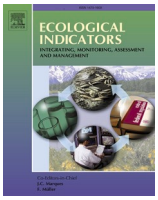
Ecological Indicators 141 (2022) 109114

Contents lists available at ScienceDirect

Ecological Indicators

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



Co-occurrence, ecological profiles and geographical distribution based on unique molecular identifiers of the common freshwater diatoms *Fragilaria* and *Ulnaria*

Maria Kahlert [a,*], Satu Maaria Karjalainen [b], Francois Keck [c], Martyn Kelly [d,e], Mathieu Ramon [f], Frederic Rimet [g], Susanne Schneider [h], Kálmán Tapolczai [i], Jonas Zimmermann [j]

[a] *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, PO Box 7050, SE-750 07 Uppsala, Sweden*
[b] *Finnish Environment Institute, Paavo Havaksen tie 3, 90570 Oulu, Finland*
[c] *Eawag: Swiss Federal Institute of Aquatic Science and Technology, Department of Aquatic Ecology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland*
[d] *Bowburn Consultancy, Bowburn, Durham DH6 5QB, UK*
[e] *Department of Geography, University of Nottingham, Nottingham NG7 2RD, UK*
[f] *Fera Science Ltd, Sand Hutton, York YO41 1LZ, UK*
[g] *INRAE, Université Savoie-Mont Blanc, UMR CARRTEL, 75bis avenue de Corzent, F-74200 Thonon-les-Bains, France*
[h] *Norwegian Institute for Water Research, Økernveien 94, 0579 Oslo, Norway*
[i] *Balaton Limnological Research Institute, Eötvös Loránd Research Network (ELKH), Klebelsberg Kuno u. 3, H-8237 Tihany, Hungary*
[j] *Research Group Diatoms, Botanic Garden und Botanical Museum Berlin, Freie Universität Berlin, Germany*

ARTICLE INFO

We dedicate this manuscript to the memory of Luc Ector, a passionate diatomist, who passed away too early.

Keywords:
Amplicon sequence variant
Freshwater
Ecological niche
Environmental assessment
Curated diatom taxonomy
*rbc*L

ABSTRACT

Diatom taxonomy has evolved in recent years, with many new species described and new approaches such as molecular genetics showing the existence of cryptic diversity within currently accepted species. This cryptic diversity is not well understood even for common freshwater genera such as *Fragilaria* and *Ulnaria.* The purpose of our study was to define taxon-specific ecological profiles and geographical distributions for unique molecular identifiers (amplicon sequence variants, ASVs) linked to curated taxonomy for these genera. Our goal is to contribute to the development of ecological assessment methods, and to the understanding why we often observe so many diatom species co-occurring in a single sample. We filtered a large (770 samples) metabarcoding dataset with linked environmental data covering several countries in Europe for genetic variants (ASVs) assigned to currently accepted species of our target genera. We studied the geographical distribution of the ASVs, and tested for ASV-pair co-occurrence. We modelled ASV-specific preferences for pH, alkalinity, total nitrogen, total phosphorus and conductivity, and analysed their preference for lakes or streams as habitat. Our study confirmed that there seems to be no general geographical barrier for the distribution of freshwater benthic diatom ASVs in Europe, but that dispersal is not rapid enough to hide historical events. The *Fragilaria* and *Ulnaria* ASVs in our study showed considerable overlap in geographical distribution, habitat and ecological preferences. We found evidence that only large differences in preferences for the analysed water chemistry variables prevented the co-occurrence of ASVs at the same sites. Instead, *Fragilaria* and *Ulnaria* ASVs co-occurred frequently in samples. We found subtle differences in ecological preferences for some ASV pairs, which might in part explain the co-occurrence by the avoidance of direct competition. However, the great overlap in distribution and ecological preferences suggests that other factors not studied here were also responsible for the observed co-occurrences and high richness of ASVs found at many sites. To our knowledge, we are the first to use ASVs in combination with a curated taxonomy to understand co-occurrence, specific ecological profiles and large-scale geographical distribution for unique identifiers not biased by the quality of reference databases, clustering methods, or non-harmonized morphological identification. Thus, our results can now be used in subsequent projects to interpret ASV occurrences, e.g. for development of ecological assessment methods.

* Corresponding author.
E-mail addresses: maria.kahlert@slu.se (M. Kahlert), Satu.Maaria.Karjalainen@syke.fi (S. Maaria Karjalainen), MGKelly@bowburn-consultancy.co.uk (M. Kelly), Mathieu.Ramon@fera.co.uk (M. Ramon), frederic.rimet@inrae.fr (F. Rimet), susi.schneider@niva.no (S. Schneider), tapolczai.kalman@blki.hu (K. Tapolczai), J.Zimmermann@bo.berlin (J. Zimmermann).

https://doi.org/10.1016/j.ecolind.2022.109114
Received 25 February 2022; Received in revised form 20 June 2022; Accepted 25 June 2022
Available online 1 July 2022

1. Introduction

Diatoms, an ecologically important group of microalgae, are frequently used worldwide for environmental assessment (Charles et al., 2021; Smol and Stoermer, 2010), and as model organisms to understand and predict biodiversity in a globally changing world (Finlay, 2002; Keck et al., 2018; Soininen et al., 2016; Vyverman et al., 2007). The basis of these analyses is knowledge of species-specific ecological profiles, which have been established studying morphospecies identified by traditional light microscopy. Diatom taxonomy has, however, evolved rapidly in recent years, with many new species described and new approaches such as molecular genetics showing the existence of cryptic diversity within currently accepted species (Evans et al., 2009; Mann et al., 2004; Sefbom et al., 2018; Sorhannus et al., 2010; Vanelslander et al., 2009). This means that ecological profiles derived from traditional light microscopy may be blurred through the inclusion of morphologically similar but genetically distinct entities within a single binomial. This problem could be resolved by deriving ecological profiles and geographical ranges directly from molecular barcodes. This novel approach could then enable us to improve ecological assessment methods, by contributing specific profiles for a range of molecular taxonomic entities, which cannot be differentiated microscopically.

Revised knowledge on diatom ecology might also help to solve open research questions in biogeographical research. Models studying the causes for modern species richness and distribution have shown that both environmental filtering and dispersal play an important role in shaping diatom assemblages, with the relative importance of these factors depending on scale (Finlay, 2002; Keck et al., 2018, and references therein; Soininen et al., 2016; Vyverman et al., 2007). These models use well-established hypotheses of community ecology, stating that species competing for the same resources cannot coexist in the long term, leading to the conclusion that co-occurring species must have different ecological niches (Verberk, 2011). However, it is still a riddle why we observe the frequent co-occurrence of diatom species with seemingly similar ecological niches (Kelly et al., 2015; Mann, 1999; Poulíčková et al., 2008). In other words, we still are "baffled by the benthos" (blog post Kelly (2015), a term adapted from Hutchinson's "Paradox of the Plankton" (Hutchinson, 1961)).

Another open question is the stability of species-specific ecological niches. Recently, Soininen et al. (2019) stated that environmental and climate models may not be transferable but seem to differ among geographical regions. Indeed, we should expect diatom species niches to change over time, through ongoing genetic fine-tuning ("microevolution", Filipchenko, 1927) of physiological mechanisms in response to changing conditions (Evans et al., 2009; Mann et al., 2004; Sefbom et al., 2018; Sorhannus et al., 2010; Vanelslander et al., 2009). Stable and unbiased ecological and geographical profiles derived from molecular units offer an opportunity to test such hypotheses.

It is important to use reproducible and clearly identified (molecular and morphological) taxonomic units for basing further research, because recent studies have shown that the morphological species concept only partly agrees with the molecular one (Kahlert et al., 2019; Pérez-Burillo et al., 2021; Tapolczai et al., 2021). The use of amplicon sequence variants (ASVs), short and unique DNA sequences, filtered from metabarcoding datasets via bioinformatics, now offers this possibility. This novel approach provides reproducible and comprehensive results of very highly resolved diversity where even one nucleotide difference between barcodes can be detected (Callahan et al., 2017). ASVs can be treated as discrete and distinct indicators of biological diversity. There is, however, also the possibility of linking ASVs to Linnean taxonomy. Using ASVs connected to the morphological species concept allows us to compare ecological profiles of species identified in two different ways (i.e. molecular or by microscope). Using ASVs also allow us to pool data from different sources and analyse patterns on large scales on a very finely resolved taxonomic level. Today, it is extremely challenging to generate taxonomically consistent datasets on large scales from morphologically identified taxa, not least because it has to take place against a backdrop of ongoing taxonomic research constantly reshuffling diatom names (Charles et al., 2021).

In the present paper, we combined data on species and ASV co-occurrence with ecological profiles of the diatom genus *Fragilaria* as well as the closely related genus *Ulnaria* to analyse whether there were as many species at our sites as suggested by morphological studies. Furthermore, it was our aim to detect differences in ecological profiles to see if these could help to explain whether coexisting *Fragilaria/Ulnaria* species can avoid direct competition by having different niches. We focused on those genera because preparatory work to connect the molecular with the morphological taxonomy has been performed already (Kahlert et al., 2019, using the rbcL marker gene). Using Kahlert et al. (2019)'s results as a starting point, our strategy was to detect species and subspecies diversity (Rimet et al., 2018) in a compiled metabarcoding dataset with accompanying data on water chemistry spanning a wide geographical range in Europe. *Fragilaria/Ulnaria* are abundant and ecologically important diatom genera in freshwater (as summarized in Kahlert et al. 2019), with species-specific ecological preferences (based on the morphological species concept) spanning the key regulatory boundary of good and moderate status as defined by the Water Framework Directive (WFD, The European Parliament and The Council of the European Union, 2000). However, there is also general agreement that the separation of species within these genera is difficult, potentially leading to confusion about the identification of species, with the consequence that our knowledge about the ecological preferences of certain species might be uncertain (Kahlert et al., 2019).

