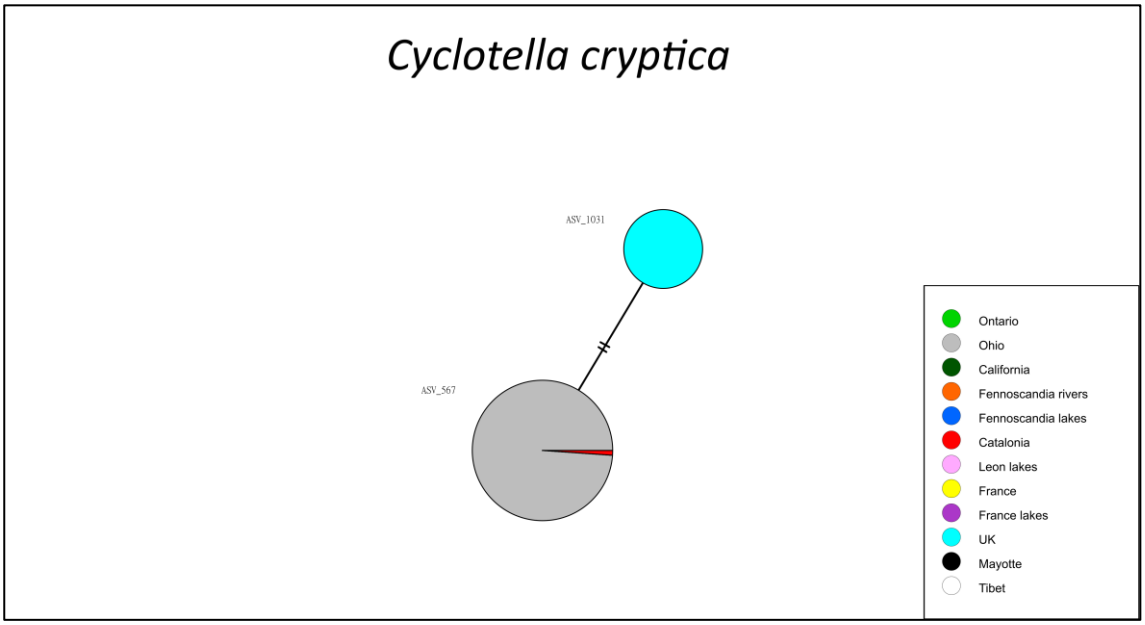




Rapid pennate species Pattern III

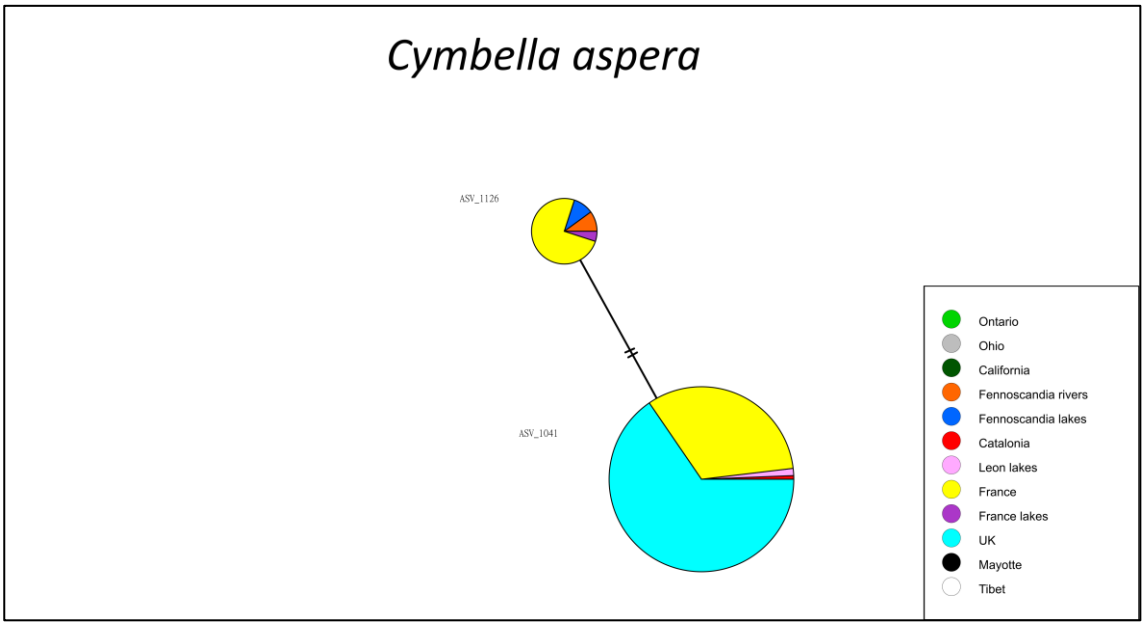
- 2-3 ASV per species
- No dominance of 1 or 2 ASVs
- No presence of rare ASVs

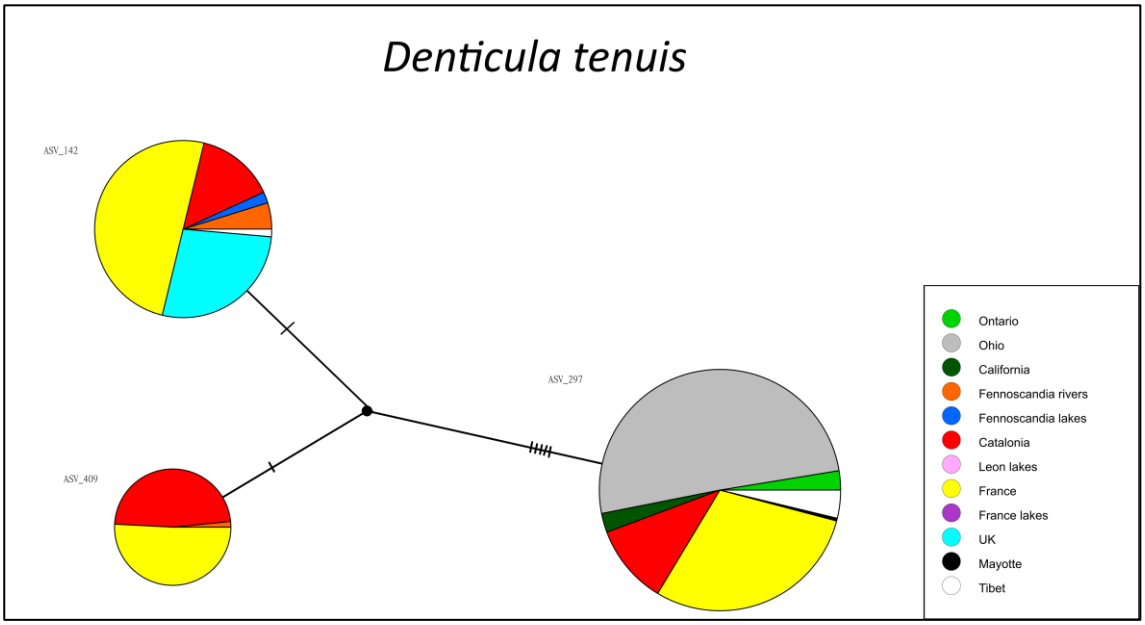




Rapid pennate species Pattern III

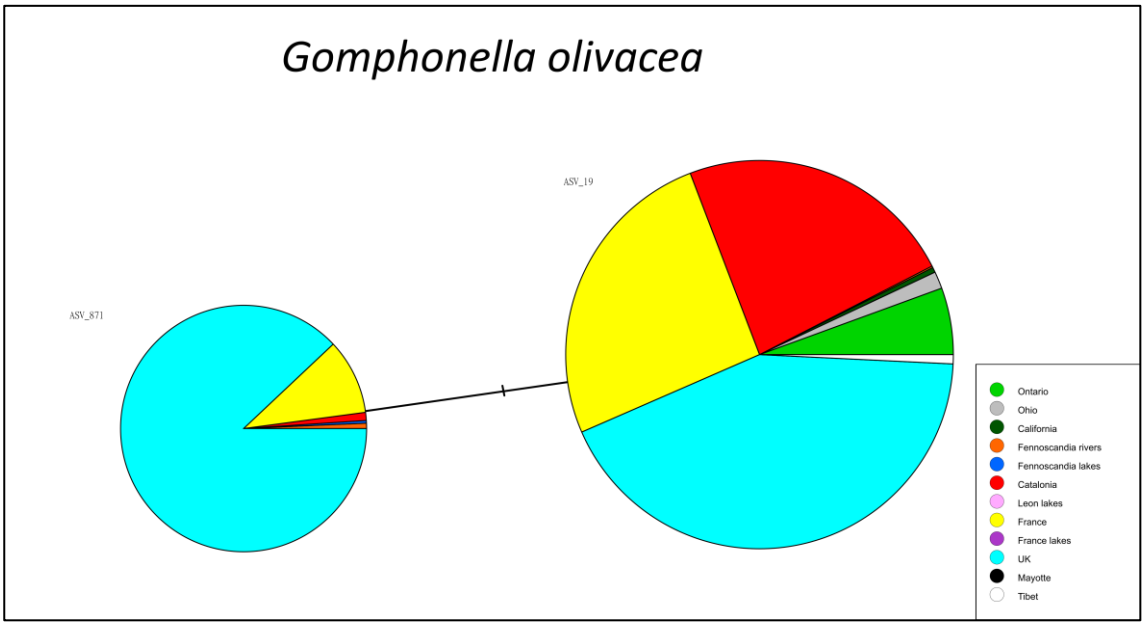
- 2-3 ASV per species
- No dominance of 1 or 2 ASVs
- No presence of rare ASVs

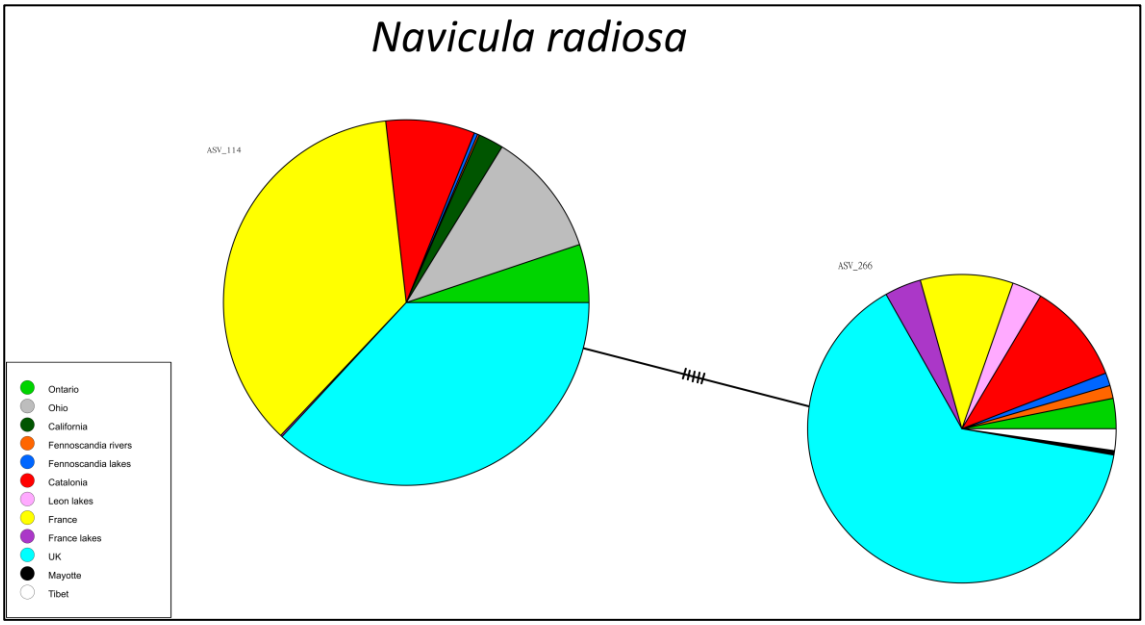




Rapid pennate species Pattern III

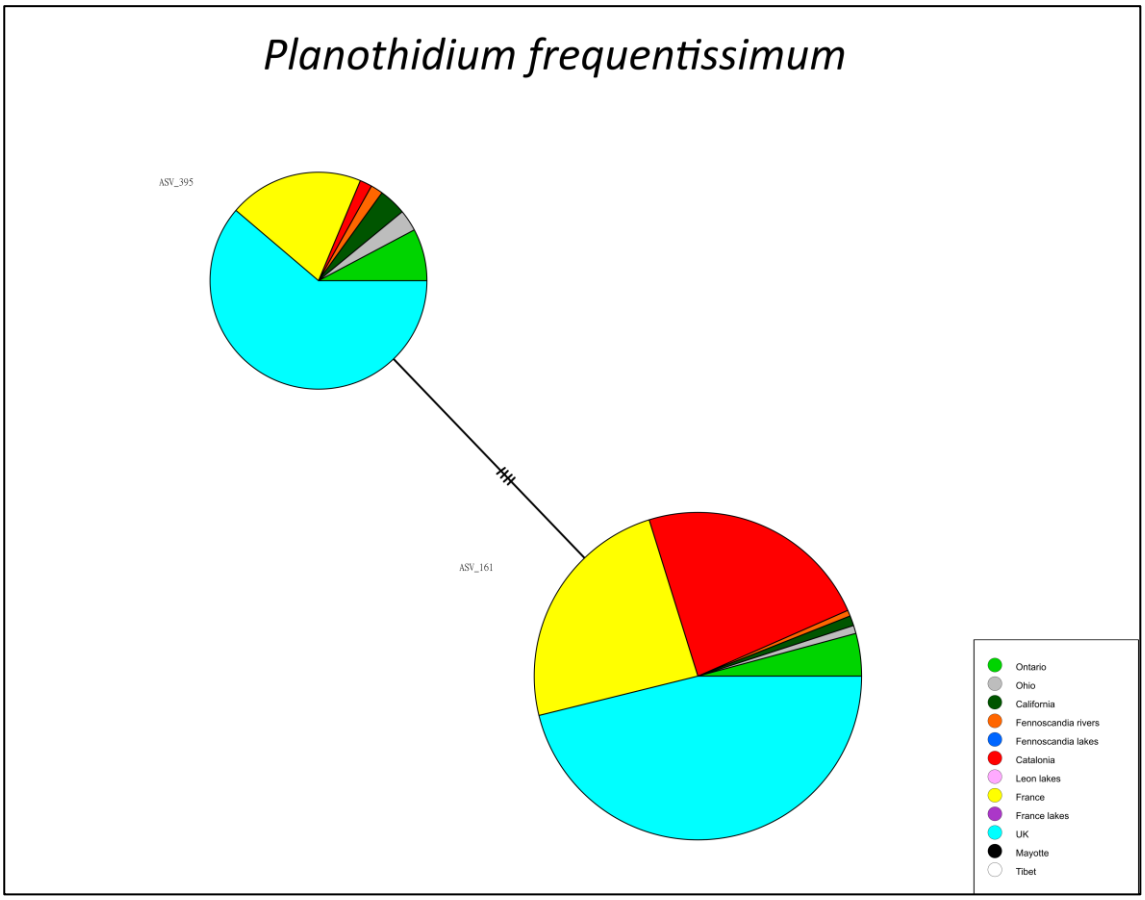
- 2-3 ASV per species
- No dominance of 1 or 2 ASVs
- No presence of rare ASVs

