Using DNA to understand river diatom communities

Chief Scientist's Group research report

Date: November 2023

Version: SC210004/R

We are the Environment Agency. We protect and improve the environment.

We help people and wildlife adapt to climate change and reduce its impacts, including flooding, drought, sea level rise and coastal erosion.

We improve the quality of our water, land and air by tackling pollution. We work with businesses to help them comply with environmental regulations. A healthy and diverse environment enhances people's lives and contributes to economic growth.

We can't do this alone. We work as part of the Defra group (Department for Environment, Food & Rural Affairs), with the rest of government, local councils, businesses, civil society groups and local communities to create a better place for people and wildlife.

Published by:

Environment Agency
Horizon House, Deanery Road,
Bristol BS1 5AH

www.gov.uk/environment-agency

Further copies of this report are available from our publications catalogue: www.gov.uk/government/publications or our National Customer Contact Centre: 03708 506 506

Email: enquiries@environment-agency.gov.uk

Author(s):
Joe Taylor
Martyn Kelly
David Mann
Stephen Juggins
Daniel Read
Jo-Anne Pitt

Keywords:
Diatoms, metabarcoding, HTS, TDI, DNA, ecological, assessment, sequencing, bioinformatics, algae

Research contractor:
Daniel Read, UK Centre for Ecology and Hydrology, MacLean Building, Benson Ln, Crowmarsh Gifford, Wallingford OX10 8BB

Environment Agency's Project Manager:
Jo-Anne Pitt

Project number: SC210004

Citation: Environment Agency (2023) Using DNA to understand river diatom communities. Environment Agency, Bristol.

Research at the Environment Agency

Scientific research and analysis underpins everything the Environment Agency does. It helps us to understand and manage the environment effectively. Our own experts work with leading scientific organisations, universities and other parts of the Defra group to bring the best knowledge to bear on the environmental problems that we face now and in the future. Our scientific work is published as summaries and reports, freely available to all.

This report is the result of research commissioned by the Environment Agency's Chief Scientist's Group.

You can find out more about our current science programmes at https://www.gov.uk/government/organisations/environment-agency/about/research

If you have any comments or questions about this report or the Environment Agency's other scientific work, please contact research@environment-agency.gov.uk.

Dr Robert Bradburne
Chief Scientist

Contents

Executive summary 7

1. Background and rationale 9

1.1 Potential for improving bioinformatics for detecting diversity in diatom assemblages 11

1.2 Looking beyond diatoms: the potential for detecting other algae 12

2. Optimising and improving bioinformatic analysis of diatom metabarcoding data 17

2.1 Introduction 17

2.2 Comparison of different pipelines using diatom mock community metabarcoding data 19

2.2.1 Background 19

2.2.2 Methods 21

2.2.3 Results and discussion 23

2.2.4 Recommendations 26

2.3 Comparison of HTS bioinformatic pipelines and LM for assessment of diatom assemblages 27

2.3.1 Introduction 27

2.3.2 Methods 27

2.3.3 Results 29

2.3.4 Discussion 36

2.3.5 Recommendations 40

3. Maximising the potential of rbcL metabarcoding data 42

3.1 Objectives 42

3.2 Analysis of non-diatom sequences within rbcL gene fragment metabarcoding libraries 42

3.2.1 Introduction 42

3.2.2 Methods 43
3.2.3 Results and discussion....43
3.2.4 Recommendations 47
3.3 Maximising the potential of rbcL metabarcoding ASVs....47
3.3.1 Introduction....47
3.3.2 Methods 47
3.3.3 Results 51
3.3.3 Discussion....64
3.3.4 Recommendations 69
4.Moving forward 70
4.1 Phase 1 – diatom metabarcoding methods 70
4.1.1 Options....70
4.1.2 Recommendation 71
4.2 Phase 2 – inclusion of non-diatom taxa....71
4.2.1 Options....71
4.2.2 Evaluation....72
4.2 Future work....73
References 74
Appendix I....84
Appendix II....86

Acknowledgements

We thank Dr Kerry Walsh, Jonathan Porter and Tim Jones at the Environment Agency for their support in shaping the project and making data available.

Executive summary

DNA-based analysis of ecological samples could provide a cost-effective means of acquiring species data and an opportunity to establish deeper insights into the make-up of biological communities. Benthic diatoms have been used in environmental assessment of aquatic environments in the UK for many years; the Trophic Diatom Index (TDI) forms the basis of the Water Framework Directive (WFD) classification tool DARLEQ, using benthic diatoms as a proxy for the wider phytobenthos. The Environment Agency, in collaboration with the Scottish Environment Protection Agency (SEPA), has developed a DNA-based method for identifying freshwater benthic diatoms. The relationship between the DNA data and nutrient pressures in rivers is similar to that for traditional light microscopy (LM) data. However, the 2 methods produce differing outputs in terms of the taxa identified and quantified, resulting in different assessments of ecological status in 35% of sites in a test data set. In this project, we investigated whether additions to the curated ribulose-bisphosphate carboxylase (rbcL) DNA barcode reference database and advances in bioinformatics processing pipelines affected differences between microscopy and DNA outputs (Phase 1). We also explored the potential to extract more information from the DNA analysis by including data from the wider phytobenthic community (Phase 2).

In phase 1, diatom rbcL metabarcoding data generated from ~1,500 UK river samples were analysed with 4 bioinformatics pipelines (UPARSE, UCLUST, UNOISE3 and DADA2), and taxonomically classified against the latest rbcL diatom barcode reference database (diat.barcode) using the Ribosomal Database Project (RDP) classifier. The recently developed next generation sequencing based DARLEQ tool (DARLEQ3) was used to calculate TDI and ecological status class. Results were compared to LM data, and to data generated using the Environment Agency's original high-throughput sequencing (HTS) pipeline. An 11 species mock community was sequenced using the rbcL marker and analysed with all pipelines, to compare their accuracy and precision at recovering known taxa.

Correlations between LM TDI and that calculated from DNA data were in the range 0.7 to 0.85, similar to the HTS pipeline (r=0.87), with UNOISE3 having the lowest correlation. All pipelines produced ecological status classes in agreement with LM for ~60 to 65% of sites, apart from UNOISE3 (43%). Based on our analyses, we recommend using the DADA2 pipeline in the future. We believe it to be an accurate, stable pipeline and with widespread use is unlikely to change within the next 3 to 5 years, which should improve user confidence. More work is still needed to understand the relationship between outputs from bioinformatic pipelines and the biological communities they represent. The differences observed between LM and DNA analyses (and subsequent derivation of ecological status) appear to be fundamental to the different forms of data they produce and are unlikely to be resolved through advances in DNA analytical techniques and improved reference databases alone.

In phase 2, additional samples collected through the Environment Agency's routine river monitoring programme (2017 to 2019) were included in the analysis. All data were analysed

using the DADA2 pipeline, which implements a high-resolution amplicon sequence variant (ASV) approach. While non-diatom taxonomic groups were detected, few contained the diversity that would be expected in riverine environments. This may be due to the performance of the current polymerase chain reaction (PCR) primers in characterising the wider phytobenthic community and influenced by the sampling method used, as both were developed specifically for diatoms.

The data set was split for further analysis into diatoms and non-diatoms. Different models were applied to the data to look at the predictive power of individual taxa and community responses to nutrients. ASVs that showed significant pressure responses were further classified against the National Center for Biotechnology Information (NCBI) GenBank database that contains rbcL sequence data from chloroplast containing species.

Comparing phase 2 results with those from phase 1 indicates that both diatom and non-diatom ASV-based models outperform the equivalent taxonomy-based HTS models. In addition, results suggest that diatom ASV models may perform as well as, or better than, the equivalent LM model. Part of this improvement is likely due to the finer taxonomic resolution offered by DNA analysis.