We analysed a) the geographical distribution, b) co-occurrence and c) ecological profiles of the ASVs, d) discussed the results in the light of the species concepts, and e) discussed possible underlying causes of the co-occurrence of taxa. We tested the following hypotheses:

(1) the distribution of diatom taxa, here studied as ASVs, is geographically restricted
(2) co-occurring diatom taxa, here studied as ASVs, have different ecological preferences

Hypothesis (2) is potentially wide-ranging, and our tests are limited to different preferences with respect to key water quality variables. These, however, are important gradients from the point of view of ecological assessment, and this paper offers an indication of the likely scale of practical improvements if the *Fragilaria/Ulnaria* taxonomy is resolved in an unbiased way using ASVs.

To our knowledge, we are the first to use ASV units in combination with a curated taxonomy to retrieve co-occurrence, specific ecological profiles and large-scale geographical distribution for unique identifiers not biased by the quality of reference databases, clustering methods, or non-harmonized morphological identification.

2. Methods

2.1. Study sites

The metabarcoding dataset with accompanying metadata used in the present study was compiled with data from different European countries: France, Sweden, Finland, Norway, Iceland, Germany and the United Kingdom (UK) (Fig. 1). This allowed us to take advantage of existing metabarcoding datasets collected as part of different environmental assessment projects that included linked environmental data (Table 1). The final dataset included 770 samples from streams and lakes where at least one of the target taxa of *Fragilaria* or *Ulnaria* was found (Table 1).

The French samples were collected in 2016–2017 from sites included in the French river monitoring, see Rivera et al. (2020) for details. 422 of the originally 447 sampling sites contained at least one of our target taxa and were thus included in the present study. The samples from the

Nordic countries (Sweden, Finland, Norway and Iceland) were collected from streams and lakes in these countries for different projects between 2006 and 2017, and were studied for a doctoral project (Bailet, 2021). The Nordic dataset includes temporal replicates of all Swedish and some of the Finnish and Norwegian stream sites, and spatial replicates of the Finnish lake sites. Of the original 181 samples, 160 samples included at least one of the target taxa and were used in our study. The German dataset was collected in 2016–2018 for a research project (German Barcode of Life 2 Project funded by Bundesministerium für Bildung und Forschung Germany [01LI1501E]). Eight lakes and one connecting channel were sampled on multiple occasions, resulting in 43 samples, of which 36 contained our target taxa and were included in the analysis. 'Fishing' for target taxa (Scherz et al., 2020) was done in 171 sampling sites from the UK river monitoring program, resulting in 152 samples for the present study.

Environmental data were available for all lake samples and most of the stream samples. We compiled and harmonized to the same measurement unit available data for pH, alkalinity, total nitrogen (TotN), total phosphorus (TotP), conductivity, and total organic carbon (TOC). In some cases, alkalinity was derived from measured calcium data (France, UK, Norway, one German site). For the UK, we derived approximate values for TotP and TotN from the available data on orthophosphate (assessed as Total reactive phosphorus) and Total oxidized nitrogen. The TotP value was derived using the relationship between TotP and orthophosphate from a Swedish dataset with >1000 available pairs of TotP and orthophosphate. The results obtained in this way were validated by testing whether the TN:TP ratios of the UK data were about the same as for other countries included in this study, and also by comparing with the equivalent equation derived from the French data. Most data were available for nutrients (TotN 663, TotP 665 measurements), less for the other variables (527 for TOC, 486 for pH and alkalinity each, 485 for conductivity). Not all environmental data were available for all samples, depending on the sampling program. The Nordic dataset was most complete, mostly with an annual mean for each variable for every sample, with a few exceptions. The UK dataset also mostly had an annual mean for each variable and site, with the exception of TOC. The French dataset was more complete regarding variables for some sites than others, with best coverage for nutrients. For the German lakes, detailed environmental data were available from samples collected from the centre of the lake, and annual mean values were calculated for each lake. However, for spatial replicates, local environmental measurements were only available for pH and conductivity, with one spot value for each taken on each occasion that diatoms were sampled. Geographic coordinates were available for all samples except for the UK. In general, our final dataset was biased towards French and UK sites, stream sites, and sites with relatively high pH and alkalinity (Supplements Fig. S20).

Details with respect to the metabarcoding methods for the different sub-datasets can be found in the original studies (Bailet, 2021; Kelly et al., 2018; Rivera et al., 2020, German Barcode of Life 2 Project funded by Bundesministerium für Bildung und Forschung Germany [01LI1501E]). In short, the French, German and Nordic countries metabarcoding was performed with the diatom-targeted primers of Vasselon et al. (2017), focusing on the short fragment of the *rbc*L plastid gene of 263 bp (312 bp with primers). The UK short barcode differs slightly from the *rbc*L barcode designed by Vasselon et al. (2017). It covers a similar region of the *rbc*L gene but is slightly longer (331 bp without primers, Kelly et al., 2018, Supplements Fig. S21). The sequencing technique was the same for the French, German and UK dataset (Illumina sequencing (MiSeq)). The Nordic countries' dataset was generated using PGM (Ion Torrent) sequencing.

2.2. Description of the studied Fragilaria and Ulnaria taxa

The entire length of the *rbc*L chloroplast gene marker of 1440 base-pairs (bp) generated with Sanger sequencing was used to establish a

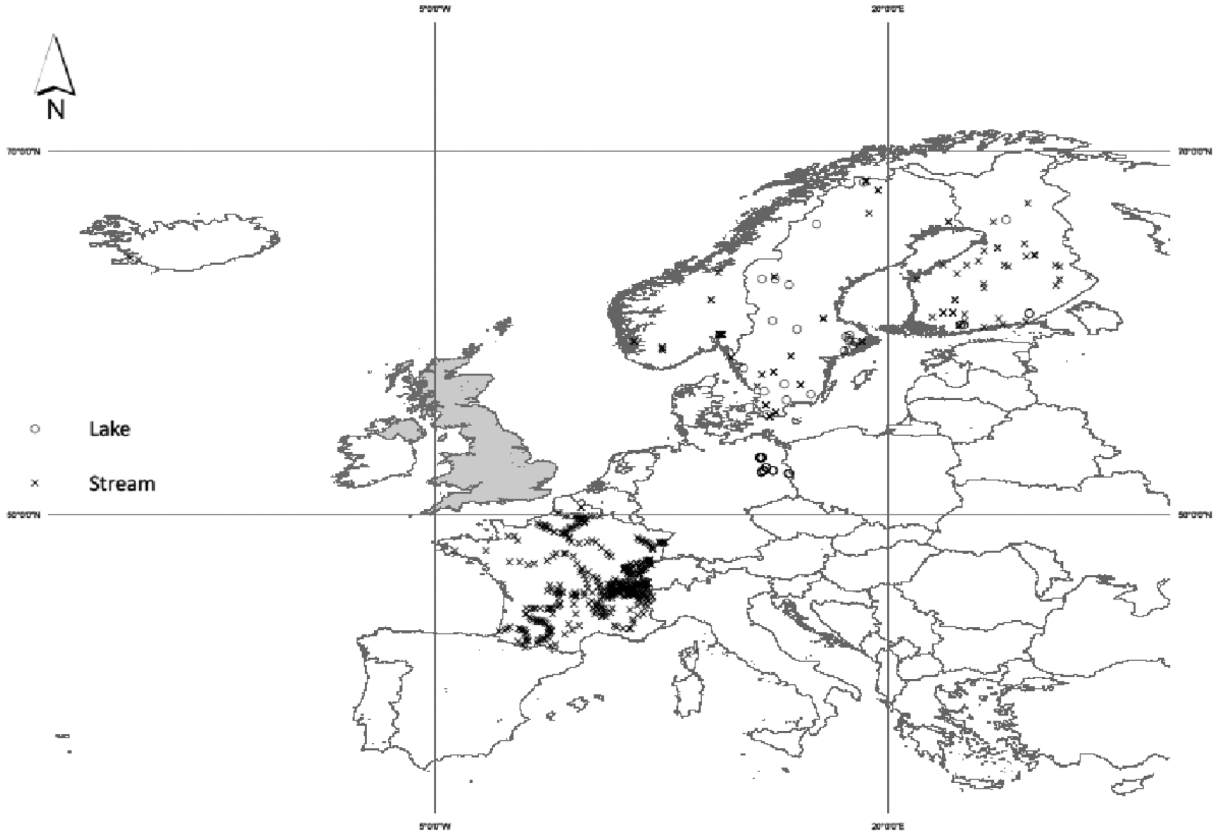


Fig. 1. Sampling sites. 643 streams, 45 lakes. (Sampling coordinates in UK not available). x stream, ○ lake.

Table 1
Dataset used for analysis of the geographical distribution and ecological preferences of unique molecular identifiers of the target taxa of *Fragilaria* or *Ulnaria*. The table shows the number of sampled sites, the total number of samples, and how many of them had environmental variables available. The information is given for streams and lakes separately. The number of samples is higher than of sites because of *temporal and **spatial replicates for some of the sites.

	streams			lakes		
	sites	total number of samples	number of samples with environmental variables available	sites	total number of samples	number of samples with environmental variables available
France	422	422	357			
Sweden	17	46*	46*	29	29	29
Finland	33	44*	44*	7	20**	20**
Norway	17	19*	19*			
Iceland	2	2	2			
Germany				9	36**	36**
United Kingdom	152	152	140			

phylogenetic tree for *Fragilaria* and *Ulnaria,* placing 194 cultured strains into clades and comparing the results with the morphological species concept. Some species defined based on morphological criteria could be confirmed using the *rbc*L marker, e.g. *F. gracilis* Østrup, *F. tenera* (W. Sm.) Lange-Bert., *F. perminuta* (Grunow) Lange-Bert. and *F. tenuistriata* Østrup. Other well-defined molecular clades did not match any currently described *Fragilaria* species. To clarify recognition of these taxa, three new species were described: *Fragilaria agnesiae* Kahlert & Rimet, *Fragilaria heatherae* Kahlert & M.G.Kelly, *and Fragilaria joachimii* Kahlert. Several other morphological taxa, mainly representing species with unclear morphological identification and a high length to width ratio, could not be resolved with the *rbc*L barcode. The genus *Ulnaria* was used as outgroup in this study. Most of the clearly separated molecular clades (= "species") identified in Kahlert et al. (2019) included several genetic variants of the long *rbc*L marker.