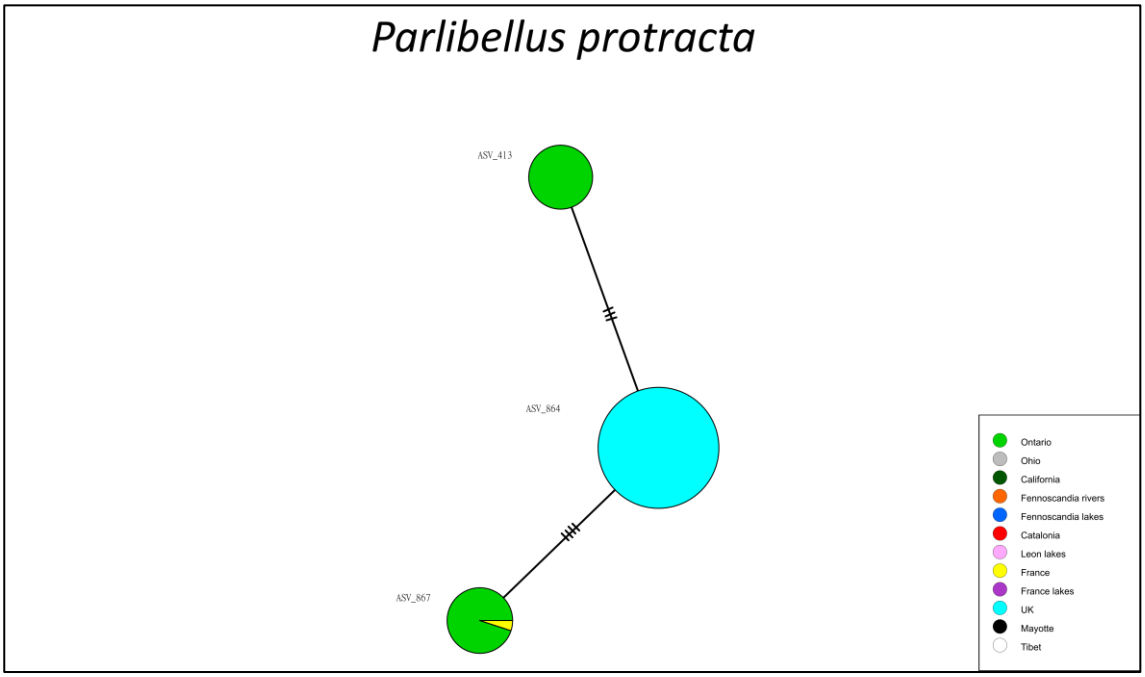




Rapid pennate species Pattern III

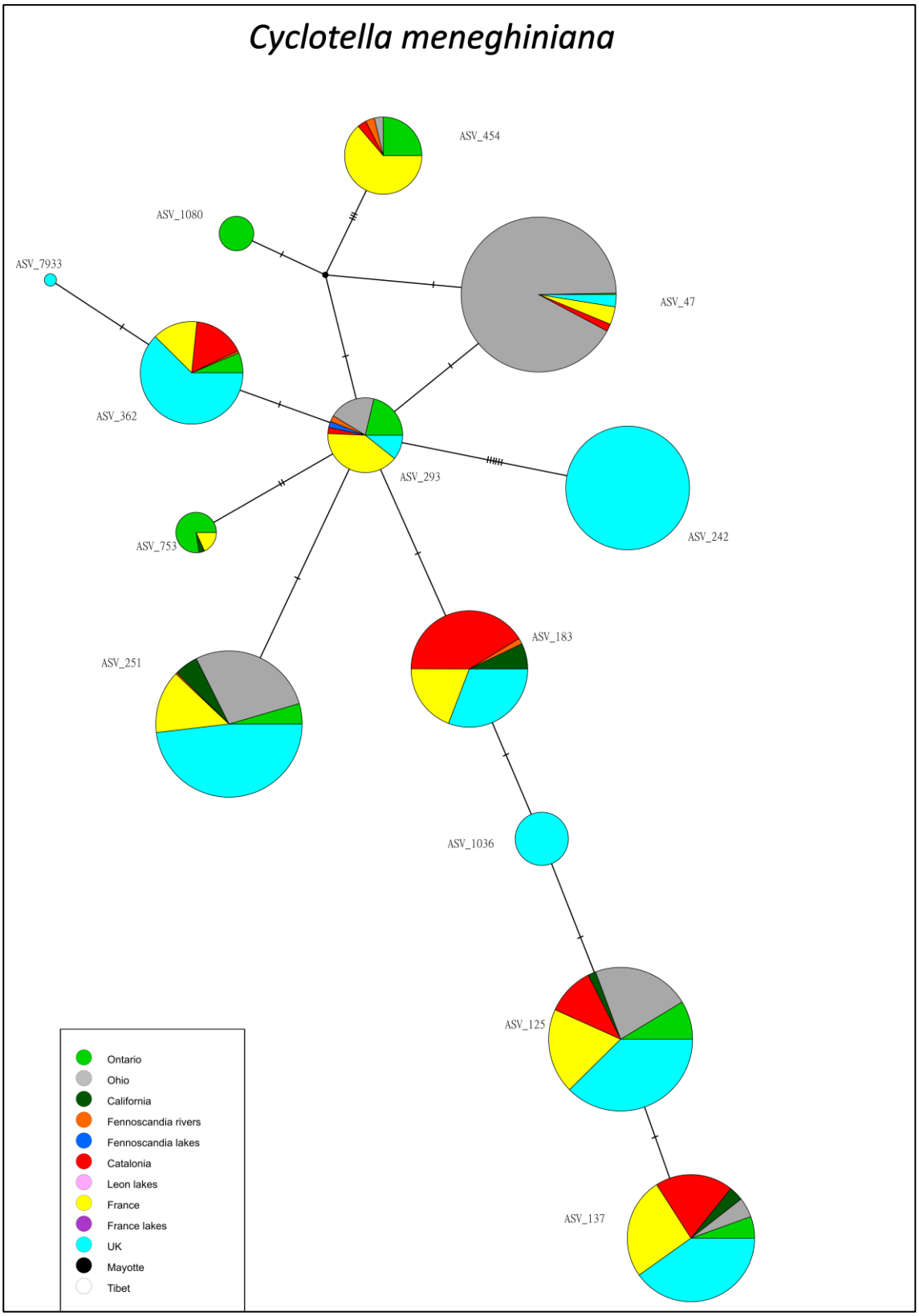
- 2-3 ASV per species
- No dominance of 1 of 2 ASVs
- No presence of rare ASVs





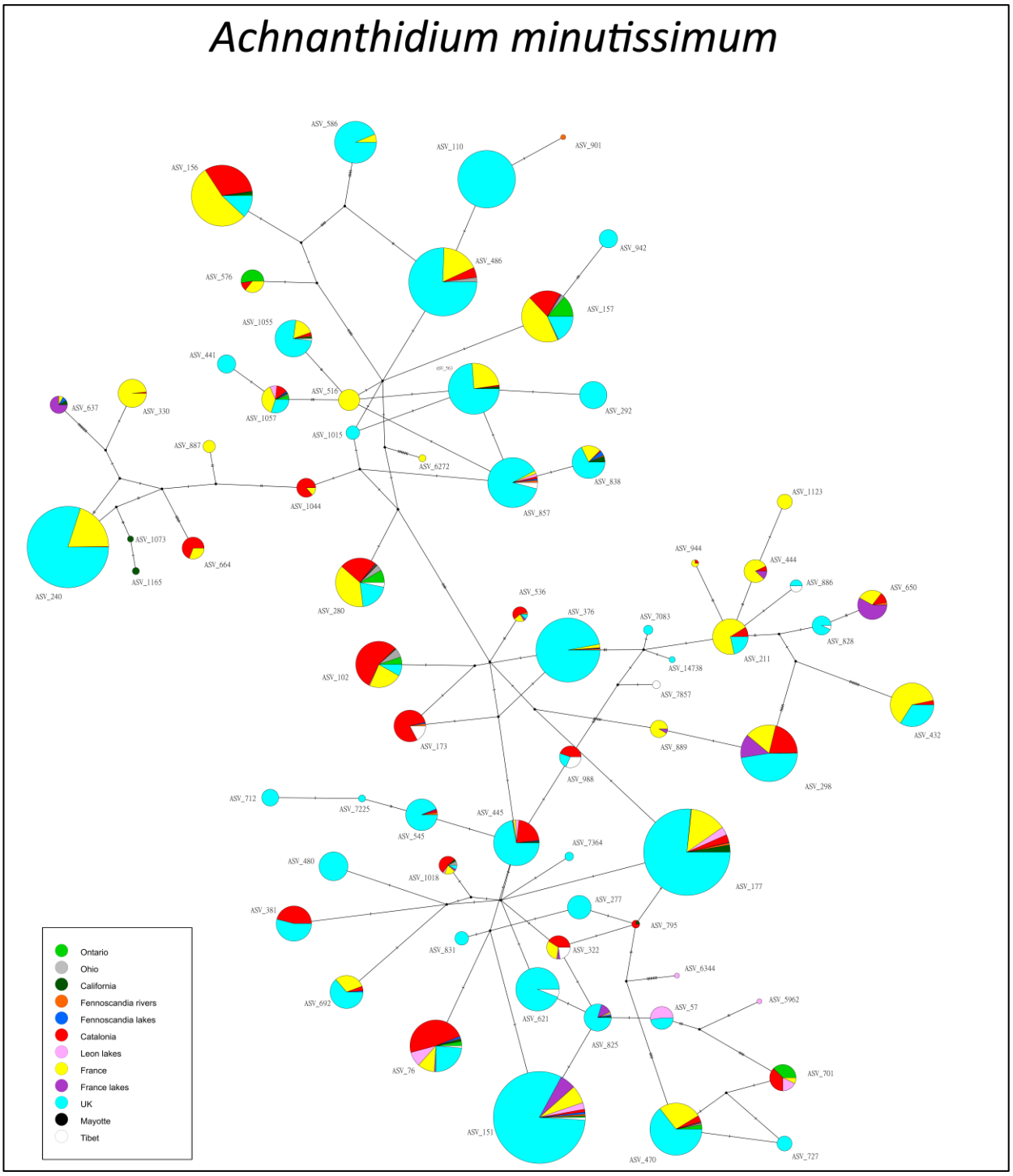
Centric species — Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



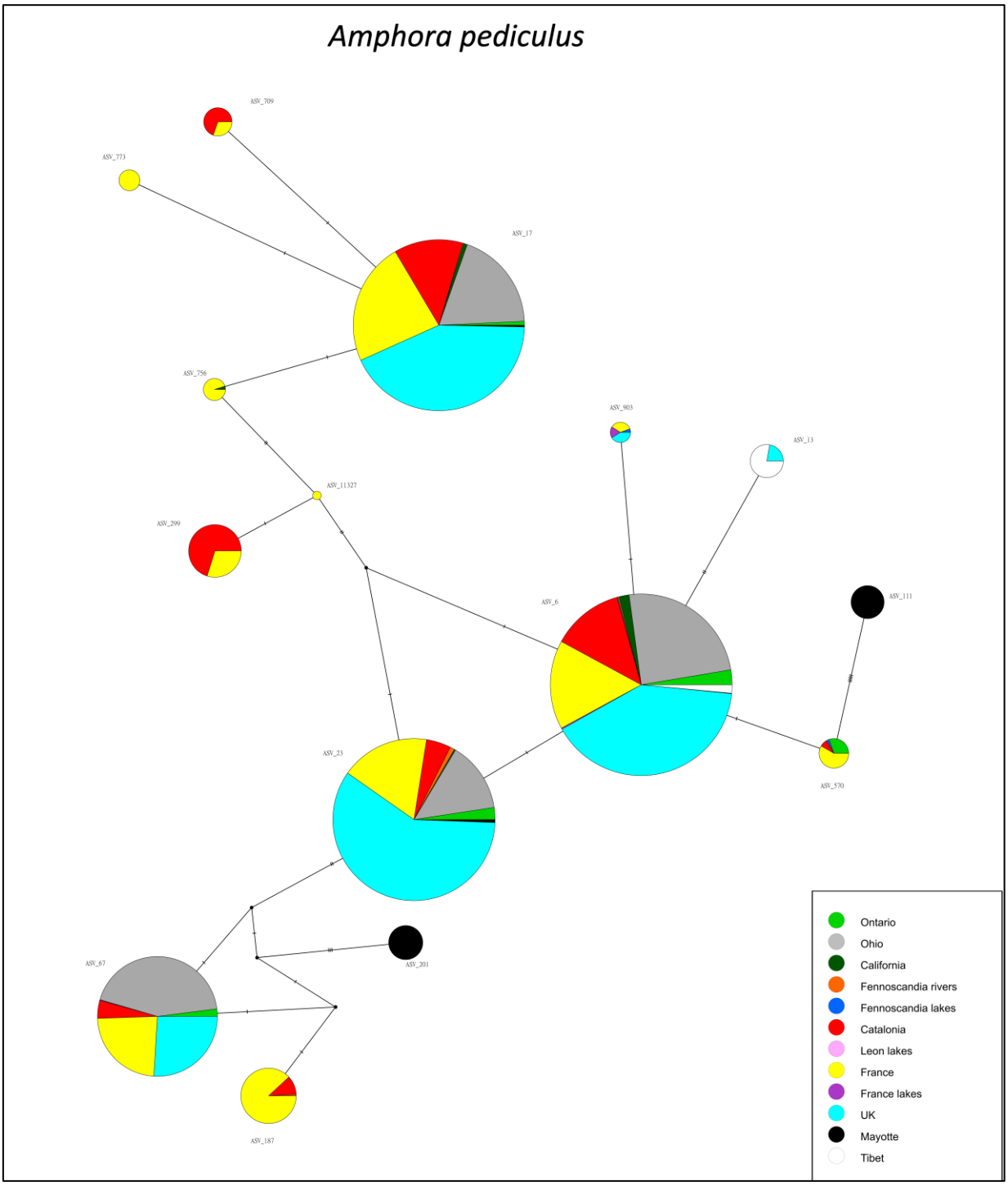
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



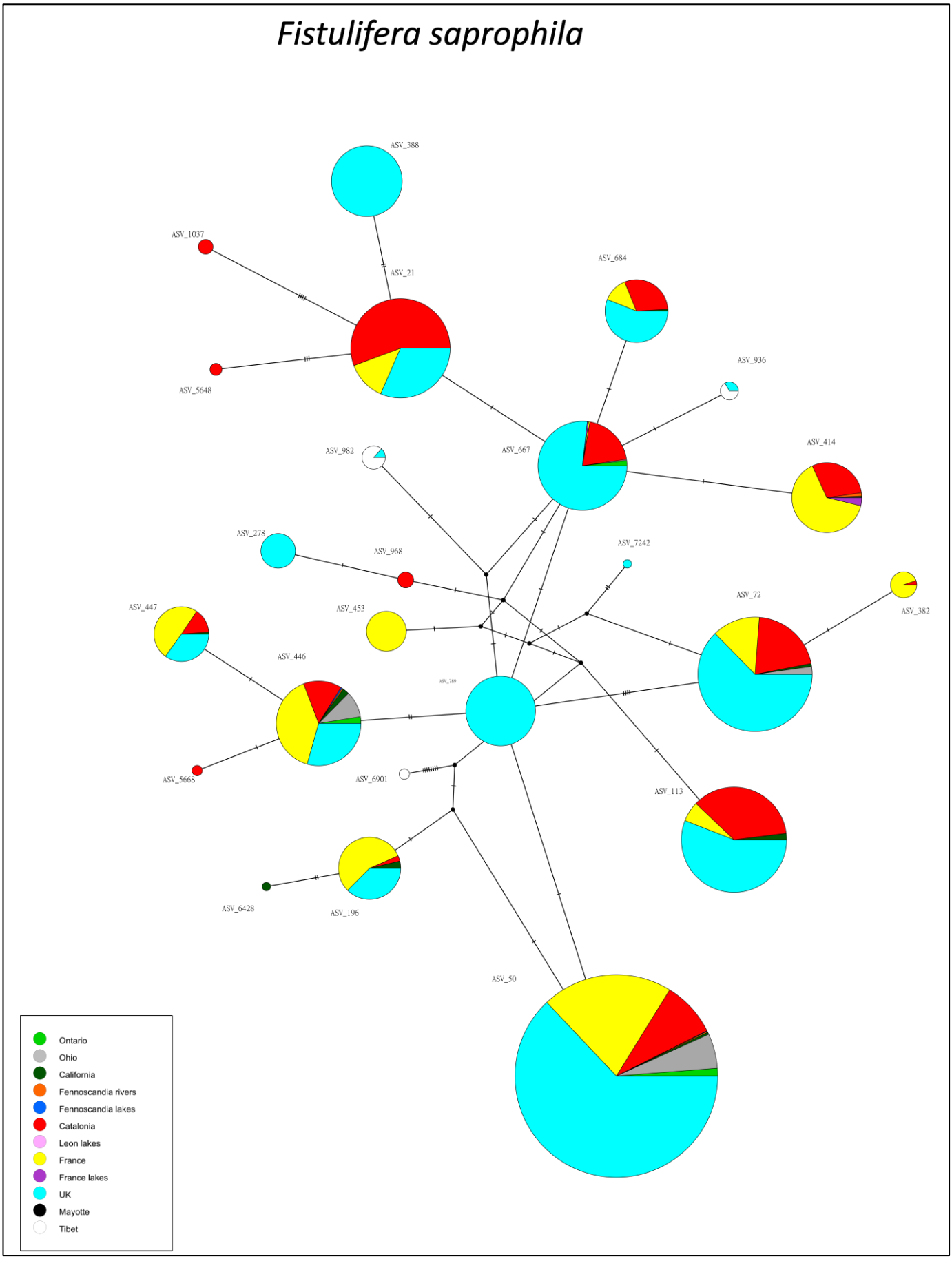
Rapid pennate species — Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