We recommend that future developments should use ASVs to calculate metrics, with links to reference databases made as a final step to generate taxa lists to support interpretation. Any further exploration of the potential of non-diatoms would benefit from access to a well-curated reference database, similar to diat.barcode. Such a database does not yet exist, and we caution against the indiscriminate use of NCBI GenBank as a taxonomic resource as the rbcL sequences deposited are not checked for errors.

The present study indicates that there is relatively little scope for improvement in the current approach (deriving ecological quality ratios (EQRs) from measures of community turnover). Therefore, the possibility of developing alternative metrics (for example, incorporating diversity) or bypassing EQR calculation and predicting status class directly should be explored.

This study also identified considerable diversity in Eustigmatophyceae (previously poorly known) and a wider distribution than previously thought for the freshwater Phaeophyceae. However, beyond the formal remit of the project, these results offer a strong case for the benefits of metabarcoding in expanding knowledge of aquatic biodiversity in the UK.

1. Background and rationale

To meet the requirements of the Water Framework Directive (WFD) (2000/60/EC), the UK regulatory and conservation agencies, under the auspices of the UK Technical Advisory Group (UKTAG), developed and subsequently refined a number of ecological assessment tools for rivers, informed by the normative definitions of ecological status in Annex V of the directive. These tools rely on the traditional approach to sampling and identifying aquatic organisms, maintaining continuity with pre-existing data and established biological identification skills.

Benthic diatoms have been used in environmental assessment in the UK for many years, principally in rivers but with some application in lakes. They are easily sampled alongside the collection of other biological and/or environmental parameters and have been shown to respond to nutrient pressure. The Trophic Diatom Index (TDI) was developed in the 1990s (Kelly and Whitton, 1995) and subsequently modified to provide an assessment tool (DARLEQ) that was compliant with the requirements of the WFD, diatoms being used as a proxy for the status of the wider phytobenthos (Kelly and others, 2008). The DARLEQ tool was derived from, and designed for use with, light microscopy (LM) data, and was intercalibrated to a common definition of good ecological status as required under the WFD.

DNA-based analysis of biological samples potentially offers a more cost-effective way of acquiring species data, and an opportunity to establish deeper insights into the make-up of biological communities than more conventional methods. The Environment Agency has been at the forefront of developments in using DNA to identify freshwater benthic diatoms in recent years, making significant progress in collaboration with the Scottish Environment Protection Agency (SEPA) and with input from other UK agencies (Environment Agency 2018, 2020 and SEPA 2018).

A relationship between the diatom DNA data and nutrient pressures has been demonstrated, which is similar to that for LM data. However, the 2 methods produce differing outputs in terms of the taxa identified and quantified. This leads, in some instances, to different TDI scores and calculated ecological quality ratio (EQR) values. Consequently, where comparative data are available, mismatches in the WFD phytobenthos status classification are seen in about 35% of river sites. It is important, from a regulatory perspective, that the ecological status of waterbodies can be assessed over time, with confidence in the reasons for observed changes and trends. The mismatch between the DNA and LM classifications is not easily explained and was considered unacceptably large when reviewed by the devolved UK administrations in 2019. Therefore, the DNA-based version of DARLEQ was not adopted for formal classification and reporting purposes for the third WFD river basin planning cycle.

In this project, we re-investigate the differences between methods, in the light of further barcode additions to the curated rbcL DNA barcode reference database (diat.barcode: Rimet and others, 2019; https://www6.inrae.fr/carrtel-collection_eng/Barcoding-database),

and advances in bioinformatics processing algorithms (Phase 1). In addition, we make an initial exploration of different analytical approaches to extract more information from the DNA data, seeking to make more effective use of the available information in an assessment of the wider phytobenthic community (Phase 2). Phase 1 of the project used samples previously collected for developing and refining the original bioinformatics pipeline (Environment Agency 2018; 2020), in this report referred to as the 'HTS pipeline'. This pipeline was based on Quantitative Insights into Microbial Ecology (QIIME) 1 and used the clustering algorithm UCLUST. Phase 2 included additional samples collected and analysed as part of routine river monitoring programmes in 2017, 2018 and 2019.

The research and application of high-throughput sequencing (HTS) for characterising the taxonomic composition of biological communities has matured in recent years, to the point at which these techniques have moved from academic research to being used by regulatory agencies. Applying these methods for ecological assessment could potentially revolutionise the way in which species distribution and biodiversity data are collected and interpreted. Recording species composition at lower cost and in less time, would allow the implementation of monitoring schemes with higher spatial and temporal resolution, encompassing a wider range of taxa and revealing patterns of biodiversity that are not possible using current approaches.

Metabarcoding assays targeting specific marker genes were originally developed by microbial ecologists for characterising bacterial and fungal communities, driven by the lack of alternative approaches to describe community composition in organisms without distinguishing morphological features. A pair of short oligonucleotides ('primers') are used to bind to and control amplification of all or part of a specific marker gene, amplifying and sequencing gene variants representing the diversity of organisms in a mixed community sample. Samples can be barcoded with unique DNA sequence tags, meaning many samples can be 'multiplexed' (run together) on an individual sequencing run. After computational processing to assign sequences to samples and perform error correction (either via clustering similar sequences or by corrections based on known sequence error profiles), sequences can be matched to a taxonomic group by comparing them against known sequences stored in a reference database. Metabarcoding has been applied to diverse communities from terrestrial and aquatic environments, including terrestrial invertebrates (Beng and others, 2016; Dopheide and others, 2019), soil fauna (Yang and others, 2014), zooplankton (Schroeder and others, 2020; Zhang and others, 2018), freshwater invertebrates (Bista and others, 2017; Hajibabaei and others, 2011; Kuntke and others, 2020), diatoms (Bailet and others, 2020; Rimet and others, 2019) and higher plants (Baksay and others, 2020; Pornon and others, 2017).

Metabarcoding of diatom assemblages is one of the most advanced applications in terms of use by regulatory agencies. This traces back to work by 2 independent groups; a UK consortium working with the Environment Agency (Environment Agency 2018, 2020; Kelly and others, 2020) and the INRAE lab at Thonon, France (for example, Vasselon and others, 2018), along with groups in Germany, Scandinavia, Hungary, Spain, Portugal and elsewhere (for example, Duleba and others, 2021). After initial trials of a fragment of the

ribulose-1,5-bisphosphate carboxylase (rbcL) gene and V4 region of 18S rRNA gene, rbcL was recommended as the marker of choice (Mann and others, 2010). The exact rbcL barcodes (and primers) developed by the UK and INRAE groups differ slightly, the INRAE barcode being slightly shorter (263 compared with 331 base pairs). However, the INRAE barcode is contained within the UK region, and it is therefore possible to directly compare metabarcoding outputs obtained by the 2 systems (the INRAE system gives slightly lower phylogenetic resolution because of the shorter barcode) and use results from both to guide further development of ecological assessment tools.

1.1 Potential for improving bioinformatics for detecting diversity in diatom assemblages

There are currently a variety of bioinformatics pipelines and approaches available for analysing metabarcoding data, including those developed specifically for diatoms based on the QIIME, Mothur and DADA2 analysis pipelines. Bailet and others (2020) demonstrated that there were differences between the outputs from different bioinformatic approaches, particularly in the assignation of sequence reads to specific taxa. The original development of the UK DARLEQ 3 tool (https://github.com/nsj3/darleq3) for high-throughput sequencing data used a custom-built pipeline based around the QIIME package of software (Caporaso and others, 2010) and curated reference database (Environment Agency 2018, 2020). There was a good correlation (r=0.77) between the LM TDI and HTS TDI EQR values. However, approximately 35% of sites would be assigned to a different WFD status class as a result of changing method (differences were both positive and negative, with no bias). Since this work was undertaken, alternative bioinformatics approaches have been developed and more barcodes added to online databases (Rimet and others, 2019). In addition, Vasselon and others (2018) showed that the number of rbcL sequence reads for diatoms is partially a function of cell biovolume, rather than of the number of cells present (the basis for LM enumeration). This means that there is a fundamental difference in the type of data collected by the 2 approaches which hinders simplistic attempts to fit metabarcoding outputs to expectations based on LM analyses.