Metabarcoding datasets are compiled lists of reads of short genetic markers, termed 'barcodes'. For the present study, we matched this short barcode in form of an ASV to the species name used in Kahlert et al. (2019). We have avoided the term "morphospecies" in this context as this term does not match all of their delimitated species. For this, we first prepared a reference database for the studied group (*Fragilaria* and *Ulnaria*) using the strains included in Kahlert et al. (2019). The reference database linked each taxon to the two short *rbc*L barcode sequences (one of 263 bp for Germany, France and the Nordic countries; and one of 331 bp for UK). The resulting reference database contained 21 different ASVs which were used to 'fish' in the metabarcoding datasets (Table 2). Due to the longer UK barcode, three taxa were represented by two different ASVs (Table 2), which were pooled for the common analysis of all datasets. Because short barcodes are not always sufficient to discriminate taxa (Porter and Hajibabaei, 2020), some ASVs in our study represented clusters of species, while in other cases several ASVs represented a single species, with the genetic variation of this species preserved within the short barcode region (Kahlert et al., 2019). In these latter cases, it is possible to evaluate whether these genetic variants within a species have different geographical distributions and ecological preferences. In cases where ASVs are shared by several species, additional molecular identifiers are needed to separate species-specific distribution and ecology. However, it is still possible to use these ASVs as unique identifiers for biodiversity analyses, including the assessment of richness and ASV-specific ecological preferences and geographical distribution (Porter and Hajibabaei, 2020).

As we were able to differentiate three distinct genetic variants of *Fragilaria gracilis* using ASVs (Table 2), it was possible to study the geographical distribution and ecological preferences of genetic variants within this species. Kahlert et al. (2019) suggested that those ASVs belong to two cryptic species within *F. gracilis*. The FGRA3 ASV represents the *F. gracilis* clade with a higher striae density (Ø 22 striae/10 µm) and a more lanceolate form of the valve than the second clade included in Kahlert et al. (2019, Ø 20 striae/10 µm, valves more linear). This second clade is represented here by the FGRA1 and FGRA2 ASVs,

Table 2
Unique identifiers (ASVs) used in the present study to explore geographical distribution and ecological preferences, and their assignment to taxonomy.

Number of strains in cluster[1]	taxa ID	species names[2]
9	FGRA1	*Fragilaria gracilis*
16	FGRA2	*Fragilaria gracilis; Fragilaria* cf. *gracilis*
53	FGRA3	*Fragilaria gracilis*
7	FPEM1	*Fragilaria perminuta*
3	FPEM2	*Fragilaria perminuta*
7	FJOA	*Fragilaria joachimii*
3	FHEA1	*Fragilaria heatherae*
1	FHEA2	*Fragilaria heatherae*
9	FAGN[3]	*Fragilaria agnesiae; Fragilaria* sp. 1
1	FBID	*Fragilaria bidens*
1	FNNO	*Fragilaria nanoides*
1	FPRU	*Fragilaria pararumpens*
1	FRAS	*Fragilaria* sp.
11	FTEN	*Fragilaria tenera; Fragilaria tenuistriata; Fragilaria* sp.
30	FMIX[3]	*Fragilaria pararumpens; Fragilaria* cf. *pararumpens; Fragilaria nanoides; Fragilaria* cf. *nanoides; Fragilaria mesolepta; Centronella reicheltii; Fragilaria crotonensis; Fragilaria* cf. *capucina; Fragilaria* cf. *rumpens; Fragilaria* sp.
5	UACU	*Ulnaria acus*
5	ULNS	*Ulnaria ulna; Ulnaria acus*
1	UULN1	*Ulnaria ulna*
1	UULN2	*Ulnaria ulna*
17	UULN3[3]	*Ulnaria ulna*
1	UULN4	*Ulnaria ulna*

[1] of 194 included in Kahlert et al. (2019).
[2] following Kahlert et al. (2019).
[3] due to the differences in length between the two barcodes, these taxa were represented by two different ASVs, which were pooled for the common analysis of all datasets.

representing populations following Kahlert et al. (2019). It was also possible to evaluate the geographical distribution and ecological preferences of the three newly described species, *F. agnesiae, F. heatherae,* and *F. joachimii*, with one reservation. Whereas the ASVs of *F. heatherae* and *F joachimii* were different and included only the target species, the ASV of *F. agnesiae* included also an unknown *Fragilaria* species named *F.* sp. 1 in Kahlert et al. (2019) (Table 2). On the other hand, it was possible to study two genetic variants within *F. heatherae,* represented by two different ASVs in our dataset (Table 2). In addition to those species discussed in Kahlert et al. (2019), we were able to study 14 ASVs of the *Fragilaria/Ulnaria* genera, some of which included only one taxon, whilst others represented a mixture of taxa (see Table 2).

Fishing for the different ASVs was performed by the holder of each of the separate metabarcoding datasets. The dataset holder searched for exact matches of the ASVs to all available reads without rarefying. For the French data, the MiSeq paired-end reads were demultiplexed by the

sequencing platform and assembled into full-length DNA sequences. The minimum bioinformatic sanity checks were then applied to ensure that the maximum information was kept. Using Mothur v.1.44.2, reads were filtered to keep only the sequences with the exact target length (312 bp, i.e. 263 bp barcode + primers) and with zero ambiguities ("N"). Finally, DNA reads were dereplicated and singletons and primers were removed. The same method was used for the Nordic dataset, except that PGM reads do not need to be assembled. The same method and settings were also used for the German and the UK datasets, except that the pipeline MetBAN was used for the German data (German Barcode of Life 2 Project funded by Bundesministerium für Bildung und Forschung Germany [01LI1501E]), and dada2 for the UK data (Callahan et al., 2016). Datasets were then matched against the *Fragilaria/Ulnaria* ASV reference database using exact matching (i.e. 100% identity) in order to extract the environmental sequences corresponding to *Fragilaria* and *Ulnaria* taxa. Because the UK uses a longer barcode, the fishing was done for all 24 unique ASVs found for the longer barcode (Table 2), but where this detected two ASVs corresponding to only one in the other datasets, these were pooled to enable a direct comparison. To ensure a robust analysis when merging the data from all countries into one dataset, we decided to use presence/absence (PA) data only.

2.3. Statistical methods

To assess whether the taxa were geographically restricted, we visually inspected ASV distribution maps, and additionally compared the ASV composition between countries. For the latter, we used the nonparametric Kruskal-Wallis test with subsequent post-hoc pairwise Mann-Whitney tests and Bonferroni corrected p-values to analyse whether richness per sample differed between countries. We also used ANOSIM (ANalysis Of Similarities) with pairwise ANOSIMs between all pairs of groups and Bonferroni corrected p-values as a post-hoc test to analyse whether ASV composition differed between countries. We used SIMPER (Similarity Percentage, Clarke, 1993) to assess which taxa were responsible for the differences. ASV presence/absence data (PA) were used, along with the Bray-Curtis distance measure. All analyses were performed with the software PAST v.2.17b (Hammer et al., 2001).

We used the R package Cooccur (Griffith et al., 2016) *to analyse whether ASVs co-occurred non-randomly*. The method applies the probabilistic model of species co-occurrence (Veech, 2013) to a set of taxa distributed among a set of survey or sampling sites. The algorithm calculates the observed and expected frequencies of co-occurrence between each pair of taxa. The expected frequency is based on the distribution of each taxon if it occurred randomly and independent of the other species. The analysis returns the probabilities that a more extreme (either low or high) value of co-occurrence could have been obtained by chance. We used the default settings, which include the removal of those taxa pair combinations which are expected to co-occur at less than one site. We tested all possible ASV pairs (i.e. also *Fragilaria/Ulnaria* combinations).

To assess the overall structure of our data, especially to see if certain ASVs cluster together at sites with similar ecology, we used NMDS with PCOrd v.6.08 (McCune and Mefford, 2011) and Bray-Curtis distance measure on PA data. Beta diversity was 3.32 which should present a medium challenge for ordination analyses (McCune and Grace, 2002). The coefficient of variation of the taxa column totals, reflecting the variability among columns, was 100%, indicating moderate to high variability. As the average skewness of taxa was 3.164, we used a non-parametric ordination which is well suited to non-normal data (McCune and Grace, 2002). The test for outliers used a cut-off of 2.0 standard deviations from the grand mean and indicated that no taxa or sites should be treated as outliers. For the NMDS, we first ran the autopilot mode with PCOrd's default settings, and then reran the NMDS with the starting configuration that was most stable. PCOrd v.6.08 returns the NMDS results with the axes ordered in falling order of importance.

To model the ecological preferences of the ASVs, we used the averaging method described in Kahlert et al. (2021), during which we modelled optimum and tolerance values along all environmental factor gradients for PA data. Means and standard deviations were calculated using the environmental values of samples where a given ASV was present. All environmental variables except pH were logarithmically transformed. We used the nonparametric Kruskal-Wallis test with subsequent post-hoc pairwise Mann-Whitney tests and Bonferroni corrected p values with PAST v.2.17b (Hammer et al., 2001) to assess whether the environmental medians of ASV occurrence were significantly different. SIMPER and ANOSIM were used to assess the ecological preference of the ASVs for lakes or streams. Indicator species analysis (IndVal) implemented in PCORD v.6.08 (McCune and Mefford, 2011) was used to test whether certain ASVs could be defined as significant indicators for extreme conditions (represented in our dataset by the values at the edges of the environmental range). We tested pH and TotP (logarithmic transformed) as examples of environmental factors known to be important drivers of diatom assemblage structure (e.g. Pajunen et al., 2020; Soininen et al., 2016). The IndVal method gives an indicator value for each taxon that weights its presence in certain environmental categories versus the presence across all sites. For pH, we used the categories pH < 5; 5–6; 6–7; 7–8 and > 8, for TotP, we used < 10; 10–100; 100–1000 and > 1000 µg/l. The statistical significance of each indicator value was tested by a Monte Carlo permutation test with 4999 randomizations. A 'significant indicator' defined by this method describes a taxon with a high and statistically significant indicator value. In our study, an indicator ASV was defined as an ASV typical of very low or high values of pH or TotP. We additionally assessed the environmental conditions, represented by pH and TotP, where the greatest richness of the target taxa ASVs can be expected, using species packing with PAST v.2.17b (Hammer et al., 2001) and the number of ASVs found at each site. The method fits Gaussian response models to taxa data along a gradient. The fitted parameters are optimum (average), tolerance (standard deviation) and maximum. The algorithm is based on weighted averaging according to ter Braak and Van Dam (1989).