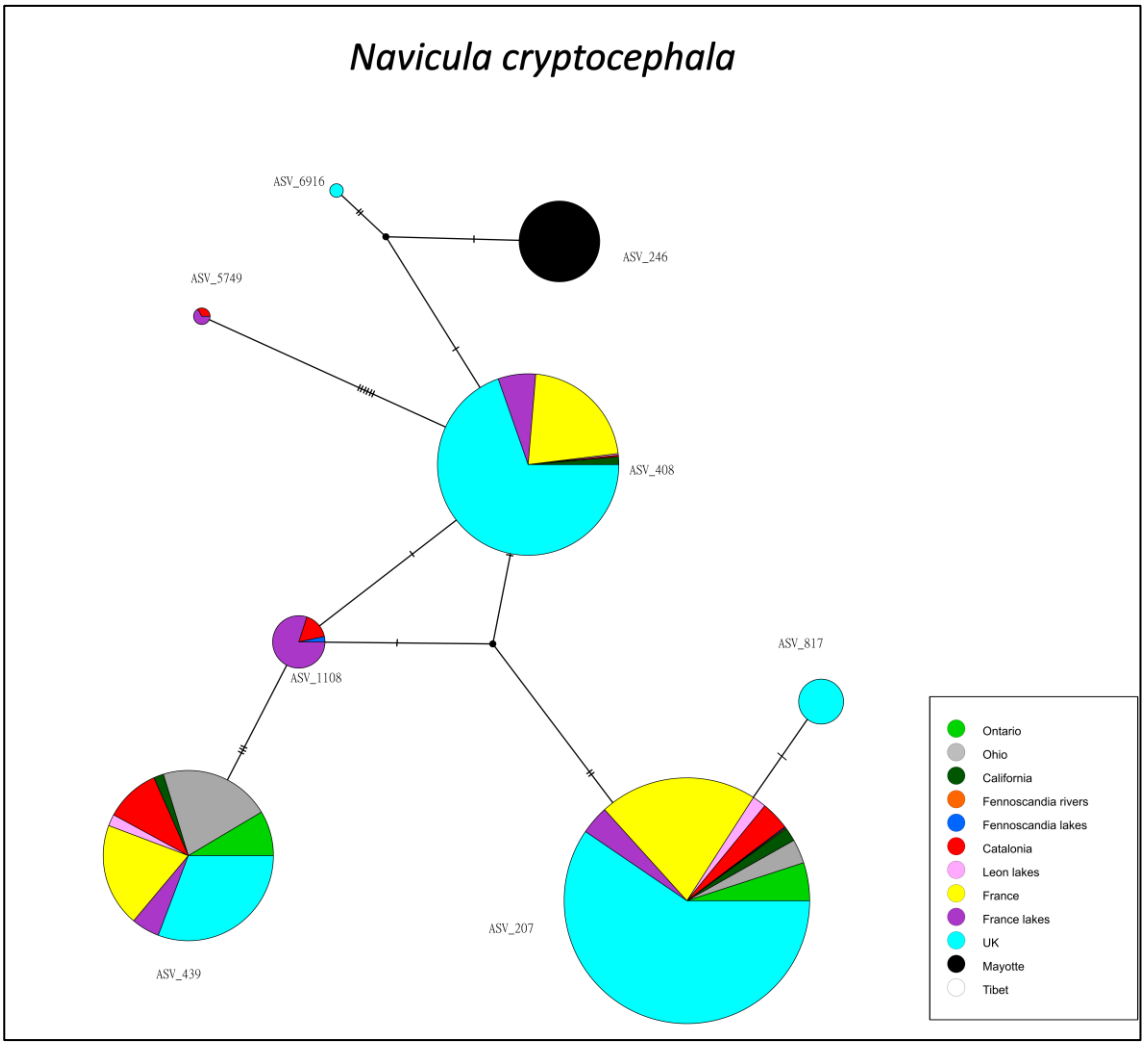
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



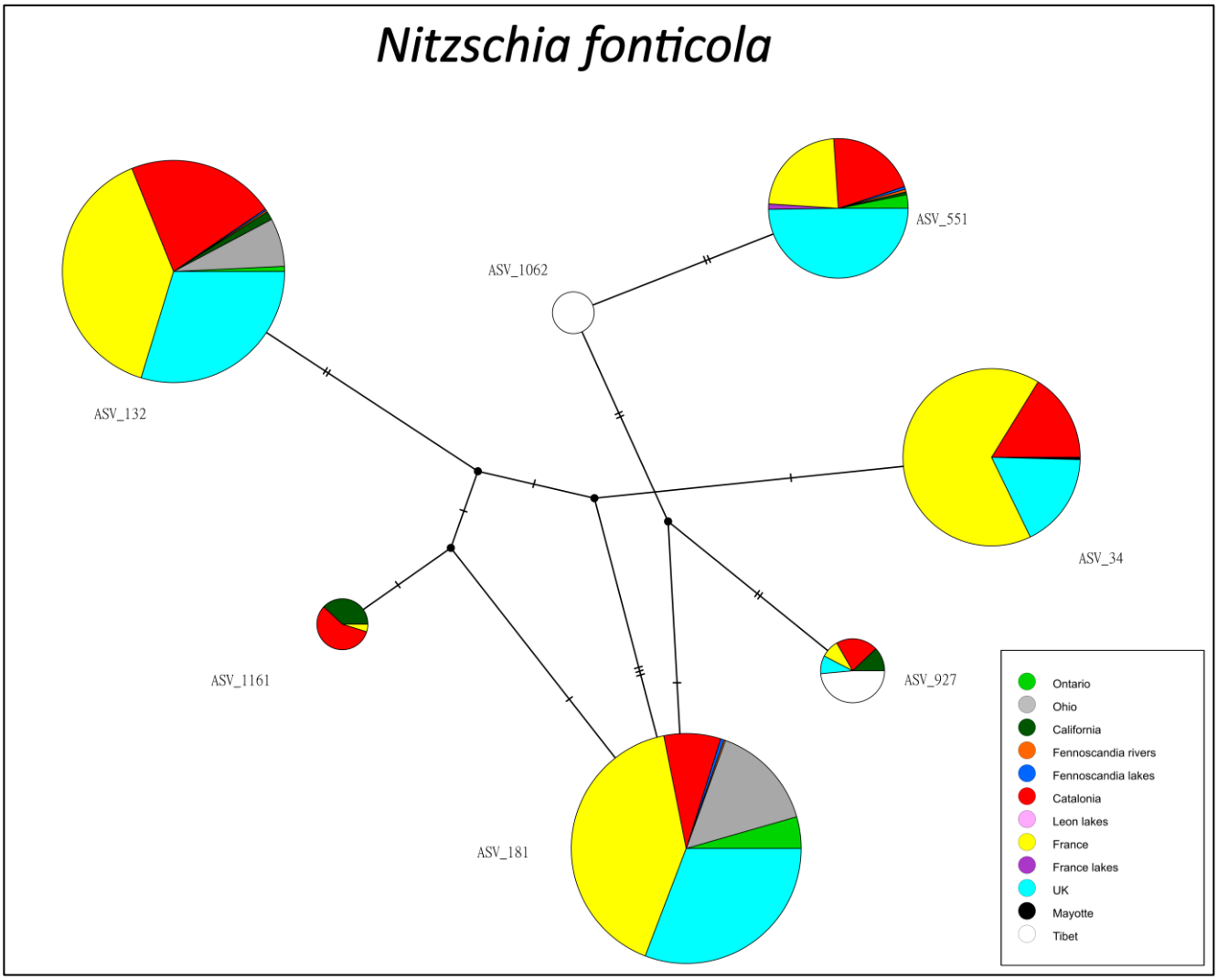
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



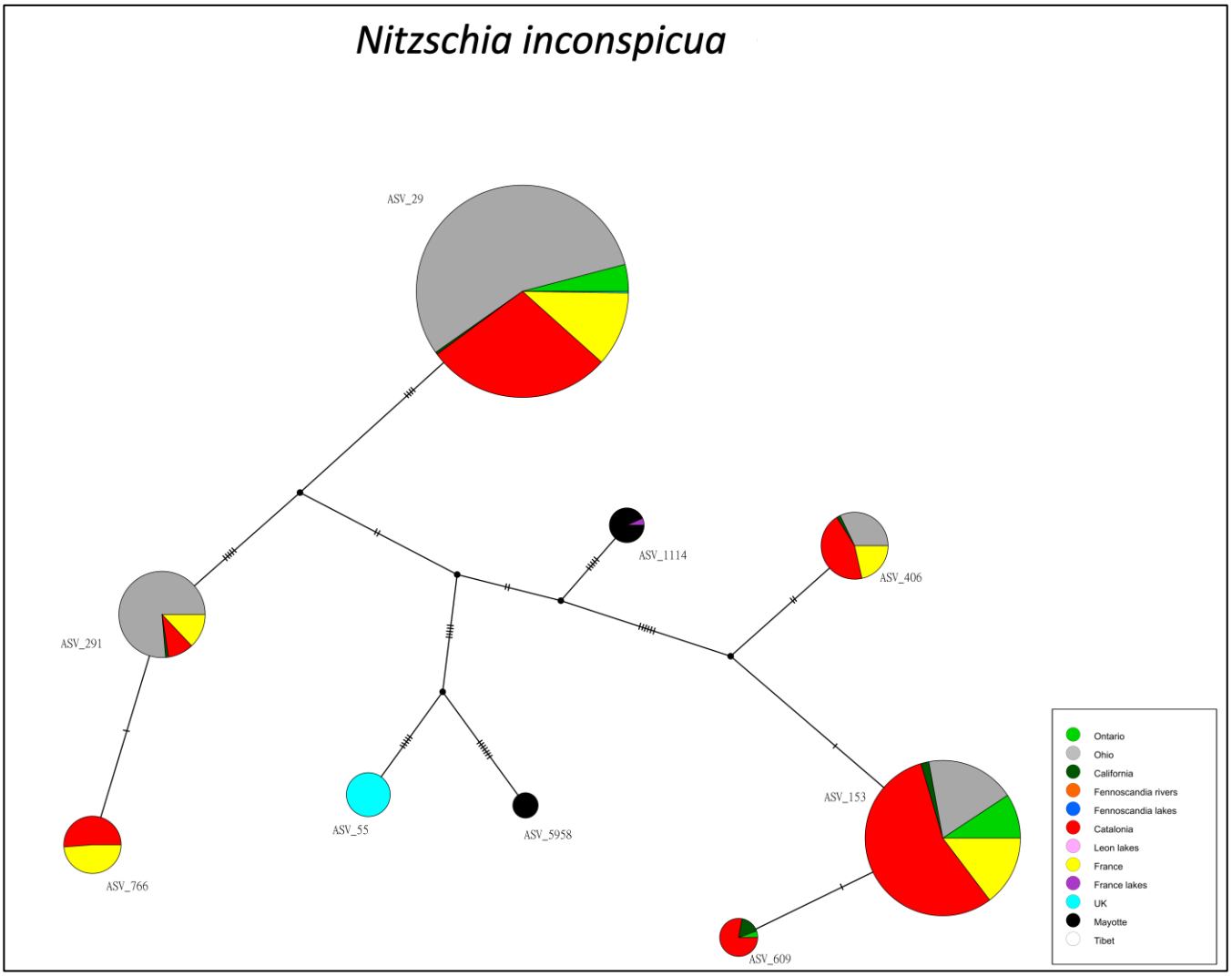
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



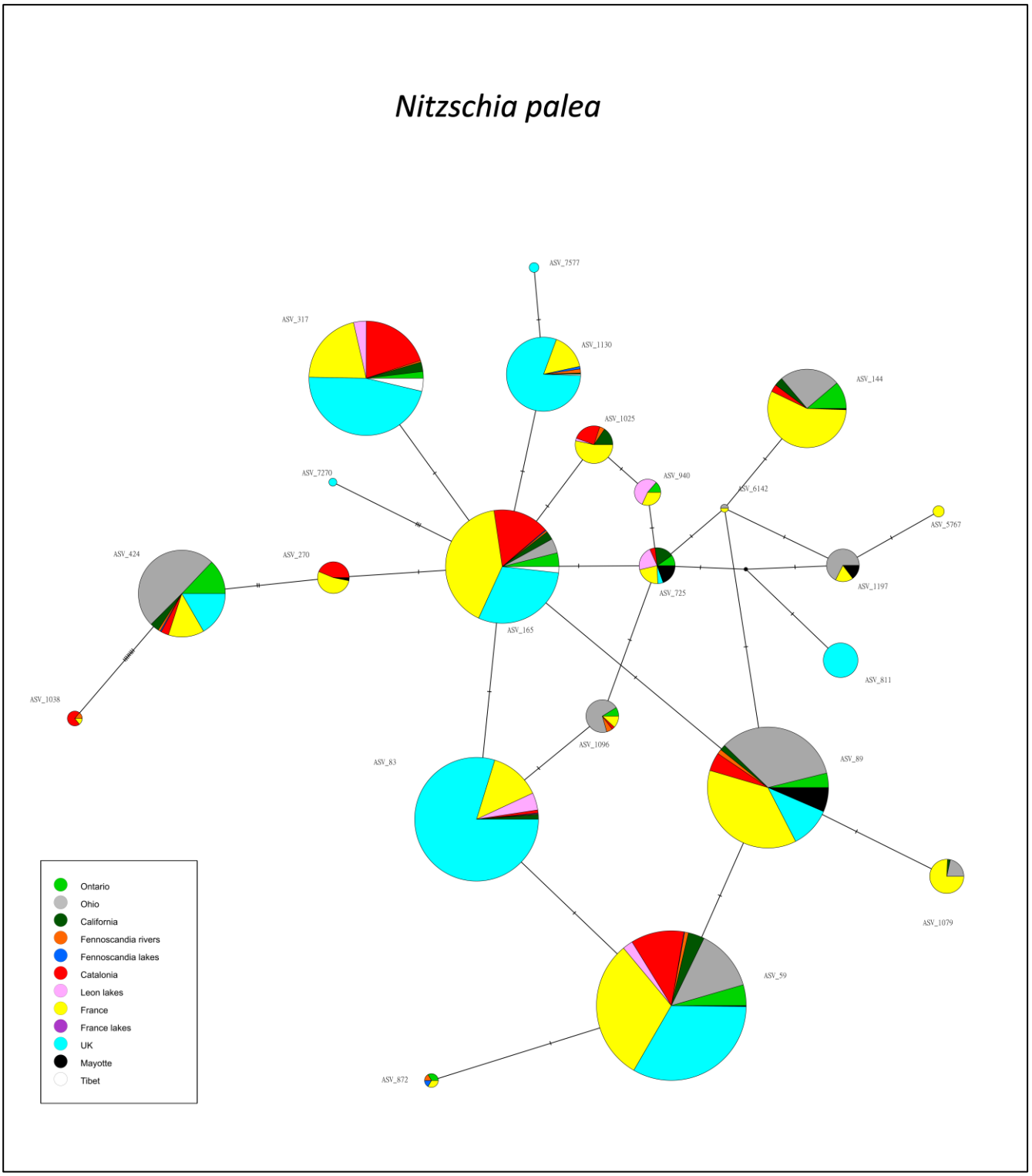
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