One of the main changes in bioinformatics approaches has been the development of metabarcoding pipelines that generate amplicon sequence variants (ASV) instead of operational taxonomic units (OTU). The definition of an OTU in molecular ecology is a group of sequences that are similar to each other, based on a threshold which for most marker genes is set at 97% similarity. This cut-off is used on the assumption that each group of 97% similar sequences has come from a single taxon/species or very closely related species. However, the similarity for some taxa may be set either higher or lower, based on what is known about their phylogenetic differences for specific marker genes. For most microbial groups and marker genes, including diatoms using the rbcL marker, 97% is used as a broad compromise to separate the majority of sequences. One of the main purposes of using OTUs has been to attempt to correct polymerase chain reaction (PCR) errors or sequencing errors produced during the generation of the sequence data, although for many earlier algorithms many of these errors remain. ASVs are fundamentally different to OTUs in that

no clustering or grouping of the sequences takes place. Instead, error correction algorithms are applied to the data so that erroneous sequences generated during PCR and sequencing are corrected or removed. ASVs are therefore a fine scale method of recovering sequences from unprocessed metabarcoding data and can detect differences as small as a single nucleotide. This means very closely related taxa or within-taxa variation can be detected.

Despite the demonstrated capabilities and potential advantages of metabarcoding as a tool for studying and monitoring the composition of biological communities, there are still technical challenges to be overcome. Although state-of-the-art at the time, the clustering algorithm originally used to generate OTUs in the UK pipeline (UCLUST) has been shown to generate many spurious OTUs in comparison to clustering algorithms used in more recent pipelines for other marker genes (16S, 18S, ITS) (Flynn and others, 2015; Majaneva and others, 2015; Prodan and others, 2020). Recent studies have used the modern version of the Mothur pipeline (Schloss 2020) to generate OTUs for diatom metabarcoding data (Rivera and others, 2021), while DADA2 has been used to generate ASVs in other diatom metabarcoding studies (Apothéloz-Perret-Gentil and others, 2021; Pérez-Burillo and others, 2021; Tapolczai and others, 2021). No comparison between different OTU/ASV methods to assess biological accuracy against mock communities has yet been performed. The most comprehensive study to date, Bailet and others, (2020), used only environmental samples to compare outputs from different OTU pipelines with each other and against light microscopy data. Comparisons have shown, for other marker genes (such as 16S rRNA), that DADA2 can also generate some spurious ASVs when compared to the generation of ASVs using other pipelines such as UNOISE3 (Nearing and others, 2018; Prodan and others, 2020). Therefore, a comparison of ASV pipelines is also needed. However, it has been shown that, although rare, individual diatom morphotaxa can hold multiple variants of the rbcL gene (Pérez-Burillo and others, 2021) and there may also be population level or biogeographic variation, which may cause problems in the ecological interpretation of individual ASVs. On the other hand, ASVs can vary at the regional or local scale, indicating the potential for improved resolution when looking at fine scale environmental changes (Pérez-Burillo and others, 2021). There may be a case for using both OTUs and ASVs to address different questions. All these pipelines have already been adapted or can be adapted for the UK rbcL barcode.

1.2 Looking beyond diatoms: the potential for detecting other algae

Initial development focused on diatoms, as this group of algae is widely used as a proxy for the wider phytobenthos for WFD-related status assessments across the EU (Charles and others, 2021). Before the development of metabarcoding, this approach reflected a combination of the practical advantages diatoms offered, an established package of working practices that developed around the use of diatoms, and the sheer scale of the challenge involved in tackling the whole range of microphytobenthos using light microscopy. Focusing on diatoms during the first explorations of the potential of metabarcoding made practical sense not only because of the expertise and knowledge of diatom ecology accumulated

over the years, but also because the body of data from LM analyses offered a unique opportunity to benchmark metabarcoding outputs.

However, this work also showed a large proportion of sequence reads were not being assigned to diatoms, raising questions about whether additional ecological information was being discarded simply because sequence reads did not belong to the phylogenetic group selected for its benefits when using light microscopy. Non-diatom algae are already used in a few countries either alongside (Germany: Schaumburg and others, 2004) or instead of (Norway: Schneider & Lindstroem, 2011) diatoms, and macroalgae are included in macrophyte assessment systems in several countries (Charles and others, 2021). Extending the scope of the metabarcoding approach to embrace other groups of algae appears, therefore, to be a logical development.

The potential for using these algae, however, is complicated by several issues, which can be summarised under 2 headings:

1. **The likelihood of an alga being found in a biofilm sample.** The Rhodophyte Lemanea fluviatilis, for example, is common in rivers and is strongly associated with high and good status. However, it tends to be attached to larger stones than are usually collected when sampling diatoms, typically in the fastest flowing sections of a reach. It is therefore likely to be underrepresented by current sampling protocols. This patchiness of stream phytobenthos assemblages means that better resolution of non-diatom algae in a biofilm sample will not automatically translate into a better understanding of the phytobenthos present in a river reach.
2. **The likelihood of rbcL from an alga being detected by the primers.** The term ‘algae’ is a catch-all phrase embracing organisms from 2 domains and 4 kingdoms, which diverged in deep geological time. Even though rbcL, as a gene for an important photosynthetic enzyme, will be highly conserved, there is still considerable variation in the structure between groups. Linked to this issue are questions about the reliability of reference libraries for several of the groups likely to be detected. Finally, as the primers have been optimised for diatoms, there are embedded issues of ‘primer bias’ when detecting these other algal groups, particularly those distantly related to diatoms. Non-diatom sequence reads are best regarded as ‘bycatch’ rather than as an ecologically coherent extension of the existing method.

Table 1 provides an overview of the major algal groups and their likelihood of being detected during the present study. It should be noted that existing knowledge of some groups is inadequate due to the limitations of collecting reliable data on tiny, morphologically plastic organisms from field samples using light microscopy. It is quite likely that, in a few cases, (for example, Eustigmatophyceae, Phaeophyceae) metabarcoding may make a significant contribution to current knowledge of the extent of these organisms in the UK.

Table 1. An overview of the major algal groups in freshwaters and their likelihood to be included in non-diatom sequence reads using the current UK primers (Forward rbcL-646F 5´-ATGCGTTGGAGAGARCGTTT-3´, Reverse rbcL-998R 5´-GATCACCTTCTAATTTACCWACAACTG-3´; Kelly and others, 2020). Information based on the authors' personal experience, John and others (2011), Adl and others (2019) and AlgaeBase