3. Result

3.1. Geographical distribution of target ASVs of Fragilaria and Ulnaria

Of the 21 ASVs, 19 were found in the different databases. Only two, FRAS (*Fragilaria* sp.) and UULN4 (*U. ulna*), were not found at all (Table 3). Overall, most (18) of the ASVs were found in the French dataset, the largest of those available to us (Table 3). The average number of ASVs per sample for each country was 3.6–4.7, with the UK having the lowest number (Table 3, significantly lower than for France and Finland (Kruskal-Wallis test, $p < 0.05$)). Visual inspection of the distribution maps showed that many ASVs were spread over all sampled geographical regions (FGRA1, FGRA2, FGRA3, FAGN, FPRU, FTEN, FMIX, UACU, ULNS, UULN3, Supplements Figs. S1-19). Certain ASVs, however, seemed to be more restricted to certain regions (FPEM1, FPEM2, FHEA1, FBID, FNNO, UULN1, Supplements Figs. 1-19). ASV composition was significantly different between countries (ANOSIM, $p < 0.005$). Assemblages in France, Finland and Germany were very different to those from Norway (ANOSIM, $R > 0.4$, pairwise ANOSIM post-hoc tests, $p < 0.005$). Equally different were assemblages of Iceland and Germany. Five ASVs were responsible for more than half of the dissimilarities between countries: FGRA2, followed by FMIX, FAGN, UACU and UULN3 (Table 3). The large deviation was mainly due to fewer records of UACU (*U. acus*) and UULN3 (*U. ulna*) in Norway compared to France, Finland and Germany (Table 3). Some ASVs were not found in certain countries at all (Table 3).

3.2. Co-occurrence of target ASVs

The analysis of co-occurrences showed that overall, most ASVs occurred more often together than would be expected by chance (Fig. 2). Of the 171 possible ASV pair combinations, 24 pairs were removed by

Table 3
ASVs found with exact matching, and differences between countries. The first rows of the table show the differences in number of samples and found ASVs per country. The following rows show which ASVs were primarily responsible for ASV composition differences between countries (SIMPER analysis on PA data, Bray-Curtis similarity). The number of occurrences of the *Fragilaria* & *Ulnaria* ASVs for all 770 samples is shown in the first column. The ASVs are sorted in descending order of contribution the differences between country, given in the 2 subsequent columns (ASV-specific, and cumulative). The last columns show the detailed country differences as average presence of an ASV per sample per country (e.g. 1: present in all samples of a country, 0.5: present in 50% of all samples of a country).

			Country	FR	SE	FI	NO	IS	DE	UK
			Number of samples	422	75	64	19	2	36	152
			Total number of ASVs	18	13	16	13	7	11	15
			Average number of ASVs present per sample	4.6	4.4	4.7	4.6	4.5	4.3	3.6
				Average presence of ASV in a sample						
ASV taxa IDta ID	Total number of occurrences in dataset	Contribution to dissimilarity%	Cumulative contribution %	FR	SE	FI	NO	IS	DE	UK
FGRA2	383	11.11	11.11	0.40	0.60	0.75	0.53	0.5	0.56	0.59
FMIX	420	11.05	22.15	0.56	0.60	0.81	0.53	0.5	0.89	0.30
FAGN	355	10.37	32.52	0.51	0.53	0.67	0.37	1	0.19	0.27
UACU	321	10.22	42.75	0.55	0.17	0.41	0.16	0	0.58	0.18
UULN3	626	10.13	52.88	0.96	0.52	0.81	0.47	1	0.67	0.63
FGRA1	257	8.322	61.2	0.42	0.28	0.09	0.16	0	0.31	0.24
ULNS	204	8.235	69.43	0.20	0.35	0.19	0.58	0	0.08	0.44
FJOA	205	7.743	77.18	0.24	0.29	0.42	0.32	0.5	0	0.31
FTEN	178	6.503	83.68	0.20	0.40	0.33	0.26	0.5	0.39	0.16
FGRA3	117	4.824	88.51	0.14	0.21	0.05	0.53	0	0	0.19
FPRU	73	3.059	91.56	0.09	0.19	0.02	0.42	0	0.08	0.05
FHEA2	83	2.912	94.48	0.11	0.19	0.03	0.11	0.5	0	0.11
FPEM1	41	2.254	96.73	0.02	0.04	0.06	0	0	0.44	0.05
FPEM2	32	1.195	97.92	0.06	0	0.03	0.16	0	0.06	0
FBID	23	0.7277	98.65	0.05	0	0	0	0	0	0.01
FHEA1	7	0.4088	99.06	0	0	0	0	0	0	0.05
UULN2	10	0.3428	99.4	0.02	0	0.02	0	0	0	0
UULN1	12	0.3382	99.74	0.03	0	0	0	0	0	0
FNNO	9	0.2578	100	0.02	0	0.03	0	0	0	0
FRAS	Not found									
UULN4	Not found									

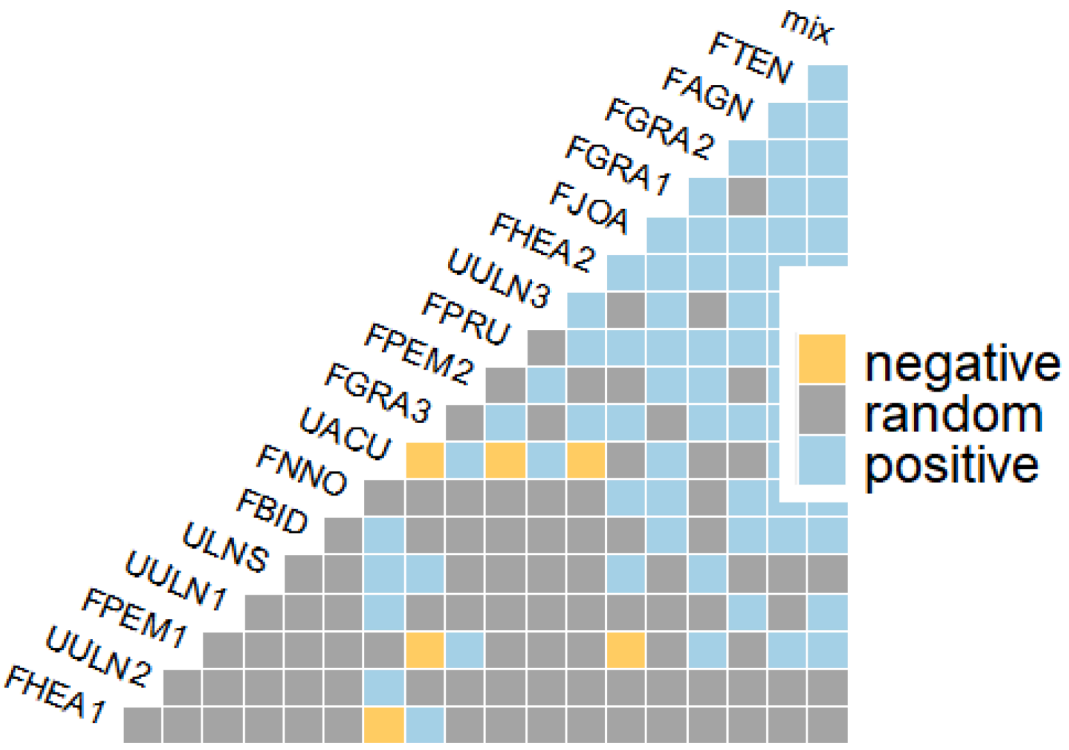


Fig. 2. Positive and negative ASV associations of the target taxa of *Fragilaria* and *Ulnaria* determined by the probabilistic co-occurrence model. ASV identifiers are positioned to indicate the columns and rows that represent their pairwise relationships with other ASVs. The most positive relationships are plotted in the upper right field of the graph.