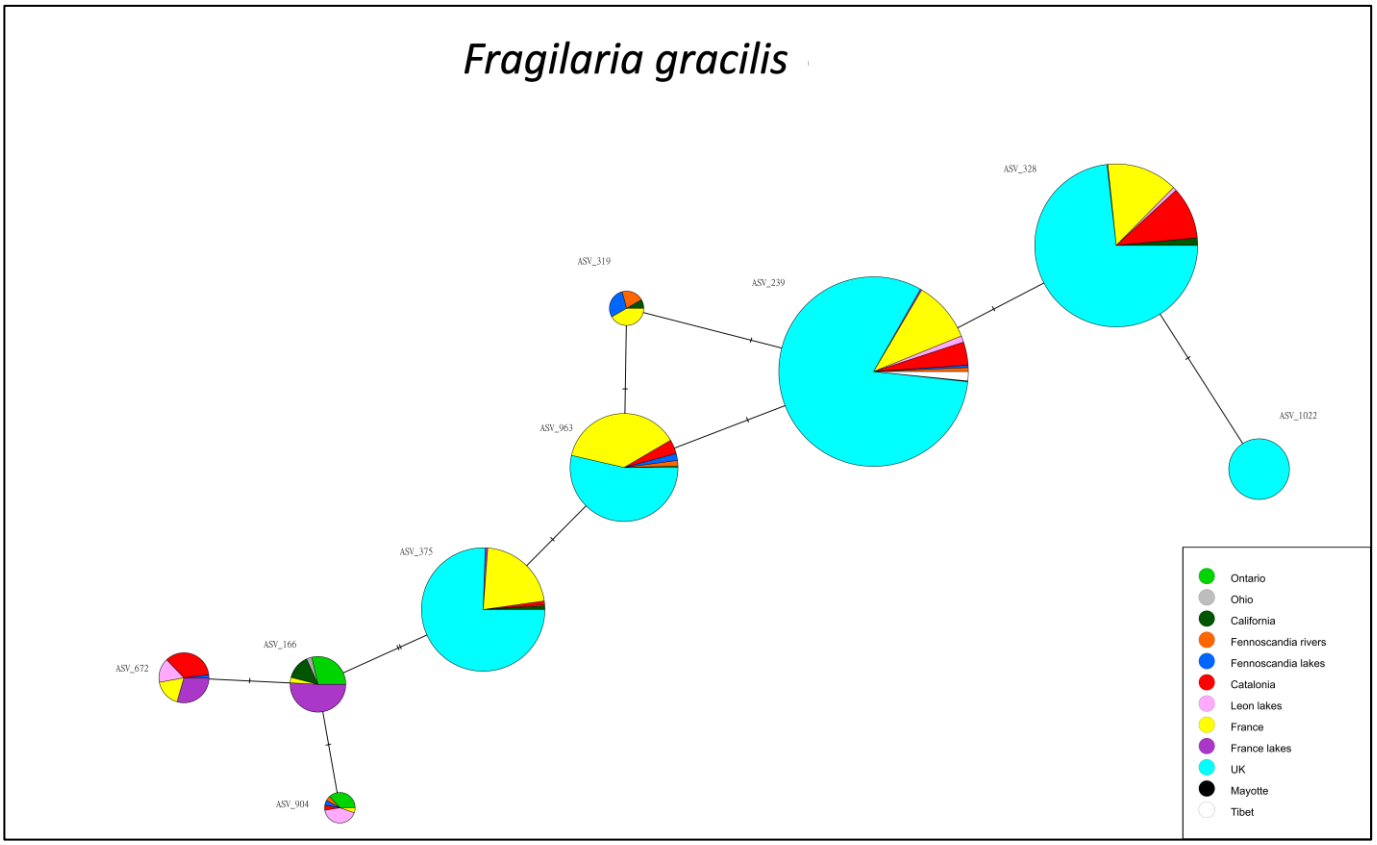
Rapid pennate species Pattern IV

- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs

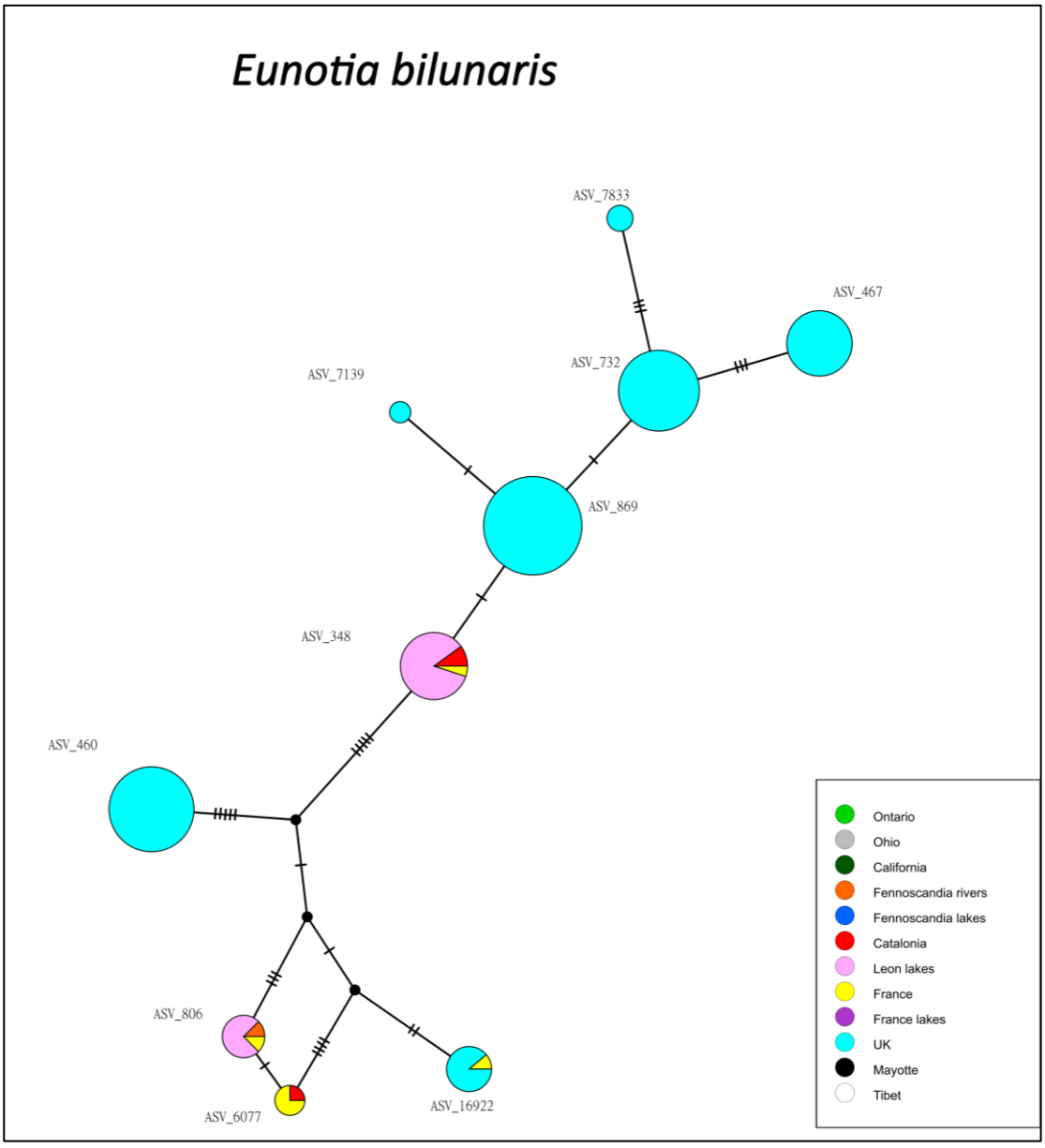


Araphid pennate species Pattern IV

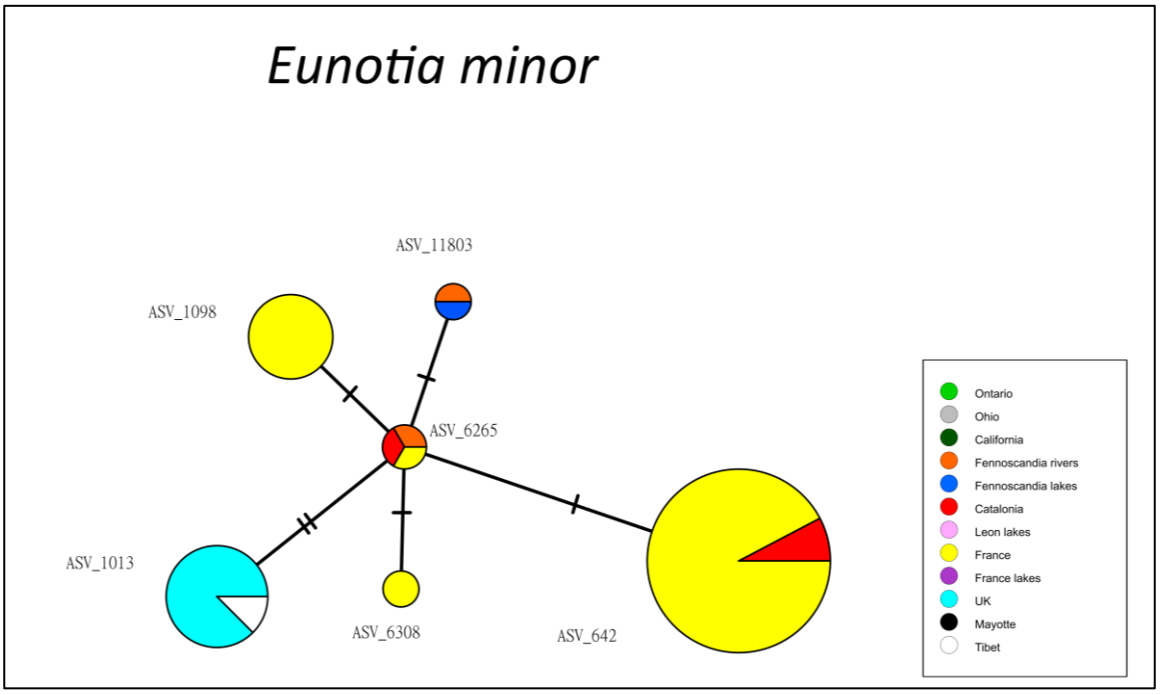
- ≥7 ASVs per species
- No dominance of 1 or 2 ASVs
- Presence of rare ASVs



Extra pattern – No fit with any of the previous patterns



Extra pattern – No fit with any of the previous patterns



Annex 2

Annex 2

Chemosphere 301 (2022) 134777

Contents lists available at ScienceDirect

Chemosphere

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

Assessment of a novel microalgae-cork based technology for removing antibiotics, pesticides and nitrates from groundwater

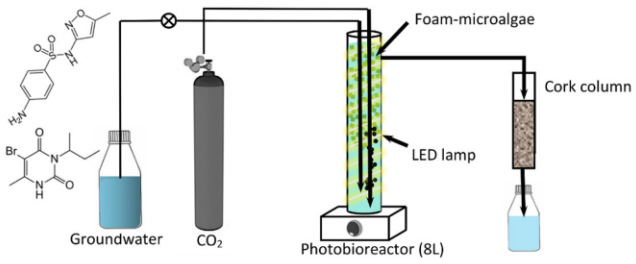
Lorenzo Rambaldo [a], Héctor Ávila [a], Mònica Escolà Casas [a], Miriam Guivernau [b], Marc Viñas [b], Rosa Trobajo [c], Javier Pérez-Burillo [c,d], David G. Mann [c,e], Belén Fernández [b], Carme Biel [f], Luigi Rizzo [g], Josep M. Bayona [a], Víctor Matamoros [a,*]

[a] Department of Environmental Chemistry, IDAEA-CSIC, c/Jordi Girona, 18-26, E-08034, Barcelona, Spain
[b] IRTA-Institute for Food and Agricultural Research and Technology, Sustainability in Biosystems Programme, Torre Marimon, E-08140, Caldes de Montbui, Barcelona, Spain
[c] 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, Catalonia, Spain
[d] Departament of Geography, Universitat Rovira i Virgili, C/Joanot Martorell 15, E43500, Vila-seca, Catalonia, Spain
[e] Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, Scotland, UK
[f] IRTA-Institute for Food and Agricultural Research and Technology, Sustainable Plant Protection Programme, Ctra. de Cabrils, Km 2, E08348, Cabrils, Catalonia, Spain
[g] Department of Civil Engineering, University of Salerno, Via Giovanni Paolo II 132, 84084, Fisciano, SA, Italy

HIGHLIGHTS

- Micropollutants were eliminated in large quantities (95%), but not nitrates (20–58%).
- The prototype's efficiency was highly reliant on the HRT.
- After the cork filter, pesticide transformation products were found.
- Microbial species with a high biodegradation of micropollutants were identified.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Tsair-Fuh

Keywords:
Antibiotics
Microalgae
Nitrates
Pesticides
Photo-biodegradation
Transformation products

ABSTRACT

Groundwater pollution has increased in recent years due to the intensification of agricultural and livestock activities. This results in a significant reduction in available freshwater resources. Here, we have studied the long term assessment of a green technology (1–4 L/day) based on a photobioreactor (PBR) containing immobilised microalgae–bacteria in polyurethane foam (PF) followed by a cork filter (CF) for removing nitrates, pesticides (atrazine and bromacil), and antibiotics (sulfamethoxazole and sulfacetamide) from groundwater. The prototype was moderately effective for removing nitrates (58%) at an HRT of 8 days, while its efficiency decreased at a HRT of 4 and 2 days (<20% removal). The combined use of PBR-CF enabled antibiotics and pesticides to be attenuated by up to 95% at an HRT of 8 days, but their attenuation decreased with shorter HRT, with pesticides being the compounds most affected (reducing from 97 to 98% at an HRT of 8 days to 23–45% at an HRT of 2 days). Pesticide transformation products were identified after the CF, supporting biodegradation as the main

* Corresponding author.
E-mail address: victor.matamoros@cid.csic.es (V. Matamoros).

https://doi.org/10.1016/j.chemosphere.2022.134777
Received 22 December 2021; Received in revised form 1 March 2022; Accepted 26 April 2022
Available online 29 April 2022

attenuation process. A gene-based metataxonomic assessment linked the attenuation of micropollutants to the presence of specific pesticide biodegradation species (e.g. genus *Phenylobacterium, Sphingomonadaceae, and Caulobacteraceae*). Therefore, the results highlighted the potential use of microalgae and cork to treat polluted groundwater.