Group	Family/Sub-group	Likely to be in sample?	Likely to be amplified?
Cyanobacteria		Yes. Very abundant in river biofilms; many good indicators of ecological conditions.	Unlikely. Also poor representation in rbcL reference libraries.
Ochrophyta (Diatom relatives)			In theory, Ochrophyta, the division which includes the diatoms, have the rbcL which is most closely related and should therefore amplify. In practice, there are several issues – see 3.2.3.
	Chrysophyceae (Golden algae)	Rarely abundant in rivers, benthic forms only common in upland streams in winter.	
	Dictyochophyceae	Small group, rarely recorded, possibly overlooked; most records from standing waters.	
	Eustigmatophyceae	Relatively little known about distribution.	
	Phaeophyceae (Brown algae)	Freshwater representatives form crusts which may not be removed with usual diatom sampling approach; likely to be widespread but under recorded.	
	Phaeosacciophyceae	Very little [nothing?] known about freshwater representatives in UK.	
	Phaeothamniophyceae	Relatively little known about distribution in UK.	
	Raphidophyceae	Relatively little known about distribution in UK.	
	Xanthophyceae	Widespread and abundant in rivers.	
Haptophyta		Widespread but mostly in standing waters; possibly overlooked.	Some Haptophyta rbcL will amplify, but the reverse

			primer is suboptimal. However, there are rather few Haptophyta in freshwaters.
Cryptophyta		Widespread in river biofilms; rarely abundant.	Though not closely related to Ochrophyta, Cryptophyta have similar rbcL, which is likely to be amplified.
Miozoa/ dinoflagellates		Widespread, but more common in standing waters.	Extremely unlikely to be amplified (has type II rbcL, which is phylogenetically distinct from rbcL region targeted by UK rbcL primers: Tabita and others, 2008).
Plantae (includes green algae)	Chlorophyta	Very widespread. Many likely to be present in biofilms. Micro and macroalgae representatives.	Evolutionarily extremely distant from diatoms in evolutionary terms, but rbcL in some groups is sufficiently similar that it may be amplified. However, rbcL is not widely used as a marker for this group so reference libraries are incomplete.
	Streptophyta	Very widespread. Many likely to be present in biofilms. Includes mosses.	
	Higher plants	Unlikely to be part of biofilm community; about 15% of biofilms are sampled from macrophytes so DNA likely to be removed in the process; ample eDNA likely within reaches.	
Rhodophyta (Red algae)		Abundant in streams, and sometimes in biofilms. Some	Some species likely to be

		form distinct crusts which may not be removed easily.	amplified, but primer mismatches are common.
Euglenophyta		Present in rivers but rarely in high numbers.	Rubiscos of Euglenophyta are phylogenetically related to Chlorophyta (q.v.); rbcL is interrupted by introns in some species (Karnkowska and others, 2018) and may not amplify/sequence.

2. Optimising and improving bioinformatic analysis of diatom metabarcoding data

2.1 Introduction

The overall objective for phase1 of the project was to re-analyse data from the original Environment Agency projects (Environment Agency 2018, 2020) to determine whether advances in bioinformatics and updates to the diatom barcode database would change the relationship between DNA and LM outputs, with the aim of improving the level of agreement between the WFD classifications produced by each method.

We set out to match rbcL metabarcoding data primarily targeting diatoms, from ~1,500 sites, to LM classification results, using the DARLEQ 3 tool to calculate TDI scores. We used the latest diatom rbcL DNA barcode reference database (diat.barcode: Rimet and others, 2019) and alternative bioinformatics pipelines that use more recent developments in amplicon sequence processing algorithms to deliver improved OTUs or ASVs. Pipelines (Table 2) were selected based on previous work with diatoms (Kang and others, 2021; Pérez-Burillo and others, 2021) or algal communities (Bombin and others, 2021) and were adapted for use with the 331bp fragment of the rbcL gene used in this project.

To assess the accuracy of the pipelines at recovering specific taxa we analysed an 11 species mock community generated in the original work (Environment Agency, 2018; Kelly and others, 2020) with the different analysis pipelines. We also tested 2 methods of sequence pair merging quality filtering. We then selected a final set of pipelines to analyse the ~1,500 environmental samples which contain molecular data, paired to LM and water column nutrient data.

Table 2. Amplicon sequence processing pipelines used in this study

Pipeline	Features
QIIME 1 (Caporaso and others, 2010) (UCLUST, Edgar 2010) **As used in original study- referred to in analysis as 'HTS' pipeline.** **Implemented in Ubuntu/Linux**	• No longer supported/updated. • Dereplicates sequences. • Clusters/picks OTUs at 97% similarity using the UCLUST algorithm which was not designed for OTU clustering originally and generates many spurious/erroneous OTUs. • Removes chimeric sequences using UCHIME.
USEARCH-UPARSE (Edgar 2013) **Implemented in Ubuntu/Linux**	• Dereplicates sequences. • Clusters/picks OTUs at 97% similarity with the UPARSE algorithm Specifically designed for OTU clustering and has been shown repeatedly to generate more biologically accurate OTUs than UCLUST. • Removes chimeric sequences using UCHIME. • Faster runtime and lower memory requirements than QIIME1.
DADA2 (Callahan 2016). **Implemented in R/Rstudio in Windows or Linux**	• Rapidly becoming the standard pipeline in microbial ecology. • First models error profiles of RAW sequencing data. • Has been used in rbcL diatom monitoring studies in Europe. • Removes sequencing errors and chimeras. • Merges paired reads. • Groups sequences at 100% similarity into amplicon sequence variants. • Fixed memory requirements.
USEARCH-UNOISE3 algorithm (Edgar 2016). **Implemented in Ubuntu/Linux**	• Becoming more widely used to generate ASVs in bacterial 16S metabarcoding as comparisons have shown less spurious/more accurate ASVs generated. • Clustering step/chimera detection is performed by UNOISE3 algorithm. • Errors are corrected by removing reads with sequencing and PCR point error. Chimeras are removed.

	• Correct biological sequences are recovered from the reads, again at an effective similarity of 100% generating amplicon sequence variants. • Never been used for rbcL- potentially optimised for 16S. • Potentially faster run time (up to 1,000x) and lower memory requirements than DADA2 (Nearing and others, 2018) - potential capacity to process large volumes of data quicker than DADA2.

2.2 Comparison of different pipelines using diatom mock community metabarcoding data

2.2.1 Background

Using mock communities to validate metabarcoding data is a standard way to assess the accuracy and precision of both the molecular and bioinformatics component in reconstructing the assemblages present in a community. Mock communities validate the molecular analyses by showing which members of the community are amplified and sequenced successfully. The bioinformatic analysis of mock communities then validates the ability of any pipeline to produce accurate and realistic numbers of OTUs or ASVs, as well as to assign the correct taxonomy to those taxonomic units (Hleap and others, 2021). Mock communities have been used effectively for various taxonomic groups such as bacteria (Bukin and others, 2019), fungi (Egan and others, 2018), invertebrates (Braukmann and others, 2019) and algae (dinoflagellates: Smith and others, 2017). Analysis of a mock community is an important step in validating bioinformatic pipelines for metabarcoding data.

The original study (Environment Agency, 2018) included an 11 species mock community in the analysis. DNA from a culture of each species was extracted separately and the DNA mixed in equal volumes for PCR amplification and sequencing. However, QIIME 1 (see Figure 1) generated thousands of OTUs for just 11 species, and identified numerous other species in addition to those included in the mock community.

This demonstrated the potential for either false positives, where sequences are annotated as species during taxonomic assignment that are not present in the sample, or false negatives, where sequences from species that are present in the sample are not annotated at all or misannotated as other taxa. Some of the mismatched species or taxa that failed to be identified in the mock community were, in part, due to an incomplete reference database at the time the study was carried out. Additionally, some of the cultures of diatoms used for the study may have been incorrectly identified, as DNA barcoding identified at least 2 cultures that were not as labelled (Table 4a). As no culture material was available for the current project, we were unable to verify the exact provenance of each culture using microscopy or by comparing the original DNA barcodes to the reference database. However, the overall precision of the original pipeline was poor due to the vast number of OTUs

generated. Theoretically, one might hope, if there is a 'barcode gap', that each species would correspond to a single OTU. However, in many cases, this gap is minimal or absent with the 331-bp rbcL marker, complicating analysis, while some species, as currently defined, contain 2 or more rbcL variants (Perez-Burillo and others. 2021). Despite some potential variation in the rbcL sequences within an individual taxon, there is not an order of magnitude difference between some taxa, as was seen in the original study. Therefore, taxa having hundreds or thousands of associated OTUs would not be biologically accurate. While the original pipeline managed to deal with this by grouping the OTUs at the species level, a huge amount of potential taxonomic resolution would have been lost.