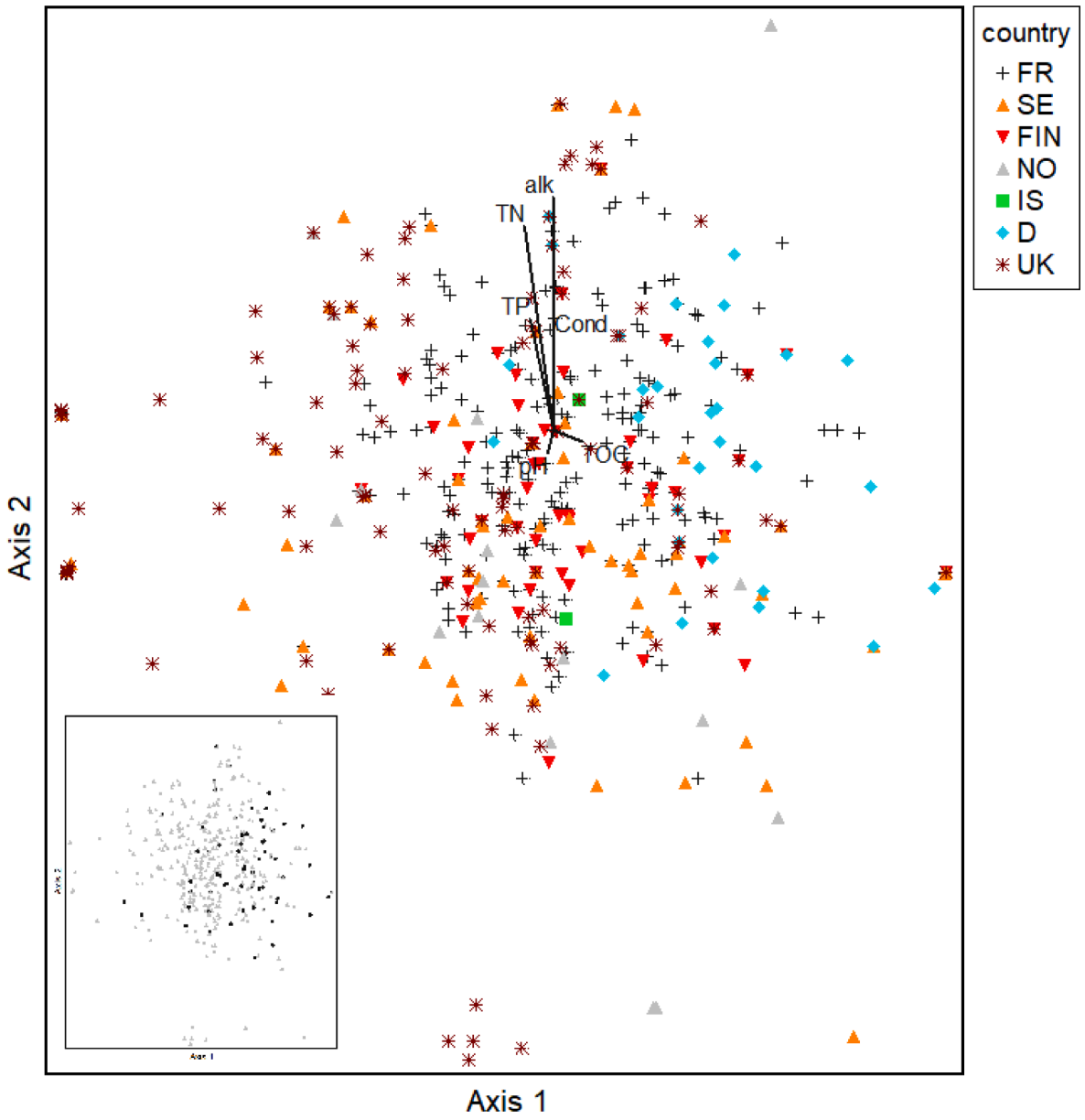


Fig. 3. Differences between sites in the ASV composition of the target diatom taxa of *Fragilaria* and *Ulnaria* (NMDS on presence-absence data, Bray-Curtis similarity). Sites classified by country. Habitat type shown in inset (grey: streams, black: lakes). Correlation of environmental variables to first and second axes added with vector length upscaled 400% for visibility. Find ASV scores in Supplements Fig. S22.

default prior to analysis because their expected co-occurrence was < 1, i.e. in less than one site. Of the 147 remaining pairs, 78 co-occurred significantly more often than randomly expected, whilst only six pairs occurred less often (Fig. 2).

3.3. Ecological conditions under which the target ASVs were present

No clear pattern of ASV distribution was found with the ordination analysis. No clusters of sites, ASVs, or a separate lake or stream cluster were found, and environmental variables were only weakly correlated with the ordination. Although caution is needed when interpreting NMDS graphs based on weak ordinations, Figs. 3 and S22 give some insights into how the ASVs and the environmental variables were correlated to the axes. The environmental variables were weakly correlated to the second ordination axis, where nutrients, alkalinity and conductivity showed positive relationships (Fig. 3, Pearson's correlation coefficient for alkalinity: 0.3, TotN: 0.3, TotP: 0.2, conductivity: 0.2). Habitat (i.e. stream or lake), was also weakly correlated with the NMDS axes, with lake sites mainly found positively related to the first and negatively to the second axis (Fig. 3). The ASVs ULNS, FJOA and FMIX showed the closest associations with the first axis, with ULNS and FJOA showing a negative and FMIX a positive association (Supplements Fig. S22). The ASVs with the closest association to axis 2, and thus the local water chemistry variables, were FGRA3, FAGN and FMIX, all with negative associations. Consequently, they were more likely to be found in nutrient poor and low alkalinity sites. By contrast, UULN3 was strongly positively associated with axis 2, indicating a preference for nutrient rich sites with high alkalinity (Supplements Fig. S22).

The subtle differences in ASV composition between lake and stream sites found in the ordination were confirmed by SIMPER (Table 4). The average occurrence of most ASVs was significantly different for lakes and streams (Table 4, ANOSIM, $R = 0.06$, $p < 0.005$). For example, FGRA3 and FHEA2 were more than four times more likely to be found in stream than in lake samples, whereas FPEM1, FTEN, and FMIX were

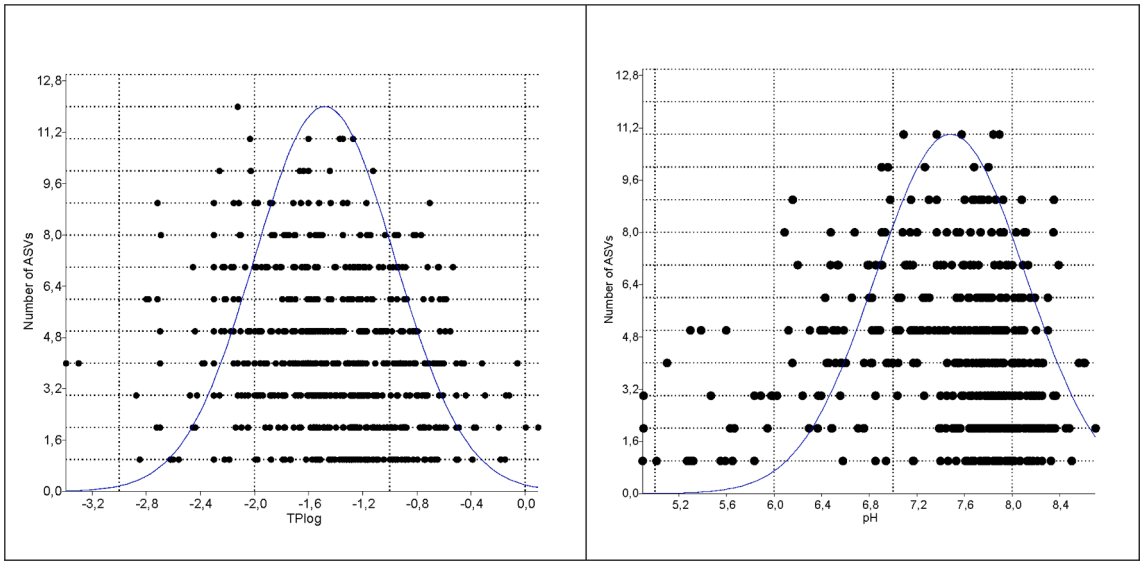


Fig. 4. Number of ASVs of *Fragilaria* and *Ulnaria* per sample versus $TotP_{log}$ [mg/l] (left) and pH (right). -with) - Gaussian response model fitted.

more often found in lakes (Table 4).

3.4. Ecological preferences

3.4.1. Ecological preferences in general

We found up to twelve *Fragilaria* and *Ulnaria* ASVs together at each sampling site. The highest richness was on average found at 0.033 mg/l TotP and pH 7.5, and lowest at the marginal environmental values (Fig. 4). IndVal analysis revealed, however, that no ASVs were typical for marginal conditions (data not shown). For example, only six out of the 19 ASVs were not found in the sites pH < 6, or in the sites with < 0.005 mg TotP/l.

The ecological preferences modelled for the separate ASVs all overlapped, with gradual shifts of the modelled optima. The large standard deviations indicated that most ASVs were found across large ranges, from low to high values of the respective environmental variables (Fig. 5, Table S1). The ASV-specific modelled optima were 7.15–7.87 for pH, 0.4–1.8 mekvl/l for alkalinity, 0.021–0.109 mg Tot-P/l, 0.6–2.0 mg Tot-N/l and 9–59 mS/cm for conductivity, respectively (for details see Fig. 5, Table S1). Although the modelled optima only showed small differences along the environmental gradients, the differences in distribution of the ASVs were still significant for all factors (Kruskal-Wallis test, $p < 0.05$). The pairwise post-hoc tests (Tables S2-S6) then specified the ASVs with significant different ecological preferences. Combining the results for ASV co-occurrence and ecological preferences, we found that two ASVs were involved in all of the few detected significantly negative co-occurrences, UACU and FPEM1 (Fig. 2). These were placed at or close to the ends of the range of the modelled nutrient preferences of all studied ASVs (Fig. 5, Table S1). The differences in ecological preferences were significant for all but one ASV pairs with a negative co-occurrence, (pairwise post-hoc tests, Tables S2-S6). Differences were found for all of the studied water chemistry variables, with three to all five variables (pH, alkalinity, TotP, TotN, conductivity) varying between ASV pairs.

3.4.2. Did co-occurring genetic variants show different ecological preferences?

We also analysed whether the ASVs showing the strongest positive associations had different ecological preferences. All of those ASVs analysed in detail occurred frequently in our dataset (Fig. 2; FTEN, FAGN, FGRA2, FGRA1, FJOA). We found significant different preferences for certain water chemistry variables for some of the ASV pairs. For example, FTEN preferred lower TotP concentrations than FAGN (Fig. 5, Table S1, Table 5). FJOA preferred higher TotP concentrations than FTEN, FGRA1 and FGRA2 (Fig. 5, Table S1, Table5). For other ASVs with strong positive association (e.g. FTEN-FGRA2, FAGN-FGRA2 or FAGN-FJOA) no significant differences of preferences for the studied water chemistry variables were found (Table 5).

3.4.3. Comparing ASV responses within species

Finally, we linked the geographical distributions and ecological preferences of ASVs to the traditional species concepts (following Kahlert et al., 2019). In summary, the results indicated that ASV variants with different optima existed for *F. gracilis*, whereas evidence for *F. perminuta* and *F. heatherae* ASV variants was weaker.

The species *F. gracilis* was represented by three genetic variants in our dataset, all of them occurring frequently. The ASV FGRA3 was found in higher frequency in Norway than FGRA1 and FGRA2 (Table 3). None of these ASVs showed a negative association with each other, but FGRA2 was significantly more associated with both of the other ASVs than expected by chance (Fig. 2). FGRA3 had very low modelled optima for all environmental variables assessed, and FGRA1 the highest values of the three ASVs (Fig. 5). The median distribution of the three ASVs was significantly different for all of the pairs for alkalinity. It was also different for TotN between FGRA3 and the other two ASVs, and for conductivity between FGRA2 and FGRA3 (Kruskal-Wallis test with posthoc pairwise comparisons, Bonferroni-corrected $p < 0.05$, Tables S1-S5). On the other hand, FGRA1 and FGRA3 were clearly more common in streams, FGRA2 was found with a slightly greater frequency in lakes (Table 4).