1. Introduction

Groundwater is the largest freshwater reserve in the world. It represents the most important source of drinking water (it supplies about half of the world's population) and contributes significantly to irrigation, and therefore to food security, in arid and semi-arid regions (Jia et al., 2019). Nevertheless, there is currently a lot of concern about groundwater pollution from nitrates (NO_3^-) and organic microcontaminants (OMCs) like pesticides and veterinary pharmaceuticals (Nguyen et al., 2020). The abundance and frequency of detection of pesticides and antibiotics has increased in recent years due to the intensification of agriculture and organic soil amendment with manure or biosolids (Kurwadkar, 2017). In Europe, 13% of groundwater monitoring stations exceed the 50 mg/L nitrate limit (91/676/EEC) (EUROSTAT, 2018), reaching concentration levels as high as 500 m/L due to intensive agricultural practices (Otero et al., 2009). Across Europe, the highest nitrate exceedance rates have been recorded in Belgium (30%), Denmark (26%), Spain (22%), and Cyprus (19%). According to a pan-European study, one or more pesticides occur at concentrations higher than 0.1 μg/L in 7% of monitored European groundwater (EUROSTAT, 2018). Although several pesticides have been identified in groundwater, atrazine and one of its metabolites, desethylatrazine, are the most frequently detected above the drinking water directive in Europe. Nevertheless, some triazine pesticides have been found in the shallow groundwaters of the United States at a concentration as high as 40 μg/L (Kolpin et al., 1998).

Current drinking water treatment technologies to remove these compounds are based on separation processes, including advanced oxidation processes, reverse osmosis, ion exchange, electrochemical reduction, electrodialysis, and activated carbon adsorption (Archna et al., 2012). However, these treatments are expensive to build and operate, and they generate a concentrated brine, which poses additional treatment and disposal operations, further increasing process costs. In contrast, biological denitrification processes remove nitrate by converting it to a harmless nitrogen gas. In recent years, biological treatments for nitrate removal have been developed based on heterotrophic microorganisms, but they require the supply of labile organic carbon as an electron donor in order to grow rapidly and take up nitrate as an electron acceptor (Rezvani et al., 2019).

Nature-based solutions based on microalgae or cork could solve this issue. Microalgae use sunlight to fix CO_2 from the atmosphere through photosynthesis (Taziki et al., 2015). Existing studies on the use of microalgae to remove nitrates from groundwater show that they can be a very effective solution, with removal efficiencies of up to 80% with an incubation time of 3 days (Rezvani et al., 2019). Recent studies from our laboratory have demonstrated that the co-immobilisation of microalgae and bacteria in polyurethane foam (PF) enhances the attenuation of nitrates, pesticides, and antibiotics from groundwater (Ferrando and Matamoros, 2020). The use of cork for the attenuation of nitrates and other pollutants from water has already been tested in various studies, but never for removing nitrates from groundwater. For example, Mallek et al. (2018) suggested that cork is a useful adsorbent for the removal of phenolic compounds and especially for the halogenated phenolic compounds PCP and 2,4-DCP, which only require 4.9 and 29 g/L of cork, respectively, to reduce their concentrations from 1 to 0.1 mg/L. In the same laboratory adsorption study, pharmaceuticals such as diclofenac and triclosan were 100% removed by using 5–10 mg of cork. However, the use of cork as a biofilter for removing nitrate and micropollutants from groundwater has never been investigated. As a consequence, combining photo-biodegradation processes in microalgae systems with biodegradation and sorption processes in cork biofilters might be extremely advantageous in terms of enhancing pollutant attenuation. Furthermore, another important gap in knowledge is the microbiological characterization of the microorganisms that drive pollutants' biodegradation. For example, recent studies have found that Sphingomonadaceae and Caulobacteraceae genera enhance the biodegradation of certain micropollutants and nutrients from water (Oh and Choi, 2019; Xu et al., 2018).

The goal of this study is therefore to explore for the first time the effectiveness of a novel groundwater treatment based on the use of co-immobilised microalgae and bacteria in combination with a cork filter (CF) for the attenuation of nitrates, pesticides (atrazine and bromacil), and antibiotics (sulfamethoxazole and sulfacetamide) at different hydraulic retention times (HRTs). Furthermore, the study will identify and link bacterial, fungal, and microalgal populations with groundwater pollutant attenuation and the formation of micropollutant transformation products (TPs).

2. Material and methods

2.1. Chemicals and reagents

Sulfamethoxazole (SMX) and atrazine (ATZ) were provided by Fluka Analytical ™, and sulfacetamide (SCM) and potassium nitrate (KNO_3) were purchased from Sigma–Aldrich, whereas bromacil (BMC) was obtained from PolyScience, Niles, USA and potassium dihydrogen phosphate single-phase (KH_2PO_4) from Merck KGaA. All described reagents were analytical grade (>95%). Carbon dioxide was supplied from an Alphagaz ™ CO_2N38 brand carbon dioxide (CO_2) cylinder with ≤10 mg/L of O_2, ≤5 mg/L of CmHm and ≤10 mg/L of H_2O.

2.2. Experimental set-up

The prototype of the water-treatment system consisted of two main units, the photobioreactor (PBR), made of methacrylate (90 mm Øext, 84 mm Øint, 1120 mm height), and the cork filter (CF), made also of methacrylate (90 mm Øext, 84 mm Øint, 400 mm height) but covered by aluminium foil to block the light (Fig. 1). Paulmann SimpLED lights 7.5 m/20 W/12 V DC were used to light the PBR in a 12 h light/12 h dark cycle. The filling height of the PBR was set at 1 m and polyurethane foam cubes (10 × 10 mm) were introduced until the upper half of the container was filled. The PBR was inoculated with a microalgae consortium obtained from a 25 L growing container fed with groundwater and inoculated with surface water from agricultural irrigation channels (Prat de Llobregat, Barcelona, Spain) containing microalgae and bacteria. The microalgal consortium was pre-acclimatised to the growth conditions for more than 6 months before the PBR was stocked with it. Natural granulated cork (2 mm in diameter) was used to fill the CF unit, which was preacclimated with groundwater at an HRT of 8 days for six months.

The most important design parameters of the photobioreactor are CO_2-dose, light intensity and quality, HRT, temperature, and stirring system (Huang et al., 2017). CO_2 was injected at 0.03% of the total water volume of the PBR (5.6 L); the light generated 105 μmol/m^2s, measured inside the methacrylate reactor by means of a photoradiometer HD2302.0 and a LP471PHOT Probe of Lighting provided by Delta Ohm. The PBR was continuously stirred by means of a magnetic stirrer (825 rpm).

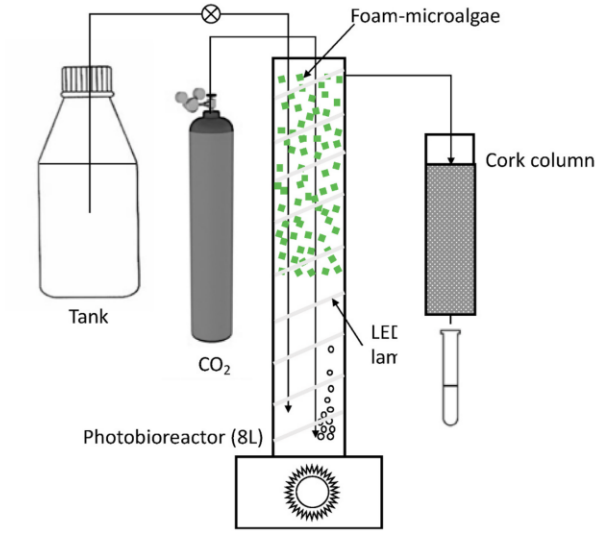


Fig. 1. Prototype design showing the 8L photobioreactor and cork filter.

The prototype was fed with real groundwater collected from a well located in the metropolitan area of Barcelona city with a chemical composition of 60 mg/L of NO_3^- and approximately 0.5 mg/L of PO_4^{3-}. Groundwater was spiked with nitrates at a concentration of 200 mg/L. Furthermore, because phosphorus is a limiting factor for the growth of microalgae, KH_2PO_4 was added to the groundwater to reach a concentration of 5 mg/L of PO_4^{3-}, as described by Rezvani et al. (2017). After 51 equilibration days of the prototype operating at an HRT of 8 days, pesticides (ATZ and BMC) and antibiotics (SMX and SCM) were added at a concentration of 100 μg/L for each one, and the study started (time = 0 days).

Three HRTs (8, 4, and 2 days) were tested by changing the peristaltic pump parameters. The prototype was tested in a temperature-controlled room at 23 ± 5 °C. The pH was monitored every week and maintained at around 7 throughout the experiment. The water, along with the contaminants added, was changed every 7 days.

2.3. Sampling strategy

Nitrates, pesticides, and antibiotics were monitored both at the inlet and in the different intermediate sections of the prototype (effluent of the PBR and CF respectively). Samples were collected on days 10, 11, 13, 17, 20, 24, and 27 after the feeding reactor was spiked with pollutants (day 0). After that, the prototype was set to a HRT of 4 days and samples were taken at days 31, 33, 35, 38, 40, 42, 47, and 49. Finally, the prototype flowrate was modified to a HRT of 2 days and samples were collected at days 52, 55, and 57.

Microbial characterization of the prototype was performed during a single sampling campaign in the last week of its operation at an HRT of 8 days. Total bacterial, fungal, and microalgal populations were characterised by collecting 3 samples of foam material and 3 of cork material (mixture of aliquots taken at a depth of between −2 and −10 cm from the CF top) for microbiological analysis. The biomass content (microalgae, bacteria, and fungus) in the foam of the PBR was on average 4 ± 1 mg in dry weight per foam cube.

2.4. General water quality parameters and micropollutants analysis

Nitrate and nitrite concentrations in the water were analysed using Hach Lange Nitrate (LCK339 and LCK340) and Nitrite (LCK341) cell tests on a Hach Lange DR 1900 Portable Spectrophotometer.

The determination of ATZ, BMC, SMX, and SCM in aqueous samples was carried out as follows. 1 mL of each water sample was filtered through a 13 mm diameter polytetrafluoroethylene (PTFE) filter with a pore size of 0.22 μm. The filtered samples were then injected into a Nexera X2 ultra high-performance liquid chromatograph (UHPLC) equipped with a photodiode array detector (SPD-M30A) (Shimadzu UK, Milton Keynes, UK). The chromatographic separation was achieved on a coreschell Ascentis® Express RP-Amide column (10 cm × 2.1 mm, 2.7 μm particle size, Supelco, Bellefonte, USA) with a guard column (0.5 cm × 2.1 mm, Supelco, Bellefonte, USA) containing the same packing material. The flow rate and injection volume were 0.35 mL/min and 25 μL, respectively. A binary gradient elution programme consisting of mobile phase A (water with 0.1% formic acid) and mobile phase B (acetonitrile with 0.1% formic acid) was set as follows: isocratic 0–1 min: 10% of B; 1–10 min: 10–100% of B; isocratic 10–15 min: 100% of B. The column oven and autosampler temperatures were set at 25 °C and 15 °C, respectively. To visualise the UPLC-UV results, Nexera Labsolutions software was used, which is from the manufacturer (Shimadzu Corporation, Kyoto, Japan). The linearity of the analytical methodology ranged from 1 to 200 μg/L with a repeatability lower than 10%.

Additionally, samples from the PBR and CF outlets, as well as groundwater samples, were analysed in an HPLC-Orbitrap to screen for possible TPs. Full scan and all ion fragmentation (AIF) acquisition modes were used with a mass range of *m/z* 50–1000. AIF as the data-independent analysis was performed using Higher-energy Collisional Dissociation (HCD) fragmentation with collision energies of 10 and 60 eV, and the resolution was set at 50,000 at a scan rate of 2 Hz. Further MS specifications as well as chromatographic conditions are presented in supplementary material (SM section). A list of possible TPs was built based on literature research of previously reported TPs for each of the tested compounds. Then, their exact masses and, when available, previously reported fragments in MassBank (https://massbank.eu) were also collected. The compiled list of compounds and masses can be seen in the SM section. The gathered data was then used in TraceFinder 3.3 EFS (Thermo Fisher Scientific, Bremen, Germany) to conduct a tentative identification of TPs in the analysed samples. Molecular ions (M+1) were detected with mass accuracy below 2 ppm and qualifier fragment ions were also provided.

2.5. Microbial community assessment

The quantification of total bacteria, fungi and microalgae from PBR and CF samples was performed by using PowerSoil™ DNeasy Isolation Kit (Qiagen). Total genomic DNAs were obtained from independent triplicate samples of each material. Gene copy numbers of 16S rRNA (total bacteria), ITS1 rRNA (total fungi) and the clade I *nosZ* gene (typical denitrifiers) were quantified by quantitative real time PCR (qPCR). The metataxonomic assessment of microbial populations (microalgal, bacterial and fungal populations) in the PBR and CF samples were characterised by paired-end High Throughput Sequencing (HTS) of V4 18S rRNA, V3–V4 16S rRNA and 5.8S-ITS2-28S rRNA amplicons, respectively. Raw data (R1 and R2 demultiplexed FASTQ files) from 16S rRNA (bacteria), ITS2 rRNA (fungi) and V4–18S rRNA (microalgae and other eukaryotes) were further processed using Cutadapt and DADA2 software. Further details on microbial community assessment and data curation are described in the SM section.

2.6. Data analysis

The experimental results were analysed statistically using the SPSS v. 24.0 software (Chicago, IL, U.S.A),. Since the samples were non-parametric and independent, the Kruskal–Wallis test, and Mann–Whitney tests were used to examine differences between the removal efficiencies of micropollutants in the different HRTs evaluated and in microbial diversity indexes between materials (foam versus cork).

3. Results and discussion

3.1. Attenuation of nitrates and nitrites

Table 1 shows the attenuation of nitrates according to the different assessed HRTs. After the prototype had been operated with the same influent nitrate concentration for 50 days (200 mg/L), the nitrate removal observed with an HRT of 8 days averaged 58%. The effectiveness of each of the studied units was similar: 28% for the PBR and 30% for the CF unit. However, the attenuation of nitrate was reduced drastically to 12 and 5% at HRTs of 4 and 2 days respectively, showing a great HRT dependence.

The attenuation of nitrates by the PBR is partly or largely attributable to microalgal assimilation for growth. Microalgae are capable of transporting and reducing nitrate, to finally incorporate inorganic N in form of organic N as glutamate (Sanz-Luque et al., 2015). Denitrification of nitrate by bacteria could also be part of the explanation, as denitrification bacteria have been identified in high abundance (see previous section). The results are also in agreement with previous studies carried out at laboratory-scale with microalgae immobilised in foam showing an attenuation of nitrates of 40–50% with an HRT of 8 days (Ferrando and Matamoros, 2020). Existing studies on the use of microalgae to remove nitrates from groundwater show that they can be a very effective solution, with removal efficiencies up to 80% with an incubation time of 3 days (Rezvani et al., 2019). Fierro et al. (2008) observed that Chitosan immobilisation of *Scenedesmus* sp. cells resulted in a 70% nitrate removal within 12 h, at a rate significantly higher than free-living cells (20% nitrate removal within 36 h of treatment). The effectiveness of our immobilised consortium in foam, including microalgae, bacteria, and fungi was lower, but this difference can be accounted for by the fact that the studies were conducted in continuous operational mode, whereas the other studies were performed in batch.

The attenuation of nitrates in the CF (30% at a HRT of 8 days) can be mainly explained by the presence of denitrifying bacteria in the biofilm. Our results suggest less effective bioremediation than that achieved by Aguilar et al. (2019), who observed that cork media used in constructed wetlands are very effective for reducing nitrates from agricultural run-off water (80–90% attenuation). This difference probably reflects the source of the water: whereas we were operating the plant with groundwater spiked at 200 mg/L of nitrates, Aguilar et al. treated agricultural run-off containing very low concentrations of nitrates but containing other nutrients and organic matter, which will have enhanced biofilm development.

Table 1
Concentration and attenuation of nitrates and nitrites in the different sampling sites in the prototype and at different HRTs. Removal efficiencies in the PBR and CF are shown in brackets.