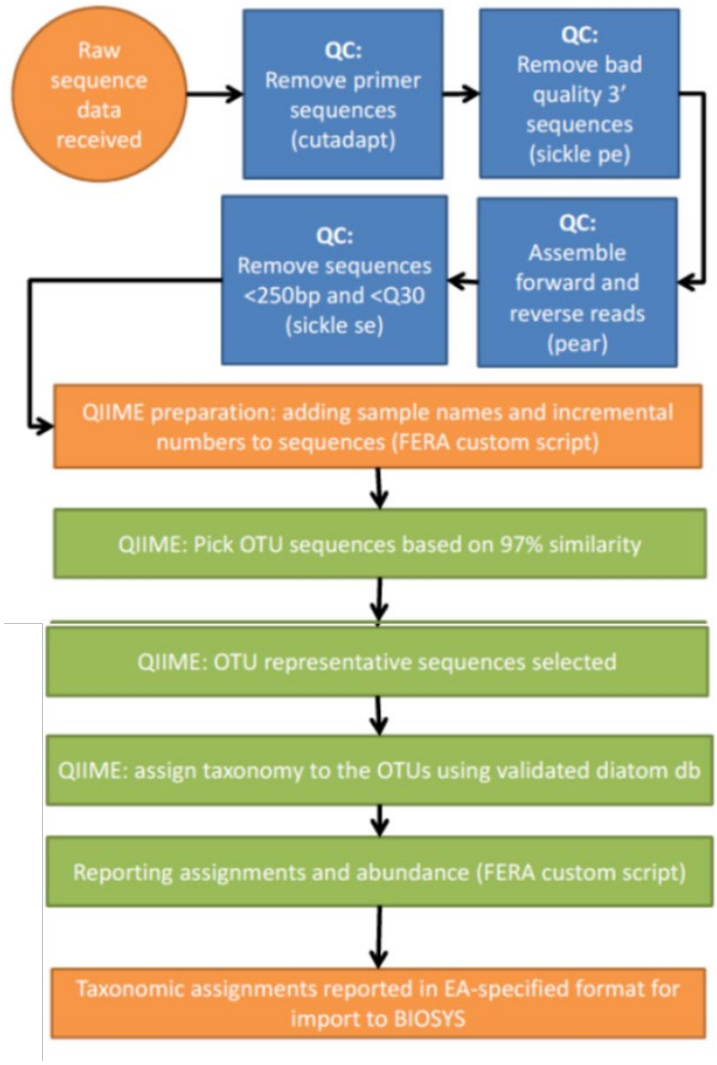


Figure 1: The original Environment Agency HTS pipeline. (Figure reproduced from SC140024 report)

Initial quality filtering steps and the algorithms used to merge paired reads in Illumina data can also impact on the outcomes of bioinformatic pipelines and the accuracy of generated OTUs. The original pipeline used Sickle (Joshi and Fass, 2011), which is a sliding window

quality filter to remove poor quality sequences. Other pipelines such as USEARCH (Edgar, 2010) filter sequences based on expected error (Edgar and Flyvbjerg, 2015). Expected error has been shown to be better at overall error correction of sequences, leading to fewer spurious OTUs (Edgar and Flyvbjerg, 2015). PEAR (Zhang and others, 2014) was also used in the original pipeline to merge paired reads. However, USEARCH has been shown to have a much more stringent read pair merger and is likely to lead to better overall data quality going into the next steps of the pipeline (Edgar and Flyvbjerg, 2015).

Based on previous work for other taxonomic groups (Flynn and others, 2015; Majaneva and others, 2015; Nearing and others, 2018; Prodan and others, 2020;), we expected the alternative pipelines to QIIME 1 (Table 2) would generate far fewer OTUs/ASVs and provide a more accurate recall of the species within the mock community.

The aims of the work described here were to assess:

- the accuracy and precision of new bioinformatics pipelines on diatom metabarcoding data using an 11-species diatom mock community
- the impact of 2 different quality filtering and pair merging algorithms on the overall number of sequence reads and numbers of OTUs/ASVs

2.2.2 Methods

Quality filtering and pair merging

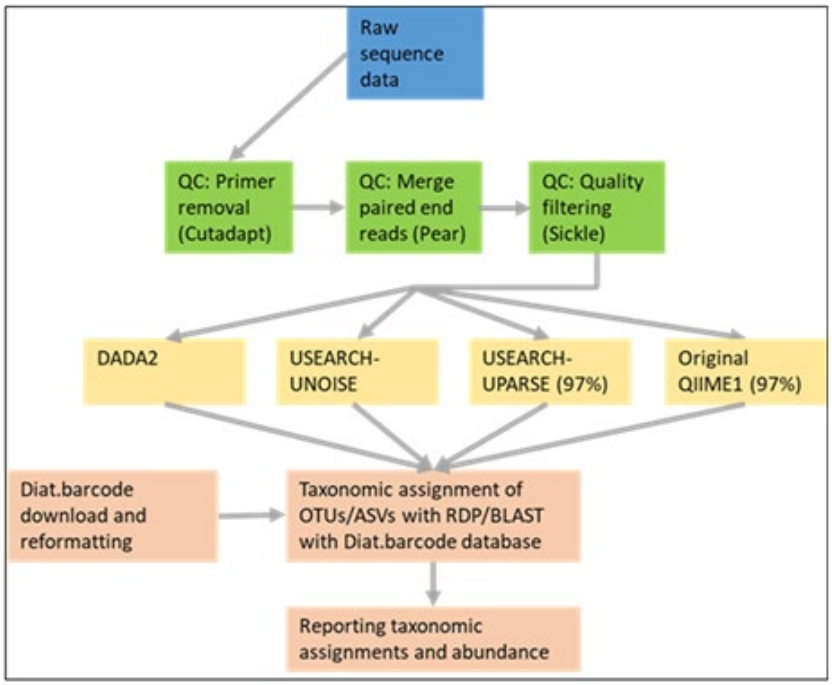
Initial processing steps (quality filtering, pair merging) varied between pipelines and therefore a standardised approach was taken for analysis. Two methods of pair merging and quality filtering were tested on the mock community data. These were:

i) the method used in the original HTS pipeline, using PEAR to merge Read 1 and Read 2, followed by 2 rounds of Sickle to quality filter the sequences (Figure 2a)

ii) USEARCH to both merge paired reads and quality filter (Figure 2b). The pipeline DADA2 applies its own quality filtering and pair merging as part of the pipeline

For taxonomic assignment, all pipelines used the Ribosomal Database Project (RDP) naïve Bayesian classifier (Wang and others, 2007) and v10 of the diat.barcode database (http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/Diat.barcode/) formatted for input into the RDP classifier.

a)



b)