F. perminuta and *F. heatherae* were the other *Fragilaria* species in our dataset represented by more than one genetic variant. ASVs of these species were less frequent in our dataset than those of *F. gracilis*. Both *F. perminuta* variants were more frequent in lakes, whereas both *F. heatherae* variants were more frequent in streams (Table 4). The *F. perminuta* variants were significantly more often found together than expected by chance but the *F. heatherae* variants were not (Fig. 2). The geographical distribution of the within-species variants was different. FPEM1 was more frequently found in Sweden, Finland, Germany and UK, whereas FPEM2 was found more frequently in France and Norway (Table 3). FHEA1 was only found in UK, whereas FHEA2 had a wider distribution. Graphical inspection of the modelled optima did not reveal a closer relationship of the within-species ASVs than of the between-species ASVs (Fig. 5). The median distribution of the two within-species ASVs was not significantly different for either *F. perminuta* or *F. heatherae* for any of the environmental variables (Kruskal-Wallis test

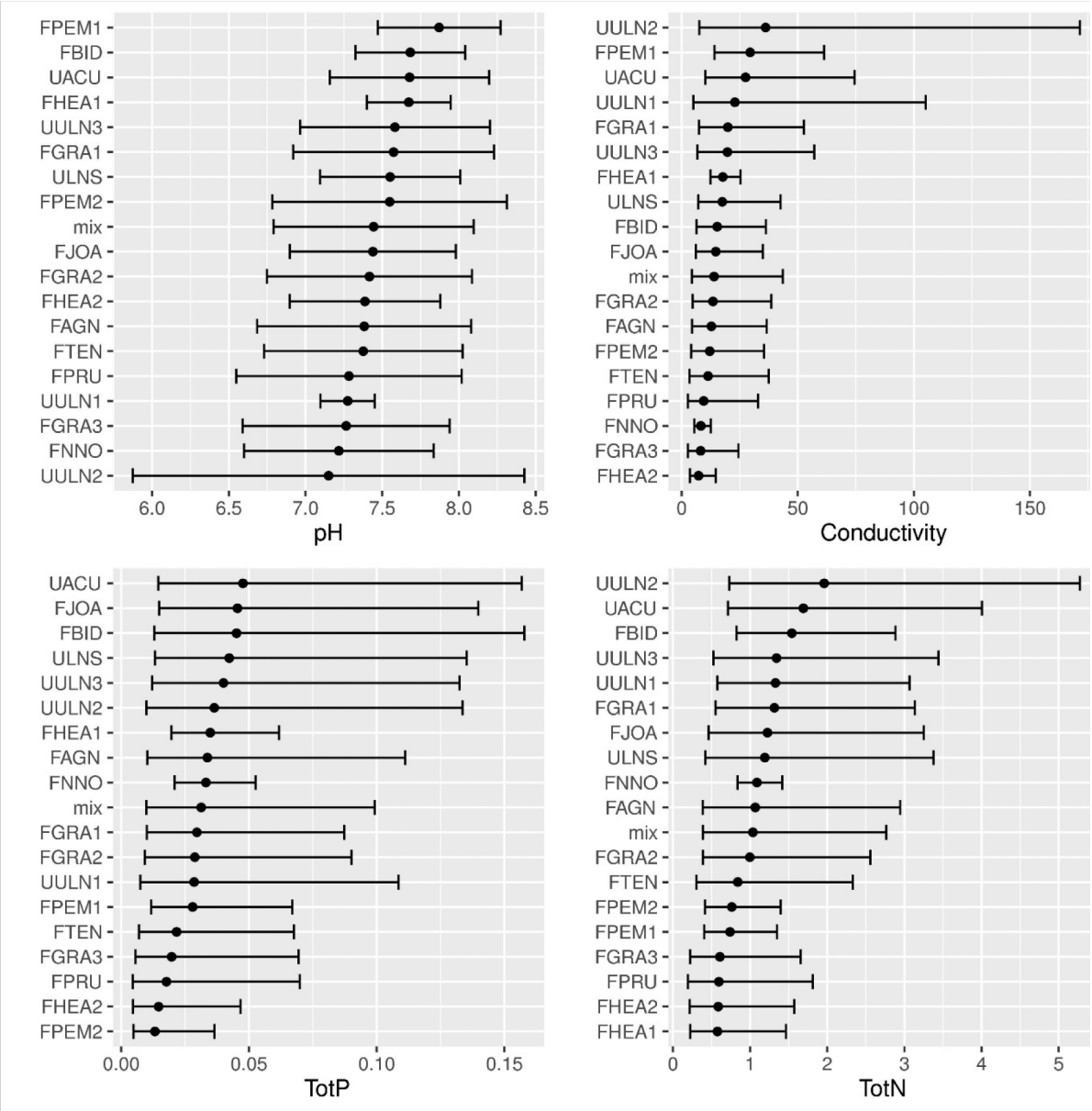


Fig. 5. Optima and tolerance (Standard deviation) of target *Fragilaria* and *Ulnaria* taxa for pH (upper left), conductivity (mS/m, upper right), TotP (mg/l, lower left), and TotN (mg/l, lower right).

with posthoc pairwise comparisons, Bonferroni corrected $p < 0.05$).

4. Discussion

In summary, we found some evidence suggesting that diatom taxa distributions, here studied as *Fragilaria* and *Ulnaria* ASVs, were restricted to certain regions. However, we did not find evidence that ASV distribution in general is geographically restricted. We cannot rule out the possibility that the co-occurrence of diatom taxa, as ASVs, might be enabled by slight differences in their ecological preferences. However, the detected differences in ecological preferences were small. Thus, neither of our hypotheses could be definitely confirmed based on our dataset of ASVs. Linking the studied ASVs to Linnaean species concepts, following Kahlert et al. (2019), indicated that the *Fragilaria* species included in our study were not restricted to certain regions, their genetic variants (ASVs) were no isolated populations, and that differences in ecological preferences among ASVs were larger than those among species.

More specifically, we did not find evidence for a general restricted dispersal of the studied *Fragilaria* and *Ulnaria* ASVs (hypothesis 1). While some ASVs had a narrow distribution, others were widely distributed. Similar results were recently found for other diatom genera (Pérez-Burillo et al., 2021), suggesting that there are no general geographical barriers for the distribution of freshwater benthic diatom ASVs in Europe, even between regions separated by the sea. For example, FBID was originally isolated from Japan, but found in France and the UK. FHEA2, isolated from Italy, was found in all countries except the locations sampled in Germany. The reasons why other ASVs (FHEA1, UULN1) had a restricted distribution (UK, France) will need further study. Maybe, those sequences are evolutionary relatively young and might not have had time to spread to a larger area, confirming the idea that diatom dispersal is not as rapid and effective as required for an ubiquitous dispersal (Poulíčková et al., 2008, and reference therein). Perhaps also the non-detection of FRAS and UULN4 could be explained by a restricted geographical distribution close to the regions of strain isolation, neither of which was included in our study (Spain, USA,

Table 4
ASVs found with exact matching, differences between habitats (streams versus lakes). Table shows the mean of the occurrences per sample for streams or lakes (SIMPER analysis on PA data, BC).

Taxon	Contribution to dissimilarity %	Cumulative contribution %	Streams (n = 685)	Lakes (n = 85)	Streams: lakes
FMIX	12.1	12.1	0.51	0.86	0.6
FGRA2	11.1	23.2	0.48	0.64	0.8
UULN3	10.8	34.0	0.84	0.59	1.4
FAGN	10.4	44.4	0.46	0.45	1.0
UACU	10.3	54.7	0.41	0.45	0.9
FTEN	8.7	63.4	0.21	0.44	0.5
FGRA1	7.5	70.9	0.36	0.13	2.8
ULNS	6.8	77.6	0.28	0.15	1.8
FJOA	6.1	83.7	0.29	0.09	3.1
FPEM1	5.2	88.9	0.03	0.27	0.1
FGRA3	3.4	92.3	0.17	0.04	4.7
FPRU	2.8	95.1	0.10	0.07	1.4
FHEA2	2.1	97.2	0.12	0.02	5.0
FPEM2	1.3	98.5	0.04	0.05	0.9
FBID	0.5	99.0	0.03	0	
FNNO	0.3	99.3	0.01	0.01	1.0
UULN1	0.3	99.6	0.02	0	
UULN2	0.2	99.8	0.01	0	
FHEA1	0.2	100.0	0.01	0	

Kahlert et al., 2019). In addition, the fact that the genetic variants within a species displayed a somewhat contrasting frequency in different subregions confirmed that ASV distribution is not ubiquitous. However, we found no clear evidence that the studied within-species ASVs of *F. gracilis*, *F. perminuta* and *F. heatherae* were isolated populations. Within-species ASVs did not show a significant negative association; rather, in some cases they co-occurred together more often than expected by chance.

We also found that a sample often hosted many of the studied *Fragilaria* and *Ulnaria* ASVs. Up to 12 out of the 19 ASVs were found at the same site. More than half of the ASV pairs co-occurred significantly more often together than would be expected by chance. The extent of these co-occurrences suggested that the taxa were not actively excluding each other per se. Only a few ASVs showed a significant negative co-occurrence, involving especially UACU and FPEM1. We found that the ecological optima of those two ASVs were placed at the ends of the range of the modelled optima and habitat preferences of all studied ASVs, meaning that differences in their ecological profiles were probably sufficient to prevent co-occurrence at sites with similar ecology altogether.