Compound Name	HRT (d)	Groundwater (mg/L)	PBR effluent (mg/L)	CF effluent (mg/L)	Global Removal (%)
Nitrate	8	202 ± 6	146 ± 11 (28%)	85 ± 31 (42%)	58 ± 10 (a)
	4	210 ± 4	195 ± 17 (7%)	186 ± 7 (5%)	12 ± 8 (b)
	2	213 ± 4	204 ± 2 (4%)	203 ± 3 (nr)	5 ± 4(b)
Nitrite	8	0.99 ± 0.43	0.11 ± 0.02 (89%)	0.11 ± 0.11 (nr)	89 ± 4
	4	0.41 ± 0.16	0.10 ± 0.07 (63%)	0.43 ± 0.33 (−330%)	nr
	2	0.19 ± 0.01	0.12 ± 0.01 (37%)	1.87 ± 0.21 (−884%)	−881 ± 89

nr no removal, different letters reflect statistical differences between HRTs (p-value<0.05).

At the end of the treatment (HRT = 8 days), the quality of the water was still slightly greater than the regulatory limit (50 mg/L) for water intended for human consumption (91/676/EEC). Nevertheless, taking into consideration the effectiveness of the assessed technology, it can be a suitable solution for nitrate polluted groundwater with a nitrate content of up to 100 mg/L.

Regarding nitrites (Table 1), their concentration was reduced by 89% at an HRT of 8 days. This reduction was due to the PBR, with only 2% being removed in the CF. Reduction of the HRT to 4 days resulted in a reduction of the nitrite attenuation by 50% in the PBR, but led to the production of nitrite in the CF, so that overall there was no N removal. Further reduction of the HRT to 2 days, resulted in a 37% decrease in the removal of nitrate and a final overall production of nitrate of 1.9 mg/L. Whereas nitrite removal in the PBR can be explained by microalgae assimilation, bacterial denitrification is suggested as the main mechanism in cork. The nitrite occurrence in the cork filter at low HRT (4 and 2 days) may suggest that the bacterial denitrification process could be hampered by nitrite accumulation (Rajta et al., 2020), due to partial denitrifying processes that could be related to a lower production of C-labile compounds from cork under low HRT.

3.2. Attenuation of pesticides and antibiotics

Fig. 2 shows the concentrations of pesticides and antibiotics in each of the sampling sites (reservoir tank, PBR outlet, and CF outlet) over time at an HRT of 8 days. In the case of the PBR, it can be shown that whereas antibiotic depletion over time was constant (around 50%), the attenuation of pesticides decreased (from 50–66% to 21–35%). These findings suggest that pesticides may accumulate in the foam biofilm until they reach biofilm sorption capacity or that their presence may have an adverse effect on microbial communities, reducing their effectiveness for removing pesticides. Nevertheless, taking into account that the selected pesticides are highly polar (log Dow <2 at pH = 7–8), no sorption into the biomass can be hypothesized. In fact, in previous studies (Ferrando and Matamoros, 2020) we observed that pesticide adsorption to the polyurethane foam was negligible and that assimilation into microalgae was very low (Matamoros and Rodríguez, 2016), suggesting that biodegradation is the main attenuation mechanism.

Regarding antibiotics, our results again agree with previous studies, which have indicated that photodegradation is the most relevant attenuation process in microalgae systems (de Godos et al., 2012; Ferrando and Matamoros, 2020). In contrast to the PBR foam, the concentrations of antibiotics and pesticides at the outlet of the CF were very constant over time, showing no time dependence (between 1 and 2 μg/L depending on the OMCs). The attenuation efficiency of selected OMCs by the cork system was greater than 90% (92–99%). This in agreement with previous laboratory sorption studies that indicate that cork has a great capacity for removing phenolic compounds, especially halogenated phenolic compounds such as PCP and 2,4-DCP, which only require 4.9 and 29 g/L of cork, respectively, to reduce their concentrations from 1 to 0.1 mg/L. de Aguiar et al. (2019) also observed that cork has a great bioadsorption capacity for atrazine and other pesticides from water (spiked at 10 μg/L), showing average removal efficiencies of between 60% and 70% following 360 min of incubation at pH 7. Mallek et al. (2018) demonstrated that pharmaceuticals such as diclofenac and triclosan at a concentration of mg/L were 100% removed by sorption to cork. New insights into the interactions between cork chemical components and pesticides indicate that lignin moieties are the main components involved in the sorption process (Olivella et al., 2015). Hence, although previous sorption studies have suggested that cork could be used for removing OMCs, our results demonstrate, for the first time, that biofilters filled with cork are a practical solution for removing OMCs from groundwater.

The decrease of the HRT to 4 and 2 days resulted in a significant reduction in the attenuation of selected micropollutants by the prototype (Table 2). Nevertheless, whereas the reduction in attenuation by

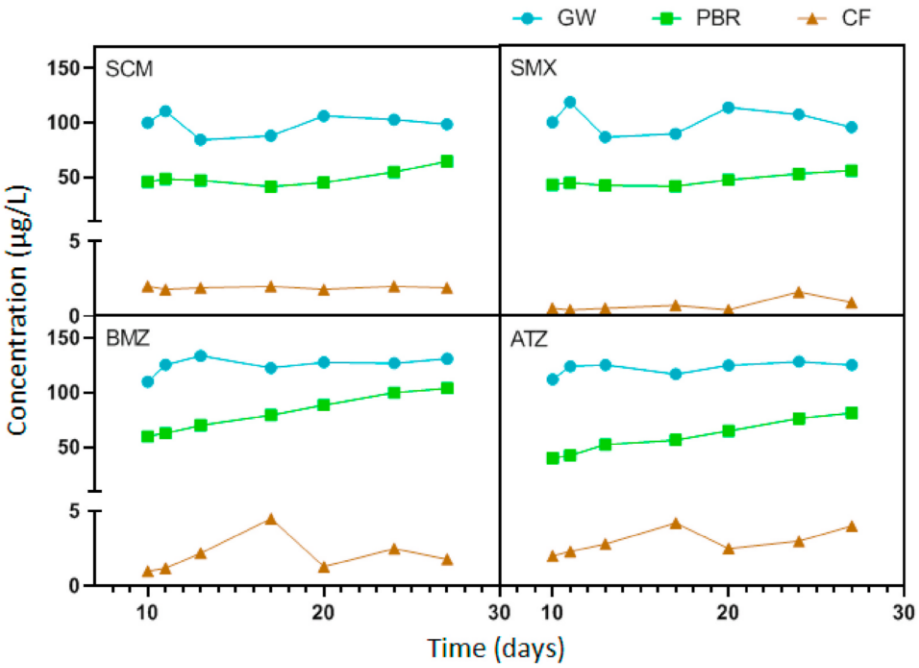


Fig. 2. Evolution of antibiotics and pesticides concentration over time at an HRT of 8 days. Groundwater (GW), photobioreactor (PBR), and cork filter (CF). Sulfacetamide (SCM), sulfamethoxazole (SMX), bromacil (BMZ), and atrazine (ATZ).

Table 2
Percentage of attenuation of pesticides (BMC and ATZ) and antibiotics (SCM and SMX) in each of the prototype units (PBR and CF). In parenthesis: the attenuation of microcontaminants considering only the CF.

Compound Name	Section	HRT = 8d (%)	HRT = 4d (%)	HRT = 2d (%)
BMC	PBR	34 ± 13 (a)	7 ± 2 (b)	1 ± 1 (c)
	CF	64 ± 12 (97 ± 12) (a)	73 ± 5 (80 ± 6) (a)	34 ± 1 (35 ± 1) (b)
	Total	98 ± 9 (a)	80 ± 5 (b)	45 ± 1 (c)
ATZ	PBR	50 ± 11 (a)	16 ± 6 (b)	6 ± 2 (c)
	CF	48 ± 11 (95 ± 8) (a)	50 ± 15 (66 ± 17) (a)	17 ± 4 (18 ± 4) (b)
	Total	97 ± 8 (a)	66 ± 4 (b)	23 ± 3 (c)
SCM	PBR	49 ± 9 (a)	26 ± 2 (b)	14 ± 1 (c)
	CF	50 ± 8 (96 ± 2) (a)	71 ± 2 (97 ± 1) (b)	83 ± 2 (96 ± 1) (c)
	Total	98 ± 6 (a)	97 ± 2 (a)	97 ± 1 (a)
SMX	PBR	53 ± 7 (a)	29 ± 3 (b)	32 ± 2 (b)
	CF	46 ± 7 (98 ± 3) (a)	61 ± 2 (89 ± 4) (b)	32 ± 4 (48 ± 8) (a)
	Total	99 ± 5 (a)	90 ± 4 (b)	64 ± 4 (c)

Different letters reflect statistical differences between HRTs (p-value<0.05).

the PBR was from 47% to 20% and finally to 13% (on average) as the HRT was reduced from 8 to 4 and then to 2 days, the CF was more resilient to HRT changes, decreasing from 97% to 83% and finally to 49%. This difference can be explained by the greater effectiveness of the CF, suggesting that sorption and biodegradation processes in the CF are more resilient than those in the PBR (photodegradation and biodegradation).

Pesticides showed the greatest dependence on HRT changes, whereas in the case of sulphonamides, the effect was much lower. This can be attributable to the attenuation mechanism of each family of compounds. As explained above, pesticides will be removed mainly by sorption and biodegradation, but antibiotics by photodegradation and biodegradation. The decrease in the HRT may therefore have a greater impact on the sorption process taking place in the CF. Previous studies carried out with co-immobilised microalgae in foam at a laboratory scale showed a similar reduction in the attenuation of micropollutants (Ferrando and Matamoros, 2020); for example, the attenuation of atrazine and bromacil decreased from 51% to 75%–31% and 57%, respectively, when the HRT was reduced from 8 days to 2 days, whereas no or a slight reduction in the attenuation of antibiotics (SMX and SM) was observed. Nevertheless, no atrazine attenuation was observed in free microalgae reactors operated at HRTs of 8, 4 or 2 days (Matamoros and Rodríguez, 2016), showing the great relevance of co-immobilisation for reducing pesticides, probably due to sorption and biodegradation processes (Ferrando and Matamoros, 2020). Conversely, the attenuation of micropollutants in the CF showed a great resilience to HRT changes, as observed also by Matamoros and Franco (2018) in pine bark biofilters, where the increase in hydraulic loading rate from 0.3 m/d to 3 m/d only reduced the attenuation of atrazine by half (from 61% to 31%) and for fenitrothion and diazinon from 90 to 89% and 87%–84%, respectively.

In the present study, the overall attenuation of antibiotics and pesticides was greater than 97% at a HRT of 8 days, and it decreased slightly at HRTs of 4 and 2 days (Table 2), highlighting the high effectiveness of the combined use of microalgae and cork biofilter for removing polar pesticides and antibiotics from groundwater.

3.3. Identification of TPs

Water samples from the system operating at a HRT of 4 days were collected for the identification of TPs. No TPs were detected following the PBR treatment. In contrast, CF treatment resulted in peaks that matched the *m/z* of two atrazine TPs (2-hydroxyatrazine and atrazine desethyl) and one bromacil TP (5-bromo-3-*sec*-butyl-6-hydroxymethyluracil). All identified peaks had a mass accuracy of below 0.6 ppm. The formation of TPs in the CF was expected because this is where the major part of the attenuation of microcontaminants took place (Table 2). The atrazine TPs found are known to be the most common results of atrazine biodegradation (Wackett et al., 2002); they have been

often identified in soil or biofilters (Lin et al., 2008) (Ulrich et al., 2017). Another TP is 5-bromo-3-*sec*-butyl-6-hydroxymethyluracil, which has been registered as a bromacil metabolite in aerobic soil conditions (Lewis et al., 2016) and is also reported as a major bromacil metabolite in plant roots (Jordan and Clerx, 1981). These results confirm that biodegradation of pesticides took place in the CF.

3.4. Overall effectiveness of the developed technology and limitations of the study

In comparison with other available nitrate-removal technologies such as ion exchange, electrochemical reduction, electrodialysis, and activated carbon adsorption (Archna et al., 2012), the nitrate attenuation achieved by the microalgae–cork prototype was only moderate. Nevertheless, the technology developed here has the advantage that there is no need for external energy (membrane-based solutions) or organic matter addition (biological denitrification treatments). Furthermore, future optimization of the prototype, such as the use of other microalgae immobilisation materials (de-Bashan and Bashan, 2010) or the replacement of the cork by wood chips or wheat straw, may increase its effectiveness for removing nitrates (Saliling et al., 2007; Schipper et al., 2010). Finally, the prototype was very effective for removing OMCs, achieving values comparable to those found in membrane groundwater technologies (Plakas and Karabelas, 2012) and greater than those found in rapid sand filtration systems (Hedegaard and Albrechtsen, 2014).

3.5. Microbial community assessment

A detailed description of the composition of the microbial communities (microalgae, bacteria, and fungus) is provided in the SM section. The results included in this section are those related to understanding the nitrate and microcontaminants biodegradation mechanisms that occurred in the PBR-cork system.

3.5.1. Microalgae population

The floating polyurethane foam cubes were rapidly colonised by visible microalgae biomass. Microscopical examination (Fig. S1) showed that most of the microalgae present were colonial and unicellular green algae, many of which could not be confidently identified at species level, and sometimes even at a genus level. The V4–18S rRNA metabarcoding data gave a much clearer picture of community composition. Both microalgal diversity and richness were quite similar in the floating foam material in the PBR and in the CF (Mann–Whitney, $P > 0.05$) (Table 3). Microalgal diversity was lower than the bacterial diversity in both materials and lower than the fungal diversity in the CF (Mann–Whitney, $P < 0.05$). This is likely due to the fact that a high proportion of the microalgal biomass (and hence DNA) present in the CF could be non-active, representing detached biomass from the foam material of the PBR. The dominant microalgal species (>70% relative abundance) was *Tetradesmus obliquus* (formerly *Scenedesmus obliquus* or *Acutodesmus obliquus*), with other *Tetradesmus*, *Chlorella,* and *Coelastrella* species also abundant but in much lower percentages (Fig. 3). *Tetradesmus obliquus* and some *Chlorella* species have been shown elsewhere to contribute significantly to the reduction of nutrients and heavy metals in different types of wastewaters (Kim et al., 2016; Rugnini et al., 2019; Yang et al., 2015).

Table 3
Richness and diversity indexes of Bacteria, Fungi and microalgae calculated from 16S rRNA, ITS2 region and V4 18S rRNA amplicon sequencing reads respectively.

Sampling point	Shannon (H)	Inv. Simpson (I/D)	Richness (OTUs)	Richness (Chao 1)
Foam-PBR (Bacteria)	3.86 ± 0.11 (a)	18.03 ± 3.15 (a)	197.97 ± 8.38 (a)	237.60 ± 19.11 (a)
Cork Filter (Bacteria)	5.58 ± 0.08 (b)	121.73 ± 23.8 (b)	579.46 ± 2.61 (b)	610.9 ± 3.5 (b)
Foam-PBR (Fungi)	0.74 ± 0.66 (a)	2.03 ± 0.98 (b)	3.67 ± 2.51 (b)	5.00 ± 3.46 (a)
Cork Filter (Fungi)	3.13 ± 0.09 (b)	7.36 ± 0.92 (b)	124.67 ± 1.15 (b)	124.67 ± 1.15 (b)
Foam-PBR (Microalgae)	1.34 ± 0.16 (a)	1.77 ± 0.25 (a)	58.12 ± 3.25 (a)	64.91 ± 1.13 (a)
Cork Filter (Microalgae)	1.31 ± 0.05 (a)	1.98 ± 0.08 (a)	45.64 ± 2.26 (b)	63.18 ± 16.84 (a)

Different letters, for each index and kingdom and materials, reflects statistical differences between materials.