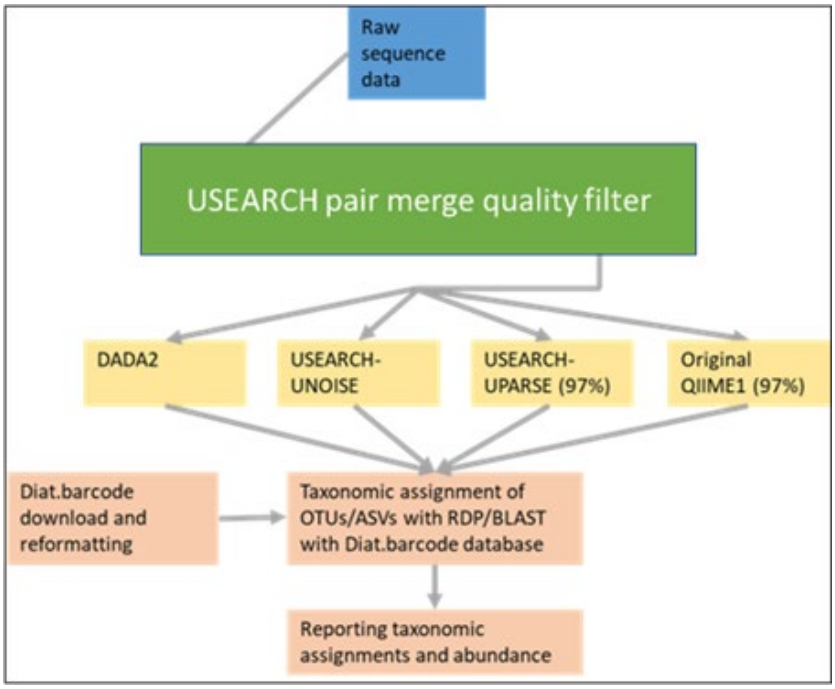


Figure 2 a) Preliminary filtering as for the original HTS pipeline with Sickle and PEAR b) USEARCH filter pipeline

2.2.3 Results and discussion

Sequence metrics between the different pipelines

Table 3 summaries the performance of the 4 pipelines and 2 filtering approaches using the 11 species diatom mock community. The Sickle and PEAR filter approach incorporated many more sequence reads into the filtered data set than the more stringent USEARCH. Overall, far fewer sequence pairs merged using USEARCH (<30%), resulting in a final data set with relatively few sequence reads. The UCLUST pipeline with Sickle and PEAR filtering recovered thousands of OTUs. These were subsequently filtered to hundreds after discarding OTUs with a cluster size <5 and a presence in all 4 mock community libraries. For the UCLUST pipeline 83 OTUs had 100% species matches. DADA2 recovered 51 ASVs, with 10 having 100% species match. UPARSE performed better than UNOISE3 in recovering a more realistic number of taxa. The UPARSE pipeline produced 10 species with 100% match using the preliminary filtering done as original HTS pipeline with Sickle, and 8 using the USEARCH filter. UPARSE using the USEARCH filtering produced 12 OTUs in total, which was the closest to the 11 species analysed. Both DADA2 and USEARCH pipelines ran in less than 15 minutes compared to the UCLUST pipeline, which took over 45 minutes.

Table 3. Total number of OTUs or ASVs: those which showed a 100% species match and total numbers of sequence reads after quality filtering in rbcL gene fragment libraries generated from an 11 species mock community.

	Sickle PEAR filter			USEARCH filter			
Pipeline	UPARSE	UNOISE 3	UCLUST	UPARSE	UNOISE 3	UCLUST	DADA 2
Total OTUs/ASVs	**129**	**234**	**809**	**12**	**44**	**116**	**51**
Numbers of OTUs/ASVs with 100% species match	**10**	**38**	**83**	**8**	**16**	**3**	**10**
Total sequence reads	**227,793**	**298,835**	**240,787**	**7,326**	**8,486**	**7,849**	**86,295**

Taxonomic composition using the different pipelines

All of the pipelines with the different filtering criteria detected either the species or genus of the taxa present in the mock community (Table 4). Cyclotella cryptica, Fragilaria sp. Nitzschia inconspicua and Gomphonema parvulum were only detected at genus level, with

no pipeline producing a 100% species match. This is likely, in part, due to uncertainty in identifying the species used for the mock community: several of them had their taxonomy revised when they were sequenced using single read Sanger sequencing (Table 4a). It is likely that some of the taxa in the mock community are new undescribed species with an erroneous taxonomic identification assigned by the culture collection, making accurate taxonomic identification problematic. This is highlighted by the fact that all pipelines detected very few additional genera or species that weren't present in the mock community. Any additional taxa that were detected are likely to be the correct taxonomy of the species used in the mock community rather than that assigned by the culture collection.

All pipelines with the different filtering criteria performed better than the original analysis of the mock community (See Environment Agency, 2018, figure 5.4). One interesting feature is that all pipelines detected genetic variation in the species Melosira nummuloides, with multiple ASVs/OTUs being assigned a 100% species match. This either shows multiple variants of the rbcL gene within a single taxon or suggests within the algal culture itself there may be a mixed community of very closely related species or variants.

While Fragilaria was not detected by the UPARSE and UNOISE3 pipelines, filtered and merged with USEARCH, this is likely to be because this taxon was present in low abundance within the mock communities and the stringent quality filter removed too much data for it to be detected.

The pair merging implemented by USEARCH was highly stringent, with successful merging in only 20 to 30% of the sequence reads. In comparison, the PEAR sequence successfully merged 60 to 80% of the sequence reads, retaining more data. USEARCH quality filtering, although stringent, generated more realistic numbers of OTUs for the 11 species. Therefore, all further analysis in the project combined pair merging using PEAR and quality filtering with USEARCH. This decision is based on the results obtained here, along with evidence that quality filtering using expected errors leads to better quality sequences than using a sliding window quality filter (Edgar and Flyvbjerg, 2015).

Table 4: a) Species used for the original mock communities and their revision after identification with Sanger sequencing. b) Diatom species mock community detection in the rbcL metabarcoding libraries, using various bioinformatic pipelines and clustering algorithms. Table shows correct species assignment in yellow, genus only in orange and absent in blue.

a)

Mock community species	Culture collection	Culture collection ID	Revised identification following Sanger sequencing
Melosira nummuloides	Bigelow	CCMP482	*Melosira nummuloides*
Cyclotella cryptica	CCAP	CCAP 1070/6	*Cyclotella meneghiniana*
Eucocconeis sp.	Bigelow	CCMP2525	*Nitzschia inconspicua* (98% identity match)
Stephanodiscus hantzschii	CCAP	CCAP 1079/4	*Cyclotella cryptica*
Tabellaria sp.	CCAP	CCAP 1081/7	*Tabellaria flocculosa*
Asterionella formosa	CCAP	CCAP 1005/7	*Asterionella formosa*
Fragilaria crotonensis	SAG Goettingen	28.96	*Fragilaria crotonensis* and *Fragilaria bidens* (99% identity match)
Gomphonema parvulum	SAG Goettingen	1032-1	*Gomphonema parvulum*
Navicula pelliculosa [1]	SAG Goettingen	1050-3	*Mayamaea permitis* (97% identity match)
Nitzschia palea	SAG Goettingen	1052-3a	*Nitzschia palea*
Sellaphora capitata	Ugent	*Sellaphora capitata* D.G. Mann and S. Droop (03x38) F1-9	*Sellaphora capitata*

b)