There are a couple of explanations for the co-occurrence of ASV pairs detected amongst our target taxa. *First*, we found some evidence that the ecological niches of the studied taxa (as ASVs), described as their preference for important water chemistry variables, and their preference for lakes or streams as habitat, were slightly, but significantly, different (hypothesis 2). This new knowledge, based on stable taxonomical units, give a first glimpse into ecological differences of a diatom group where separation has been difficult until now. We are aware that our results need to be interpreted cautiously. Overall, we found that the spatial pattern of ASV distribution was weak, so any attempts to explain the observed patterns need to be made with care. Furthermore, the modelled ecological preferences for the environmental variables showed only gradual shifts between ASVs when in rank order of their sensitivity to a variable. On the other hand, we observed significant differences in ecological preferences from the ASV pairs with the strongest positive relationships, all of which occurred frequently in our study. Additionally, we based our model on strong gradients known to influence diatom assemblage structure (Juggins and Birks, 2012). It is unlikely that combining several different datasets caused major batch effects as discussed by (Bailet et al., 2020). ASV data did not cluster per country, and we did not find outliers or certain country deviations when testing correlations of water chemistry variables prior to the analyses. Therefore, we believe that our findings of slight differences in the preference for important water chemistry variables indeed may indicate that ASVs might partly co-occur due to subtle differences in use of resources or other dimensions of the ecological niches as hypothesized in community ecology.

Several other causes for the observed frequent detected ASV co-occurrence are possible. One important factor is time. Not only does dispersal take time (as mentioned above), but competition and evolution take place over time, and environmental conditions are constantly changing too. This means that there is never a stable equilibrium for long enough for a taxon to exclude a co-occurring taxon purely through competition. The observed composition of taxa at any point in time is just a snapshot from an ever-changing continuum of possibilities. Our data also do not allow insights into the seasonal succession of taxa or their activity. Taxa might compete less and thus co-occur when their life cycles are not synchronous, as has been shown for e.g. the marine planktonic genus *Pseudo-Nitzschia* (Ruggiero et al., 2015; Turk Dermastia et al., 2020). At least in theory, it is possible that their abundances would peak at different times (due to yearly variation in temperature, light availability or seasonal dynamics of grazer populations), but their growth seasons may still overlap. Moreover, metabarcoding data do not provide information about whether the observed populations were active at the time of sampling. Resting cells will appear along with actively dividing ones, so it is possible that one of the co-occurring taxa was completely passive at the time of sampling.

It is also possible that ASVs were found side-by-side because the genetic variants have yet not formed distinct ecological niche preferences. This study has focussed on a single marker, albeit one representing a key photosynthetic enzyme. Links between this gene and the environmental variables on which we focus are complex, so it does not necessarily follow that small differences in the optima associated with ASVs necessarily represent the outcome of evolutionary processes. We can for example not fully rule out the possibility that some of the studied ASVs may represent species other than those listed in Kahlert et al. (2019). The analysis of additional markers and mating experiments would be needed to support species delineation and, in turn, species-specific distribution and ecology.

Space is another important factor, which need to be taken into account when ASV co-occurrence at a site is observed. Resource availability, or the presence of stress, is patchy not only in time, but also in space. This patchiness can perpetuate co-occurrence even if species are competing, as shown by Mitchell et al. (2013). We are also aware that the standard spatial scale used for sampling in our study covers several

Table 5
Probability that the modelled optima of two coexisting ASVs of *Fragilaria* and *Ulnaria* were significantly different ($p < 0.05$). Results of the post-hoc pairwise Mann-Whitney tests with Bonferroni corrected p values (PAST v.2.17b (Hammer et al., 2001)).

	FTEN-FAGN	FTEN-FGRA2	FTEN-FGRA1	FTEN-FJOA	FAGN-FGRA2	FAGN-FJOA	FGRA2-FGRA1	FGRA2-FJOA	FGRA1-FJOA
pH	1	1	1	1	1	1	1	1	1
Alkalinity	1	1	**0,01156**	1	1	1	**0,002313**	1	0,1429
TotN	0,9687	1	**0,000201**	**0,02018**	1	1	0,0537	1	1
TotP	**0,02469**	1	1	**0,000003047**	1	1	1	**0,004822**	**0,0113**
conductivity	1	1	**0,02881**	1	1	1	0,1332	1	1

substrata within a transect of several metres. Thus, each sample might pool a number of ecological microniches differing in their ecological, chemical, physical, and biological properties, contributing to the apparent co-occurrence we observed. To detect competition, it will be necessary to study diatoms at smaller scales than is usual for environmental assessment (Mitchell et al., 2013; Pérez-Burillo et al., 2021),

Future studies need to look beyond local water chemistry or spatial distribution to fully understand a niche. Water temperature and hydrology also need to be taken into account, as well as biological factors. Certainly the ability to withstand parasites and diseases (e.g. Crawford et al., 1985), or grazers (Hillebrand et al., 2007; Worm et al., 2002) are traits which are important for diatom competition. Biotic interactions are, however, much less studied than geographic distribution and responses to physical and chemical variables. Furthermore, we do not know much about how the relationships with single stressors discussed here are influenced by the wide range of other natural and anthropogenic stressors encountered by phytobenthos assemblages, each experienced at different amplitudes and frequencies.

The significance of these results for environmental assessments in general, and specifically for those associated with the WFD, is not yet clear. Certainly, the results of ASV-specific sensitivity to water quality variables of the target genera *Fragilaria* and *Ulnaria* could be used to further develop environmental assessment based on metabarcoding. Most ASVs represented taxa with a pH optimum slightly above 7, and low nutrient concentrations probably consistent with at least "good ecological status" (Poikane et al., 2019). However, the distribution of some ASVs might extend into more nutrient rich conditions. Our data give clear support to the idea that ASVs as stable and unbiased molecular units should be the base for ecological studies or environmental monitoring using molecular data, as has been recommended by Pérez-Burillo et al. (2021). Our findings of significant differences in preferences for alkalinity, TotN and conductivity between FGRA1 and FGRA3 in *F. gracilis,* or the different habitat preferences of the two ASVs of *F. perminuta*, indicate that the seemingly broad ecological tolerances of some diatom species might be a result from overlapping preferences among genetic variants, or cryptic species. These more restricted preferences and distributions of the within-species genetic variants has been earlier suggested by Poulíčková et al. (2008), and was found for other diatom genera by Pérez-Burillo et al. (2021), Poulíčková et al. (2008); Tapolczai et al. (2019) and Tapolczai et al. (2021). Combining the ASV approach with the use of metabarcoding datasets is a promising way to develop ecological assessment using barcodes. When used with a reliable reference library, metabarcoding should remove the differences between laboratories that occur when light microscopy results are pooled. The diatom metabarcoding method has been benchmarked for its use in environmental assessment (Bailet et al., 2019; Charles et al., 2021; Kelly et al., 2020; Valentin et al., 2019) and a curated reference barcoding library (Diat.barcode, Rimet, 2020; Rimet et al., 2019) is available for this marker. Using ASVs to generate metrics is the next logical step (Kahlert et al., 2021), with the barcode libraries being used to assign taxonomy post hoc, to aid interpretation.

5. Conclusions

Our study confirmed that there seem to be no general geographical barriers for the distribution of freshwater benthic diatom ASVs in Europe, but that dispersal is not sufficiently rapid to hide historical events during the studied time scale. There was evidence that only large differences in preferences for the analysed water chemistry variables prevented the co-occurrence of *Fragilaria* and *Ulnaria* ASVs at the same sites. Instead, the studied ASVs occurred frequently together at a site, and even showed positive relationships. We found subtle differences in ecological preferences amongst some pairs, which suggests that their co-occurrence may be due to avoidance of direct competition. However, the great overlap in geographical distribution and preferences for ecological conditions suggests that factors other than physico-chemical conditions were at least partly responsible for the observed co-occurrences and high richness of ASVs found at many sites. Our results are based on a large dataset, however we acknowledge that there are still biases regarding water chemistry, habitat and regions, some of which were better represented than others. We also emphasise that our conclusions are based on ASVs defined by a relatively short barcode region, and that our knowledge on the molecular diversity of diatoms, and the genetic variation of species and clades, is still in its infancy. Our study is just a first step to understand the seeming paradox that closely-related freshwater benthic diatoms show sympatric distributions in apparent contradiction to ecological theory (e.g. competitive exclusion principle). New molecular data are generated en masse as we are writing, and thus we will have the possibility to study the patterns of distribution of molecular diatom units in more detail in the near future. More detailed data will be required in order to test whether diatom taxa are competing. These data must describe other dimensions of the ecological niche than preferences for commonly measured water chemistry variables or habitat, and must be acquired at ecologically-realistic scales.

CRediT authorship contribution statement

Maria Kahlert: Conceptualization, Methodology, Formal analysis, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Satu Maaria Karjalainen:** Investigation, Writing – review & editing. **Francois Keck:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Martyn Kelly:** Conceptualization, Writing – original draft, Writing – review & editing. **Mathieu Ramon:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Frederic Rimet:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Susanne Schneider:** Investigation, Writing – review & editing. **Kálmán Tapolczai:** Formal analysis, Writing – review & editing. **Jonas Zimmermann:** Conceptualization, Investigation, Writing – original draft, 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.

Data availability

Data will be made available on request.

Acknowledgement

We thank Jón S. Ólafsson (Marine and Freshwater Research Institute, Iceland), for the collection of data from Iceland and helpful feedback. We thank the Norwegian Institute for Water Research (NIVA) for the water chemistry data for the Norwegian samples. We thank Agnès Bouchez (INRAE, UMR CARRTEL, France) for fruitful discussions and guidance, and the INRAE molecular platform (UMR CARRTEL, France) for technical support in processing the Nordic dataset. We thank Bonnie Bailet for a first curation of the Nordic dataset within her PhD project. We thank Stefan Hellgren (SLU) for creating the maps. Specific thanks regarding the collection of the German samples to Petra Werner. Warms thanks to Valentin Vasselon and Simon Hallstan (SLU) who contributed pilot analyses, and Martin Pfannkuchen and Ana Baric who fished for the target ASVs in their marine dataset without getting a hit. We are grateful to David G. Mann and Shinya Sato who contributed with many new sequences of the *Fragilaria/Ulnaria* in an UK barcoding project, and were involved in the taxonomic curation as coauthors of Kahlert et al. (2019). We are also grateful to Dr. Alexey Morozov for fruitful discussions regarding methods, analysis, and conclusions of the present article. This article was in parts inspired by the COST Action DNAqua-Net

(CA15219).