3.5.2. Bacterial diversity

Both bacterial diversity and richness were significantly higher (Mann–Whitney, $p < 0.05$) in the CF (H: 5.58 ± 0.08, and Chao1: 610.9 ± 3.5) than in the PBR foam (H: 3.86 ± 0.11 and Chao1: 237.61 ± 19.1) (Table 3). Interestingly, bacterial biomass (measured as 16S rRNA copies) was abundant and comparable in both materials (*t*-Test, $p = 0.309$), accounting for 4–6 × 10^9 16S rRNA copies g^{-1} (Fig. S2). The bacterial population in CF was statistically more diverse in terms of alpha diversity and in the number of taxonomic groups (at the family level), but not significantly at ASV level (beta diversity, ANOSIM, $R = 1$, $p = 0.1$). Fig. 4 shows that the main assigned genera that were significantly abundant in the PBR foam based on LDA (LefSe analysis, Kruskal–Wallis, $p < 0.05$, LDA score above 4.5) (Fig. S4; Table S5) were *Rhodanobacter* and *Rhodopseudomonas*; the first one is a Gammaproteobacteria and comprised 10.6% of the reads, whereas the second one is an Alphaproteobacteria and comprised 5.6% of the reads. It is noteworthy that these genera have been previously described as denitrifying bacteria (Dunstan et al., 1982; Gao et al., 2021; Lee et al., 2002; Liu et al., 2020). Furthermore, their abundance could be related to the high *nosZ* gene population detected by qPCR in the foam material (Fig. S2). Therefore, we suggest that denitrifying activity could take place in the inner parts of the foam biofilms where oxygen would be less available.

In the PBR foam, the dominating taxa was Alphaproteobacteria class (Bradyrhizobiaceae, Sphingomonadaceae, Phyllobacteriaceae, Caulobacteraceae, and Rhizobiaceae genus) (Fig. S5). For example, recent studies have observed that Sphingomonadaceae and/or Caulobacteraceae enhance the biodegradation of certain micropollutants and nutrients from water (Oh and Choi, 2019; Xu et al., 2018). Conversely, the most abundant genus in the CF was *Simplicispira* (Betaproteobacteria, with 6.2% relative abundance), which has recently been described as containing several denitrifying species (Siddiqi et al., 2020). In another recent study, *Simplicispira* spp. were highly enriched in sequencing batch reactors exposed to synthetic wastewater containing antibiotics such as trimethoprim/sulfadiazine (Kruglova et al., 2019). *Simplicispira* did also occur in the PBR foam, at relative abundances of 1–3%. A further noteworthy aspect is the common occurrence in the CF of *Phenylobacterium* (Alphaproteobacteria), *Asprobacter* (Alphaproteobacteria), *Actinomarinicola* (Actinobacteria), and *Steroidobater* (Gammaproteobacteria), all of which could also be related to the degradation of micropollutants. Recent studies, such as Liao et al. (2016) and Espín et al. (2020) have described *Phenylobacterium* as a ciprofloxacin- and atrazine-degrading bacterium. The denitrifier *Steroidobacter* (2% of relative abundance in CF) has been correlated positively with sulphonamide degradation (Zhang et al., 2021) including the family Iamiacea (3% of relative abundance in CF), which includes the genus *Actinomarinicola*, already mentioned above.

The constant lighting received by the PBR foam and the availability of O_2 in excess, due to photosynthesis, could hamper the denitrifying activities of such high facultative denitrifying populations. In contrast, the absence of light and lower availability of O_2 (in the deepest regions of biofilms) in the interstitial water among cork particles in the CF could lead to bacterial denitrifying activity (especially high in TRH-8 days). Total denitrifying bacterial populations were higher in the PBR foam (*t*-test, $p = 0.03$), achieving values of 9 × 10^8 *nosZ* copies g^{-1} compared to

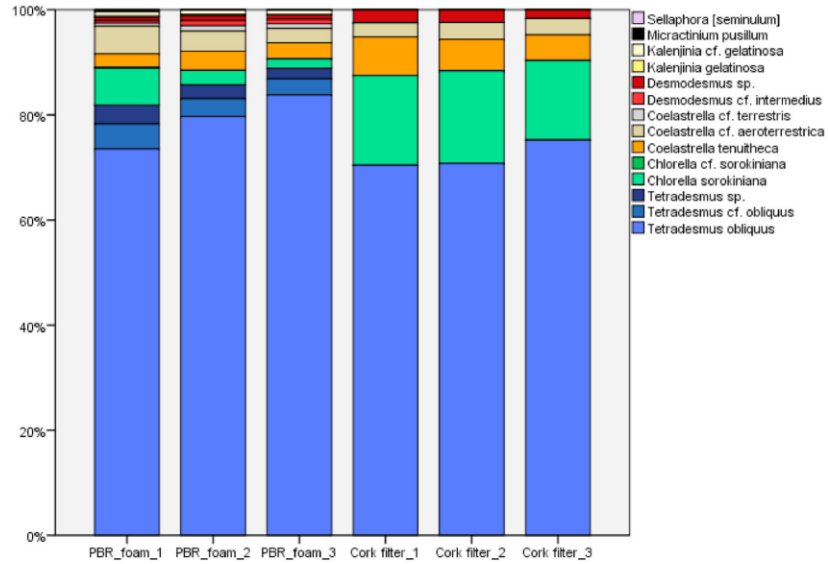


Fig. 3. Relative abundance of microalgal sequencing reads at species level (from V4–18S metabarcoding) in the photobioreactor (PBR) foam and the cork filter. The relative abundances are the number of reads (sequences) assigned to any given taxon, divided by the total number of reads per sample that are assignable to any autotrophic eukaryote.

the values in the CF, even though these were also high (5.7×10^7 *nosZ* copies g^{-1}).

In this work, functional gene copy numbers of the *nosZ* clade I were used to assess the abundance of denitrifying populations because typical denitrifiers are known to reduce N_2O to nitrogen gas at different rates (Hallin et al., 2018). Also, it is described that clade I *nosZ* consists almost exclusively of the Alpha-, Beta-, and Gammaproteobacteria classes. Considering qPCR and NGS results, we can conclude that both bioreactors possessed high denitrification capacity, not only because of the *nosZ* gene copy numbers, but also because the predominant bacterial classes have denitrification potential. Nevertheless, although both systems have a high denitrification capacity, our results show that nitrate attenuation is low (20–60%). This is in agreement with recent findings stating that the transcription rate of denitrification genes depends strongly on environmental conditions (Chon and Cho, 2015), and therefore, having a high oxygen concentration in both systems would reduce it. Consequently, future studies should explore the denitrification effectiveness of the CF operated under oxygen-limiting conditions (water-saturated columns).

3.5.3. Fungal diversity

Our results indicate a high fungal diversity in the CF, probably due to the high organic content of the cork material and cell debris (Table 3). Fungal populations in the CF (Fig. 5) were dominated by *Humicola nigrescens* (30% of relative abundance). This belongs to the Sordariomycetes, and some genera in this group are producers of phytase to improve phosphorus bioavailability from organic matter (Bala et al., 2014). Also, unclassified Ascomycota (10%), unclassified Chaetomiaceae (8%, Sordariomycetes), *Penicillium* sp. (3%) and *Conlarium duplumascospora* (3%) were abundant. It is noteworthy that *Conlarium duplumascospora* has been shown to play an important role in degrading woody debris and leaves in submerged freshwater environments (Liu et al., 2012). The fungal community in the filter was diverse and could be linked to the utilisation of cork material together with cell debris from detached PBR biofilms that flow to the filter acting as periphyton (a mixture of algae, cyanobacteria, heterotrophic microbes, and detritus attached to solid substrates that are able to release C-labile exudates in aquatic ecosystems). Microbial interactions in the CF could produce additional degradable organic matter from the cork, contributing to additional denitrification processes (Mendonça et al., 2004). Cork material could initially be metabolised by fungi and bacteria, and degradable organic matter could be further used as an electron source by denitrifying microbiota, which have been detected as highly abundant both in PBR and CF.

Previous studies have identified that the Ascomycota (the predominant phylum in the CF) encompasses species (besides Basidiomycota) that also possess ligninolytic enzymes such as laccase (EC 1.10.3.2) (Osono, 2020). Laccase is a well-known enzyme with a high pollutants biodegradation capacity, especially for recalcitrant ones (García-Delgado et al., 2018; Medaura et al., 2021). For example, Cupul et al. (2014) performed enzymatic assays on Basidiomycota ligninolytic fungi with atrazine and observed an increase in laccase activity and atrazine removal. Furthermore, Esparza-Naranjo et al. (2021) showed that different lignin-degraders isolated from leaf litter (such as *Fusarium* sp., Ascomycota phylum) were also able to degrade atrazine without expressing laccase activity. These findings revealed the importance of other enzymatic capacities that deserve attention, such as cytochrome P450 monooxygenase or unspecific peroxygenases (UPOs), where the latter were studied in sulfonamide degradation by Basidiomycota UPOs (Lemańska et al., 2021).

Overall, our results suggest that isolated bacterial communities from PBR and CF would have high denitrification capacity, but our results indicate that they were only capable of removing nitrates moderately, with removal efficiencies of around 58% at a HRT of 8 days. This is in agreement with the fact that although denitrification capacity existed, bacteria were not capable of performing it due to the predominant aerobic conditions of both systems. Nevertheless, the presence of bacteria and fungus with a high demonstrated capacity for removing micropollutants is in agreement with the high effectiveness observed by both the PBR-CF systems for removing pesticides and antibiotics (>95% at a HRT of 8 days).

4. Conclusions

The results of the study show that the combined use of microalgae and cork filtration is effective for the treatment of groundwater

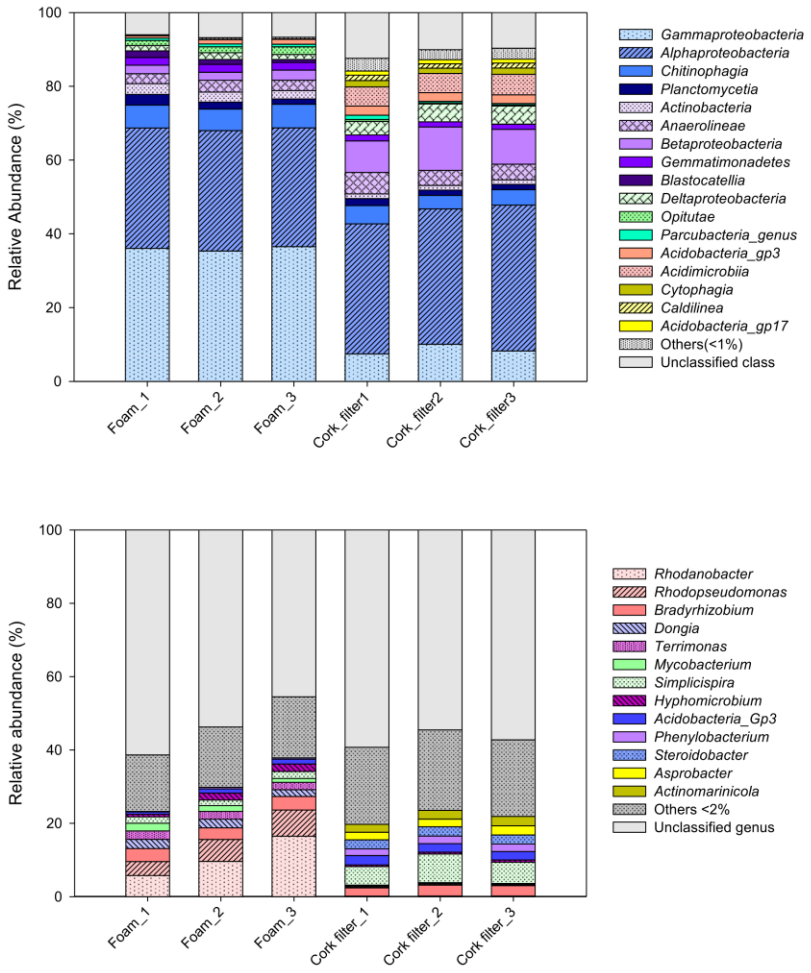


Fig. 4. Relative abundances of taxonomically assigned bacterial reads, at class level (above) and genus level (below), in the photobioreactor (PBR) foam and the cork filter of the NDN bioreactor. Relative abundance was defined by the number of reads (sequences) affiliated with any given taxon, divided by the total number of reads per sample. Phylogenetic groups with relative abundance >1% and >2% were categorized as 'Others'.

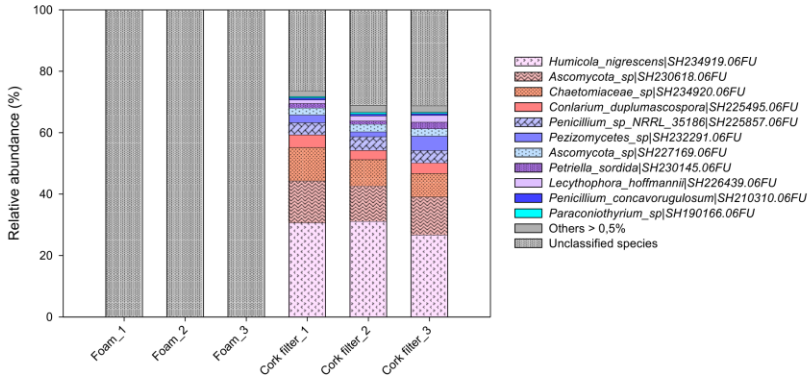


Fig. 5. Relative abundance of taxonomically assignedfungal reads, at species level, in the photobioreactor (PBR) foam and the cork filter of the NDN bioreactor. Relative abundance was defined by the number of reads (sequences) affiliated with any given taxon, divided by the total number of reads per sample. Phylogenetic groups assigned to the Fungi with relative abundances >0.5% were categorized as 'Others'. 'Unclassified species' include non-fungal reads, including microalgae.

contaminated by nitrates, and tested pesticides and antibiotics. The main results and key conclusions can be summarised as follows:

- The PBR-CF prototype removes nitrates (58%) and nitrites (89%) at an HRT of 8 days, but it fails at lower HRT (<20%).
- The combined use of PBR and CF enabled attenuation of antibiotics and pesticides up to 95% at an HRT of 8 days, but this decreased with decreasing HRT, with pesticides being the compounds most affected (changing from 97 to 98% attenuation to 23–45% with a reduction of the HRT from 8 to 2 days). We hypothesise that the release of C-labile molecules from PBR and cork material can promote denitrification and micropollutant degradation.
- The identification of atrazine and bromacil TP in CF outputs indicated that biodegradation was the main attenuation process.
- The most abundant microbiological species were the green alga *Tetradesmus*, in both the PBR and the CF. Nevertheless, molecular analysis confirmed that both bioreactors were enriched in denitrifying populations able to perform denitrification.
- The CF had more bacterial and fungal diversity than the PBR, indicating a higher potential for pollutant biodegradation. The attenuation of micropollutants was linked to the presence of certain microorganism genera and species.

The results are very promising. However, the low efficiency of the system in terms of nitrate attenuation at low HRT values will require further studies, such as testing other materials for both microalgae immobilisation and improving the biofilter system.

Credit author statement

Lorenzo Rambaldo: Investigation, Data curation, Writing – original draft preparation. **Héctor Ávila:** Investigation, Data curation, Writing – original draft preparation. **Monica Escolà;:** Data curation, Writing- Reviewing and Editing. **Miriam Guivernau:** Investigation, Data curation, Writing – original draft preparation. **Marc Viñas:** Writing- Reviewing and Editing. **Rosa Trobajo:** Writing- Reviewing and Editing. **Javier Pérez-Burillo:** Writing- Reviewing and Editing, **David G. Mann:** Writing- Reviewing and Editing. **Belén Fernández** Writing- Reviewing and Editing. **Carme Biel:** Funding acquisition, Writing- Reviewing and Editing. **Luigi Rizzo:** Writing- Reviewing and Editing. **Josep M. Bayona:** writing- Reviewing and Editing. **V. Matamoros:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

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

Acknowledgements

The authors wish to thank the financial support of the European Union through the project LIFE18 ENV/ES/000199 and the Spanish Ministry of Science, Innovation and Universities through Project CTM2017-91355-EXP. Finally, European Commission (Erasmus program) and Government of Chile for supporting Lorenzo Rambaldo and Héctor Avila Cortés's visit at IDAEA-CSIC in Barcelona, Spain.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.134777.