(SP)= species present 100% match
(GP)=Genus present at 100%
Abscent

		EA Sickle PEAR filter			Usearch filter			
		UPARSE	UNOISE3	UCLUST	UPARSE	UNOISE3	UCLUST	DADA2
	Total OTUs/ASVs	**129**	**234**	**809**	**12**	**44**	**116**	**51**
	Numbers of OTUs/ASVs with							
Total species in mock community =11	**100% species match**	**10**	**38**	**83**	**8**	**16**	**3**	**10**
	Total reads (4 mocks, 4							
	different runs, dilutions)	**227793**	**298835**	**240787**	**7326**	**8486**	**7849**	**86295**
Asterionella formosa		**(SP) 1**	**(SP) 5**	**(SP) 16**	**(SP) 1**	**(SP) 3**	**(SP) 1**	**(SP) 1**
Cyclotella cryptica		(GP)	(GP)	(GP)	(GP)	(GP)	(GP)	(GP)
Cyclotella meneghiniana		(SP)	(SP)	(SP)	(SP)	(SP)	(GP)	(SP)
Fragilaria crotonensis and Fragilaria bidens		(GP)	(GP)	(GP)			(GP)	(GP)
Gomphonema parvulum		(GP) 0.99	(GP)	(GP)	(GP)	(GP)	(GP)	(GP)
Mayamaea permitis (97% identity match)		(SP)	(SP)	(SP)	(SP)	(SP)		(SP)
Melosira nummuloide		**(SP) 2**	**(SP) 25**	**(SP) 51**	**(SP) 1**	**(SP) 6**		**(SP) 3**
Nitzschia inconspicua (98% identity match)		(GP)	(GP)	(GP)	(GP)	(GP)		(GP)
Nitzschia palea		(SP)	(SP)	(SP)	(SP)	(SP)		(SP)
Sellaphora capitata		(SP)	(GP)	(SP)	(GP)		(GP)	(GP)
Tabellaria flocculosa		(SP)	(SP)	(SP)	(SP)	(SP)	(SP)	(SP)
Additional species (100% ,match)		*Navicula_lanc*	*Discostella_pseudost*	*Discostella*	*Discostella*	*Discostella_pseudost*		*Navicula_l*
		Melosira_varians						
Additional Genus (100% match)								
		Discostella		*Discostella*				*Discostella*

2.2.4 Recommendations

- Using PEAR to merge sequences combined with a stringent quality filter in USEARCH will improve the accuracy of the USEARCH based pipelines.
- Future validation work should use known numbers of cells, forming a community with absolute numbers of each taxon. This would allow accurate representation of relative abundance to validate the whole process and improve the precision of relative abundance estimates using molecular data. It would also allow for better correction of relative abundances by applying factors related to cell volume (Vasselon and others, 2018).
- Species used in future mock community validation studies should have better taxonomic certainty. For instance, by using strains that have been used directly to add additional sequences to the diat.barcode database.
- Mock communities should ideally comprise species likely to be encountered in environmental samples.
- Mock communities should include non-diatom representatives if these are part of the study aims.

It should be noted that working with algal cultures, or PCR products from cultures, does raise potential issues for contamination. They represent a source of concentrated DNA from specific species that may contaminate environmental samples, which typically have much lower DNA concentrations. Contamination can occur in the laboratory or particularly when environmental samples are sequenced alongside mock communities. This can be mitigated by good laboratory practice and protocols. If this not possible (for example, contracting work

to a lab of unknown provenance), then mock communities should be comprised of species that will not be present in the sample (for example, marine species for freshwater work).

2.3 Comparison of HTS bioinformatic pipelines and LM for assessment of diatom assemblages

2.3.1 Introduction

Comparisons between morphotaxonomic and molecular approaches are often carried out to validate the utility of molecular approaches in replacing or complementing traditional analysis techniques. In many cases, molecular approaches perform as well as morphological analysis in predicting environmental pressure gradients (Hinz and others, 2022; Keck and others, 2022). However, there are always inherent differences, with molecular profiles being often complementary, but not identical, to morphological profiles (Keck and others, 2022; Pérez-Burillo and others, 2022). This was noted in the development of the UK DARLEQ 3 tool where microscopy and molecular methods showed strong correlation when calculating TDI scores, but the mismatch between the 2 techniques resulted in a significant proportion of sites being assigned different ecological status classes (Environment Agency 2018, 2020; and SEPA 2018). Much of the difference is due to fundamental differences in the type of data generated by the 2 approaches, but another factor that can influence taxonomic outputs from molecular data is the way the data are processed. We sought to improve agreement between the 2 methods in calculating TDIs by analysing the original data using 4 different bioinformatics pipelines (Figure 2b). Three more recently developed pipelines (UPARSE, UNOISE3 and DADA2) were compared with the original clustering algorithm (UCLUST) but with different quality filtering and taxonomic assignment applied in line with the other pipelines to be tested. These were compared with outputs from the original HTS pipeline, which classified the sequences using an older version of the diat.barcode database.

2.3.2 Methods

Sequence processing

The program USEARCH version 11 (Edgar, 2010) was used to analyse the data using UPARSE, UCLUST and UNOISE3 (Appendix 1), allowing a direct comparison of the performance of the clustering algorithm from the HTS pipeline (UCLUST) with newer clustering algorithms. All data were analysed together. Forward and reverse reads were first merged using PEAR (Zhang and others, 2014) using the default settings. Cutadapt (Martin, 2011) was then used to trim primer pairs from the data. Low quality expected error >0.5 and short sequences <200bp were removed from the fastq files using USEARCH. Sequence headers were replaced with a unique sample identifier and the fastq files converted to FASTA files. The FASTA files were dereplicated, abundance sorted and singleton sequences removed. At this point, the 3 pipelines deviated: OTUs were clustered using either UPARSE (Edgar, 2013) or UCLUST at 97%. Chimeras were filtered using UCHIME

(Edgar, and others, 2011) run against the diat.barcode v10 database as a reference. Separately, UNOISE3 was used to generate ASVs. All OTUs/ASVs were then mapped back to the original reads and a separate ASV/OTU table produced for each of the 3 pipelines.

To analyse the sequences using DADA2, demultiplexed Illumina MiSeq data were analysed in RStudio (version 2021.09.0 Build 351), implementing R version 4.0.0 within Microsoft Windows using the DADA2 package (Callahan and others, 2016). Scripts for analysing diatom metabarcoding data were obtained from https://github.com/fkeck/DADA2_diatoms_pipeline/blob/master/pipeline.R (Appendix 2). For published method see Pérez-Burillo and others, (2021). As DADA2 denoises the data per individual sequencing run, each run was first analysed separately. Primers were removed from the R1 and R2 reads using Cutadapt. The resulting R1 and R2 reads were truncated to 200 and 170 nucleotides respectively, based on 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 rate model to infer ASVs. At this point the data were saved as .rds files. RDS files were merged from different sequencing runs for the formation of finalised ASVs and removal of chimeric sequences. ASVs detected as chimeras were discarded using the function 'removeBimeraDenovo' implemented in DADA2. This final step then produced a final ASV table for further analysis. OTUs or ASVs with sequence clusters of <8 sequences and appearing in less than 3 samples were removed.

Taxonomic assignment and filtering of OTUs

To ensure consistency in taxonomic assignment all pipelines were classified in the same way. ASVs or OTUs were classified using the RDP classifier (Wang and others, 2007) against the most recent version (version 10) of the diat.barcode reference database (Rimet and others, 2019), formatted for input into the RDP classifier (http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/Diat.barcode/). The classifier assigns bootstrap confidence to each taxonomic level. All ASV/OTU tables were filtered to remove all taxonomic units with a sequence cluster size of less than 8 sequences and all taxonomic units that appeared in less than 3 samples. This was done to remove residual PCR errors or spurious taxonomic units.

The taxonomically assigned data from each of the 4 pipelines were processed as follows. Firstly, samples from lakes and those with total read counts of fewer than 500, along with all ASVs/OTUs assigned to non-diatom groups, were excluded from further analyses. Secondly, diatom ASVs/OTUs with species assignment confidence of at least 0.97 were retained as species-level assignments. Thirdly, ASVs/OTUs with species confidence of less than 0.97 but genus assignment confidence of at least 0.97 were retained as genus-level assignments. Finally, ASVs/OTUs with genus-level confidence of less than 0.97 were recorded as unknown diatoms. This filter of uncertain taxonomic assignment was set much lower in the original study (>0.90), with a change to >0.95 for further work and methods refinement. The original study also used top BLAST hit as the method of classification; it is likely that many OTUs were wrongly assigned in the original study. For the updated pipelines

read counts were normalised to sample total (relative abundance) to account for the differences in total reads (read depth, sample depth) among samples, as this method has been shown to be superior to other methods of HTS data normalisation for comparing microbiological communities (McKnight and others, 2019).