Funding

Part of the work of Maria Kahlert was supported by the Swedish Agency for Marine and Water Management [dnr 4303-2020]. Part of the work of Jonas Zimmermann was supported by the Bundesministerium für Bildung und Forschung [01LI1501E] for GBOL2 Diatoms.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2022.109114.

References

Bailet, B., 2021. New methods for improving water management - Exploring the role of diatoms in ecosystems. Sveriges lantbruksuniv. Acta Universitatis Agriculturae Sueciae 1652–6880.

Bailet, B., Apotheloz-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 https://doi.org/10.1016/j.scitotenv.2020.140948.

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, e34002.

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. https://doi.org/10.1038/ismej.2017.119.

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.

Charles, D.F., Kelly, M.G., Stevenson, R.J., Poikane, S., Theroux, S., Zgrundo, A., Cantonati, M., 2021. Benthic algae assessments in the EU and the US: Striving for consistency in the face of great ecological diversity. Ecol. Ind. 121 https://doi.org/10.1016/j.ecolind.2020.107082.

Clarke, K.R., 1993. Non-parametric multivariate analysis of changes in community structure. Aust. J. Ecol. 18, 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x.

Crawford, R.M., Canter, H.M., Jaworski, G.H.M., 1985. A study of 2 morphological variants of the diatom fragilaria-crotonensis kitton using electron-microscopy. Ann. Bot. 55, 473–485. https://doi.org/10.1093/oxfordjournals.aob.a086925.

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.

Filipchenko, Y., 1927. Variabilität und variation. Gebrüder Borntraeger, Berlin.

Finlay, B.J., 2002. Global dispersal of free-living microbial eukaryote species. Science 296, 1061–1063. https://doi.org/10.1126/science.1070710.

Griffith, D.M., Veech, J.A., Marsh, C.J., 2016. cooccur: Probabilistic Species Co-Occurrence Analysis in R. J. Stat. Software Code Snippets 69, 1–17. https://doi.org/10.18637/jss.v069.c02.

Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontol. Electronica 4 (1), 9.

Hillebrand, H., Gruner, D.S., Borer, E.T., Bracken, M.E.S., Cleland, E.E., Elser, J.J., Harpole, W.S., Ngai, J.T., Seabloom, E.W., Shurin, J.B., Smith, J.E., 2007. Consumer versus resource control of producer diversity depends on ecosystem type and producer community structure. PNAS 104, 10904–10909. https://doi.org/10.1073/pnas.0701918104.

Hutchinson, G.E., 1961. The Paradox of the Plankton. Am. Nat. 95, 137–145.

Juggins, S., Birks, H.J., 2012. Quantitative Environmental Reconstructions from Biological Data, in: Birks, H.J.B., Lotter, A.F., Juggins, S., Smol, J.P. (Eds.), Tracking Environmental Change Using Lake Sediments. Springer Netherlands, pp. 431-494, 10.1007/978-94-007-2745-8_14.

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.

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.

Kelly, M. 2015. Baffled by the benthos (1). microscopesandmonsters [Online]. Available from: https://microscopesandmonsters.wordpress.com/2015/01/10/baffled-by-the-benthos-1/ (accessed 26.01. 2022).

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D., Pass, D., Sapp, M., Sato, S., Glover, R., 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. 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. Ind. 118 https://doi.org/10.1016/j.ecolind.2020.106725.

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, 27–33. https://doi.org/10.1080/0269249X.2014.951398.

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., McDonald, S.M., Bayer, M.M., Droop, S.J.M., Chepurnov, V.A., Loke, R.E., Ciobanu, A., du Buf, J.M.H., 2004. The Sellaphora pupula species complex (Bacillariophyceae): morphometric analysis, ultrastructure and mating data provide evidence for five new species. Phycologia 43, 459–482. https://doi.org/10.2216/i0031-8884-43-4-459.1.

McCune, B., Grace, J.B., 2002. Analysis of ecological communities.

McCune, B., Mefford, M.J., 2011. PC-ORD. Multivariate Analysis of Ecological Data. Version 6.08, 6.08 ed. MjM Software, Gleneden Beach, Oregon, U.S.A.

Mitchell, J.G., Seuront, L., Doubell, M.J., Losic, D., Voelcker, N.H., Seymour, J., Lal, R., 2013. The Role of Diatom Nanostructures in Biasing Diffusion to Improve Uptake in a Patchy Nutrient Environment. PLoS ONE 8, e59548.

Pajunen, V., Kahlert, M., Soininen, J., 2020. Stream diatom assemblages as environmental indicators - A cross-regional assessment. Ecol. Ind. 113 https://doi.org/10.1016/j.ecolind.2020.106183.

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. 798, 149029 https://doi.org/10.1016/j.scitotenv.2021.149029.

Poikane, S., Kelly, M.G., Salas Herrero, F., Pitt, J.-A., Jarvie, H.P., Claussen, U., Leujak, W., Lyche Solheim, A., Teixeira, H., Phillips, G., 2019. Nutrient criteria for surface waters under the European Water Framework Directive: Current state-of-the-art, challenges and future outlook. Sci. Total Environ. 695, 133888 https://doi.org/10.1016/j.scitotenv.2019.133888.

Porter, T.M., Hajibabaei, M., 2020. Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. Front. Ecol. Evol. 8 https://doi.org/10.3389/fevo.2020.00248.

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

Rimet, F. 2020. Diat.barcode: a curated barcoding database [Online]. INRAE. Available: https://www6.inrae.fr/carrtel-collection/Barcoding-database (Accessed 21 Jan 2020).

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, 15116. https://doi.org/10.1038/s41598-019-51500-6.

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

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

Ruggiero, M.V., Sarno, D., Barra, L., Kooistra, W.H.C.F., Montresor, M., Zingone, A., 2015. Diversity and temporal pattern of Pseudo-nitzschia species (Bacillariophyceae) through the molecular lens. Harmful Algae 42, 15–24. https://doi.org/10.1016/j.hal.2014.12.001.

Scherz, M.D., Rasolonjatovo, S.M., Köhler, J., Rancilhac, L., Rakotoarison, A., Raselimanana, A.P., Ohler, A., Preick, M., Hofreiter, M., Glaw, F., Vences, M., 2020. 'Barcode fishing' for archival DNA from historical type material overcomes taxonomic hurdles, enabling the description of a new frog species. Sci. Rep. 10, 19109. https://doi.org/10.1038/s41598-020-75431-9.

Sefbom, J., Kremp, A., Rengefors, K., Jonsson, P.R., Sjöqvist, C., Godhe, A., 2018. A planktonic diatom displays genetic structure over small spatial scales. Environ. Microbiol. 20, 2783–2795. https://doi.org/10.1111/1462-2920.14117.

Smol, J.P., Stoermer, E.F.e., 2010. The diatoms: applications for the environmental and earth sciences. Cambridge University Press.

Soininen, J., Jamoneau, A., Rosebery, J., Leboucher, T., Wang, J., Kokociński, M., Passy, S.I., 2019. Stream diatoms exhibit weak niche conservation along global environmental and climatic gradients. Ecography 42, 346–353. https://doi.org/10.1111/ecog.03828.

Soininen, J., Jamoneau, A., Rosebery, J., Passy, S.I., 2016. Global patterns and drivers of species and trait composition in diatoms. Glob. Ecol. Biogeogr. 25, 940–950. https://doi.org/10.1111/geb.12452.

Sorhannus, U., Ortiz, J.D., Wolf, M., Fox, M.G., 2010. Microevolution and speciation in Thalassiosira weissflogii (Bacillariophyta). Protist 161, 237–249. https://doi.org/10.1016/j.protis.2009.10.003.

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 https://doi.org/10.3389/fevo.2019.00409.

Tapolczai, K., Selmeczy, G.B., 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. Ecological Indicators 122, 107322, https://doi.org/10.1016/j.ecolind.2020.107322.

ter Braak, C.J.F., Van Dam, H., 1989. Inferring pH from diatoms - a comparison of old and new calibration methods. Hydrobiologia 178, 209–223. https://doi.org/10.1007/BF00006028.

The European Parliament and The Council of the European Union, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official J. Eur. Union L327, 1–73.

Turk Dermastia, T., Cerino, F., Stanković, D., Francé, J., Ramšak, A., Žnidarič Tušek, M., Beran, A., Natali, V., Cabrini, M., Mozetič, P., 2020. Ecological time series and integrative taxonomy unveil seasonality and diversity of the toxic diatom Pseudo-nitzschia H. Peragallo in the northern Adriatic Sea. Harmful Algae 93, 101773. https://doi.org/10.1016/j.hal.2020.101773.

Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., Agnès, B., 2019. Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: experience and developments from France Water Framework Directive networks. Metabarcoding Metagenomics 3. https://doi.org/10.3897/mbmg.3.39646.

Vanelslander, B., Creach, V., Vanormelingen, P., Ernst, A., Chepurnov, V.A., Sahan, E., Muyzer, G., Stal, L.J., Vyverman, W., Sabbe, K., 2009. Ecological differentiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* (Bacillariophyceae). J. Phycol. 45, 1278–1289. https://doi.org/10.1111/j.1529-8817.2009.00762.x.

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. Ind. 82, 1–12. https://doi.org/10.1016/j.ecolind.2017.06.024.

Veech, J.A., 2013. A probabilistic model for analysing species co-occurrence. Glob. Ecol. Biogeogr. 22, 252–260. https://doi.org/10.1111/j.1466-8238.2012.00789.x.

Verberk, W., 2011. Explaining General Patterns in Species Abundance and Distributions. Nat. Educ. Knowl. 3, 38.

Vyverman, W., Verleyen, E., Sabbe, K., Vanhoutte, K., Sterken, M., Hodgson, D.A., Mann, D.G., Juggins, S., Vijver, B.V.D., Jones, V., Flower, R., Roberts, D., Chepurnov, V.A., Kilroy, C., Vanormelingen, P., Wever, A.D., 2007. Historical processes constrain patterns in global diatom diversity. Ecology 88, 1924–1931. https://doi.org/10.1890/06-1564.1.

Worm, B., Lotze, H.K., Hillebrand, H., Sommer, U., 2002. Consumer versus resource control of species diversity and ecosystem functioning. Nature 417, 848–851. https://doi.org/10.1038/nature00830.