References

Aguilar, L., Gallegos, Á., Arias, C.A., Ferrera, I., Sánchez, O., Rubio, R., Saad, M. Ben, Missagia, B., Caro, P., Sahuquillo, S., Pérez, C., Morató, J., 2019. Microbial nitrate removal efficiency in groundwater polluted from agricultural activities with hybrid cork treatment wetlands. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2018.10.426.

Archna Sharma, S.K., Sobti, R.C., 2012. Nitrate removal from ground water: a review. E-J. Chem. https://doi.org/10.1155/2012/154616.

Bala, A., Sapna Jain, J., Kumari, A., Singh, B., 2014. Production of an extracellular phytase from a thermophilic mould Humicola nigrescens in solid state fermentation and its application in dephytinization. Biocatal. Agric. Biotechnol. https://doi.org/10.1016/j.bcab.2014.07.002.

Chon, K., Cho, J., 2015. Abundance and expression of denitrifying genes (narG, nirS, norB, and nosZ) in sediments of wastewater stabilizing constructed wetlands. Environ. Eng. Res. https://doi.org/10.4491/eer.2014.069.

Cupul, W.C., Abarca, G.H., Vázquez, R.R., Salmones, D., Hernández, R.G., Gutiérrez, E. A., 2014. Response of ligninolytic macrofungi to the herbicide atrazine: dose-response bioassays. Rev. Argent. Microbiol. https://doi.org/10.1016/s0325-7541(14)70094-x.

de-Bashan, L.E., Bashan, Y., 2010. Immobilized microalgae for removing pollutants: review of practical aspects. Bioresour. Technol. https://doi.org/10.1016/j.biortech.2009.09.043.

de Aguiar, T.R., Guimarães Neto, J.O.A., Şen, U., Pereira, H., 2019. Study of two cork species as natural biosorbents for five selected pesticides in water. Heliyon. https://doi.org/10.1016/j.heliyon.2019.e01189.

de Godos, I., Muñoz, R., Guieysse, B., 2012. Tetracycline removal during wastewater treatment in high-rate algal ponds. J. Hazard Mater. https://doi.org/10.1016/j.jhazmat.2012.05.106.

Dunstan, R.H., Kelley, B.C., Nicholas, D.J.D., 1982. Fixation of dinitrogen derived from denitrification of nitrate in a photosynthetic bacterium, Rhodopseudomonas sphaeroides forma sp. denitrificans. J. Bacteriol. https://doi.org/10.1128/jb.150.1.100-104.1982.

Esparza-Naranjo, S.B., da Silva, G.F., Duque-Castaño, D.C., Araújo, W.L., Peres, C.K., Boroski, M., Bonugli-Santos, R.C., 2021. Potential for the biodegradation of atrazine using leaf litter fungi from a subtropical protection area. Curr. Microbiol. https://doi.org/10.1007/s00284-020-02288-6.

Espín, Y., Aranzulla, G., Álvarez-Ortí, M., Gómez-Alday, J.J., 2020. Microbial community and atrazine-degrading genetic potential in deep zones of a hypersaline Lake-Aquifer system. Appl. Sci. https://doi.org/10.3390/app10207111.

EUROSTAT, 2018. Archive:Agri-environmental indicator - pesticide pollution of water [WWW Document]. URL. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Agri-environmental_indicator_-_pesticide_pollution_of_water.

Ferrando, L., Matamoros, V., 2020. Attenuation of nitrates, antibiotics and pesticides from groundwater using immobilised microalgae-based systems. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2019.134740.

Fierro, S., del Pilar Sánchez-Saavedra, M., Copalcúa, C., 2008. Nitrate and phosphate removal by chitosan immobilized Scenedesmus. Bioresour. Technol. https://doi.org/10.1016/j.biortech.2007.02.043.

Gao, Y., Mania, D., Mousavi, S.A., Lycus, P., Arntzen, M., Woliy, K., Lindström, K., Shapleigh, J.P., Bakken, L.R., Frostegård, Å., 2021. Competition for electrons favours N2O reduction in denitrifying Bradyrhizobium isolates. Environ. Microbiol. https://doi.org/10.1111/1462-2920.15404.

García-Delgado, C., Eymar, E., Camacho-Arévalo, R., Petruccioli, M., Crognale, S., D'Annibale, A., 2018. Degradation of tetracyclines and sulfonamides by stevensite- and biochar-immobilized laccase systems and impact on residual antibiotic activity. J. Chem. Technol. Biotechnol. https://doi.org/10.1002/jctb.5697.

Hallin, S., Philippot, L., Löffler, F.E., Sanford, R.A., Jones, C.M., 2018. Genomics and ecology of novel N2O-reducing microorganisms. Trends Microbiol. https://doi.org/10.1016/j.tim.2017.07.003.

Hedegaard, M.J., Albrechtsen, H.J., 2014. Microbial pesticide removal in rapid sand filters for drinking water treatment - potential and kinetics. Water Res. https://doi.org/10.1016/j.watres.2013.09.024.

Huang, Q., Jiang, F., Wang, L., Yang, C., 2017. Design of photobioreactors for mass cultivation of photosynthetic organisms. Engineering 3, 318–329. https://doi.org/10.1016/J.ENG.2017.03.020.

Jia, X., Connor, D.O., Hou, D., Jin, Y., Li, G., Zheng, C., Sik, Y., Tsang, D.C.W., Luo, J., 2019. Science of the Total Environment Groundwater depletion and contamination : spatial distribution of groundwater resources sustainability in China. Sci. Total Environ. 672, 551–562. https://doi.org/10.1016/j.scitotenv.2019.03.457.

Jordan, L.S., Clerx, W.A., 1981. Accumulation and metabolism of bromacil in pineapple sweet orange (Citurs sinensis) and Cleopatra Mandarin (Citrus reticulata). Weed Sci. https://doi.org/10.1017/s0043174500025716.

Kim, H.-C., Choi, W.J., Chae, A.N., Park, J., Kim, H.J., Song, K.G., 2016. Evaluating integrated strategies for robust treatment of high saline piggery wastewater. Water Res. 89, 222–231. https://doi.org/10.1016/j.watres.2015.11.054.

Kolpin, D.W., Barbash, J.E., Gilliom, R.J., 1998. Occurrence of pesticides in shallow groundwater of the United States: initial results from the National Water-Quality Assessment program. Environ. Sci. Technol. https://doi.org/10.1021/es970412g.

Kruglova, A., Mikola, A., Gonzalez-Martinez, A., Vahala, R., 2019. Effect of sulfadiazine and trimethoprim on activated sludge performance and microbial community dynamics in laboratory-scale membrane bioreactors and sequencing batch reactors at 8°C. Biotechnol. Prog. https://doi.org/10.1002/btpr.2708.

Kurwadkar, S., 2017. Groundwater pollution and vulnerability assessment. Water Environ. Res. https://doi.org/10.2175/106143017x15023776270584.

Lee, D.Y., Ramos, A., Macomber, L., Shapleigh, J.P., 2002. Taxis response of various denitrifying bacteria to nitrate and nitrite. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.68.5.2140-2147.2002.

Lemańska, N., Felis, E., Poraj-Kobielska, M., Gajda-Meissner, Z., Hofrichter, M., 2021. Comparison of sulphonamides decomposition efficiency in ozonation and enzymatic oxidation processes. Arch. Environ. Protect. https://doi.org/10.24425/aep.2021.136443.

Lewis, K.A., Tzilivakis, J., Warner, D.J., Green, A., 2016. An international database for pesticide risk assessments and management. Hum. Ecol. Risk Assess. 22, 1050–1064. https://doi.org/10.1080/10807039.2015.1133242.

Liao, X., Li, B., Zou, R., Dai, Y., Xie, S., Yuan, B., 2016. Biodegradation of antibiotic ciprofloxacin: pathways, influential factors, and bacterial community structure. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-016-6054-1.

Lin, C.H., Lerch, R.N., Garrett, H.E., George, M.F., 2008. Bioremediation of atrazine-contaminated soil by forage grasses: transformation, uptake, and detoxification. J. Environ. Qual. https://doi.org/10.2134/jeq2006.0503.

Liu, F., Hu, D.-M., Cai, L., 2012. Conlarium duplumascospora gen. et. sp. nov. and Jobellisia guangdongensis sp. nov. from freshwater habitats in China. Mycologia 104, 1178–1186. https://doi.org/10.3852/11-379.

Liu, T., He, X., Jia, G., Xu, J., Quan, X., You, S., 2020. Simultaneous nitrification and denitrification process using novel surface-modified suspended carriers for the treatment of real domestic wastewater. Chemosphere. https://doi.org/10.1016/j.chemosphere.2020.125831.

Mallek, M., Chtourou, M., Portillo, M., Monclús, H., Walha, K., Salah, A. ben, Salvadó, V., 2018. Granulated cork as biosorbent for the removal of phenol derivatives and emerging contaminants. J. Environ. Manag. https://doi.org/10.1016/j.jenvman.2018.06.069.

Matamoros, V., Franco, J., 2018. Assessing the use of sand, peat soil, and pine bark for the attenuation of polar pesticides from agricultural run-off: a bench-scale column experiment. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-018-2213-x.

Matamoros, V., Rodríguez, Y., 2016. Batch vs continuous-feeding operational mode for the removal of pesticides from agricultural run-off by microalgae systems: a laboratory scale study. J. Hazard Mater. 309 https://doi.org/10.1016/j.jhazmat.2016.01.080.

Medaura, M.C., Guivernau, M., Moreno-Ventas, X., Prenafeta-Boldú, F.X., Viñas, M., 2021. Bioaugmentation of native fungi, an efficient strategy for the bioremediation of an aged industrially polluted soil with heavy Hydrocarbons. Front. Microbiol. https://doi.org/10.3389/fmicb.2021.626436.

Mendonça, E., Pereira, P., Martins, A., Anselmo, A.M., 2004. Fungal biodegradation and detoxification of cork boiling wastewaters. Eng. Life Sci. https://doi.org/10.1002/elsc.200420018.

Nguyen, H.T., Yoon, Y., Ngo, H.H., Jang, A., 2020. Technology the application of microalgae in removing organic micropollutants in wastewater. Crit. Rev. Environ. Sci. Technol. 1–34. https://doi.org/10.1080/10643389.2020.1753633, 0.

Oh, S., Choi, D., 2019. Microbial community enhances biodegradation of bisphenol A through selection of Sphingomonadaceae. Microb. Ecol. https://doi.org/10.1007/s00248-018-1263-4.

Olivella, M.A., Bazzicalupi, C., Bianchi, A., Fiol, N., Villaescusa, I., 2015. New insights into the interactions between cork chemical components and pesticides. The contribution of π-π interactions, hydrogen bonding and hydrophobic effect. Chemosphere. https://doi.org/10.1016/j.chemosphere.2014.08.051.

Osono, T., 2020. Functional diversity of ligninolytic fungi associated with leaf litter decomposition. Ecol. Res. https://doi.org/10.1111/1440-1703.12063.

Otero, N., Torrentó, C., Soler, A., Menció, A., Mas-Pla, J., 2009. Monitoring groundwater nitrate attenuation in a regional system coupling hydrogeology with multi-isotopic methods: the case of Plana de Vic (Osona, Spain). Agric. Ecosyst. Environ. https://doi.org/10.1016/j.agee.2009.05.007.

Plakas, K.V., Karabelas, A.J., 2012. Removal of pesticides from water by NF and RO membranes - a review. Desalination. https://doi.org/10.1016/j.desal.2011.08.003.

Rajta, A., Bhatia, R., Setia, H., Pathania, P., 2020. Role of heterotrophic aerobic denitrifying bacteria in nitrate removal from wastewater. J. Appl. Microbiol. https://doi.org/10.1111/jam.14476.

Rezvani, F., Sarrafzadeh, M.H., Ebrahimi, S., Oh, H.M., 2019. Nitrate removal from drinking water with a focus on biological methods: a review. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-017-9185-0.

Rezvani, F., Sarrafzadeh, M.H., Seo, S.H., Oh, H.M., 2017. Phosphorus optimization for simultaneous nitrate-contaminated groundwater treatment and algae biomass production using Ettlia sp. Bioresour. Technol. https://doi.org/10.1016/j.biortech.2017.08.053.

Rugnini, L, Ellwood, NTW, Costa, G, Falsetti, A, Congestri, R, Bruno, L, 2019. Scaling-up of wastewater bioremediation by Tetradesmus obliquus, sequential bio-treatments of nutrients and metals. Ecotoxicol. Environ. Saf. 172, 59–64. https://doi.org/10.1016/j.ecoenv.2019.01.059.

Saliling, W.J.B., Westerman, P.W., Losordo, T.M., 2007. Wood chips and wheat straw as alternative biofilter media for denitrification reactors treating aquaculture and other wastewaters with high nitrate concentrations. Aquacult. Eng. https://doi.org/10.1016/j.aquaeng.2007.06.003.

Sanz-Luque, E., Chamizo-Ampudia, A., Llamas, A., Galvan, A., Fernandez, E., 2015. Understanding nitrate assimilation and its regulation in microalgae. Front. Plant Sci.

Schipper, L.A., Robertson, W.D., Gold, A.J., Jaynes, D.B., Cameron, S.C., 2010. Denitrifying bioreactors-An approach for reducing nitrate loads to receiving waters. Ecol. Eng. https://doi.org/10.1016/j.ecoleng.2010.04.008.

Siddiqi, M.Z., Sok, W., Choi, G., Kim, S.Y., Wee, J.H., Im, W.T., 2020. Simplicispira hankyongi sp. nov., a novel denitrifying bacterium isolated from sludge. Antonie van Leeuwenhoek. Int. J. Gen. Mol. Microbiol. https://doi.org/10.1007/s10482-019-01341-0.

Taziki, M., Ahmadzadeh, H., Murry, A., Lyon S, M.R., 2015. Nitrate and nitrite removal from wastewater using algae. Curr. Biotechnol. https://doi.org/10.2174/2211550104666150828193607.

Ulrich, B.A., Vignola, M., Edgehouse, K., Werner, D., Higgins, C.P., 2017. Organic carbon amendments for enhanced biological attenuation of trace organic contaminants in biochar-amended stormwater biofilters. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.7b01164.

Wackett, L., Sadowsky, M., Martinez, B., Shapir, N., 2002. Biodegradation of atrazine and related s-triazine compounds: from enzymes to field studies. Appl. Microbiol. Biotechnol. https://doi.org/10.1007/s00253-001-0862-y.

Xu, X.J., Lai, G.L., Chi, C.Q., Zhao, J.Y., Yan, Y.C., Nie, Y., Wu, X.L., 2018. Purification of eutrophic water containing chlorpyrifos by aquatic plants and its effects on planktonic bacteria. Chemosphere. https://doi.org/10.1016/j.chemosphere.2017.10.171.

Yang, J., Cao, J., Xing, G., Yuan, H., 2015. Lipid production combined with biosorption and bioaccumulation of cadmium, copper, manganese and zinc by oleaginous microalgae Chlorella minutissima UTEX2341. Bioresour. Technol. 175, 537–544. https://doi.org/10.1016/j.biortech.2014.10.124.

Zhang, G., Zhao, Z., Yin, X.A., Zhu, Y., 2021. Impacts of biochars on bacterial community shifts and biodegradation of antibiotics in an agricultural soil during short-term incubation. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2020.144751.