Statistical analysis

Species and genus assignments in the final HTS data from each pipeline, transformed to proportions, were harmonised to the taxonomy and nomenclature used in DARLEQ 3. This stage involved linking synonyms between the barcode database and taxon names used in DARLEQ 3 and aggregating some varieties into the nominate type. To examine the congruence between the HTS and LM data sets the assemblage data were ordinated using non-metric multidimensional scaling (nMDS), summarising the main species distributional patterns with a 2-dimensional solution. The resulting ordinations of each pipeline were then compared to that from the LM data and the congruence between the 2 ordinations quantified using Procrustes analysis (Peres-Neto and Jackson, 2001). This method rotates the HTS ordination to maximise similarity with the LM ordination. The degree of congruence is quantified by the correlation between the 2 sets of ordination scores. TDI metrics (TDI5 HTS), EQRs and WFD status classes were calculated according to Kelly and others (2008; 2020) using the R package darleq3 (https://github.com/nsj3/darleq3). The resulting metrics were then compared with similar metrics derived from LM and HTS data sets from previous analysis of the same set of samples reported in Environment Agency, 2018, 2020 and Kelly and others, 2020.

2.3.3 Results

Table 5 summarises the number of samples and ASVs/OTUs in each pipeline and original HTS and LM data sets. Numbers of assigned taxa (ASVs/OTUs matched to the barcode database with a species or genus confidence of at least 0.97), varied from 223 (UPARSE) to 271 (UNOISE) and were lower than assignments in the original HTS data set (331), likely due to the less stringent taxonomic assignment in the original study. All HTS pipelines have substantially fewer taxa than the LM data (485).

Table 5: Numbers of samples, ASVs/OTUs and assigned taxa for each pipeline and light microscopy (LM) data set

Pipeline	Number of samples	Number of ASVs/OTUs	Number of assigned taxa
DADA2	1,850	6,107	265
UCLUST	1,685	17,646	230
UNOISE3	1,687	16,605	271
UPARSE	1,685	10,401	223
HTS (Original)	1,736		331
LM	1,578		485
Common to all	1,071		

Figure 3 compares the proportion of the total diatom assemblage assigned to species, genus or 'unknown diatom' categories using the 4 different pipelines. DADA2 and UPARSE have broadly similar patterns of assignments, with over 50% of the assemblage assigned to species in 75% of samples (Figure 3, density plots, top left). For UCLUST and UNOISE3 the corresponding figures are 53% and 31% respectively. The overall proportion of genus-level assignments is broadly similar for the 4 pipelines, although there is considerable variability in the pattern of genus assignments among samples (Figure 3, scatter plots, top right). Patterns of unassigned ASVs/OTUs are broadly similar between DADA2, UCLUST and UPARSE but very different for UNOISE3 (Figure 3, bottom left).

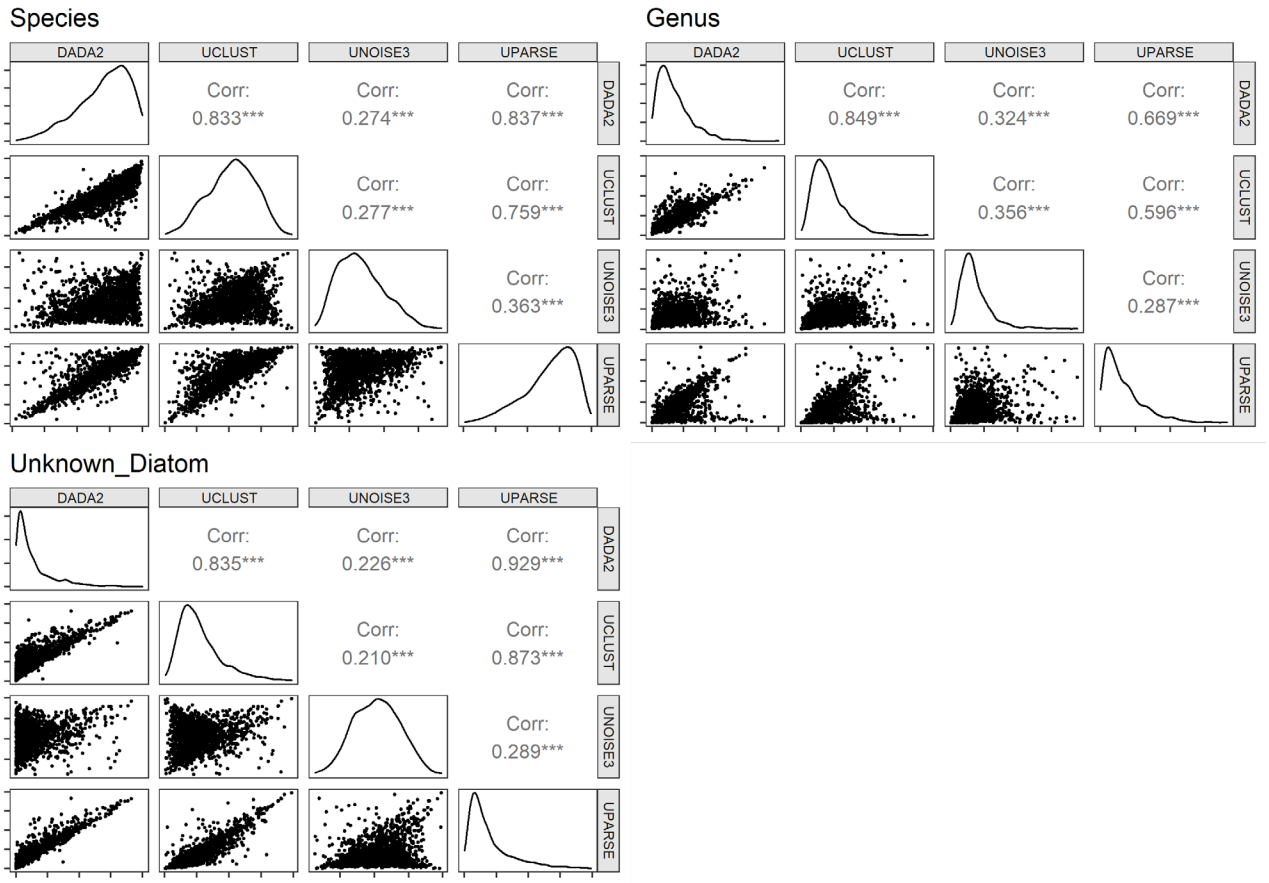


Figure 3: Scatterplot matrices comparing the proportion of the assemblage assigned to species, genera or unknown diatoms among pipelines. Density plots on the diagonal show the distribution of each group of assignments. x- and y-axis scales pan 0 to 1.0

Table 6 lists Procrustes correlations between ordinations of the LM and HTS data sets generated from each of the pipelines. Except for UNOISE3, correlations for the original HTS and new pipelines are all high (around 0.83), suggesting that these pipelines capture the same species/sample relationships that exist in the LM data. The correlation for UNOISE3 is somewhat lower (r=0.72).

Table 6: Procrustes correlations between ordinations of LM and HTS data sets

Pipeline	Procrustes correlation
HTS	0.837
DADA2	0.833
UCLUST	0.833
UNOISE3	0.720
UPARSE	0.828

Figure 4 summarises the relationships between TDI5 scores for the different pipelines and LM data sets. There is very close agreement between TDI5 scores produced from DADA2, UCLUST, UPARSE and the original HTS pipeline (including the use of the older barcode reference database) ($r > 0.95$). UNOISE3 is again a clear outlier with correlations with other HTS pipelines of <0.8. Correlations between LM TDI5 scores for DADA2, UCLUST and UPARSE are very similar ($r = 0.845$ to 0.857) and lower for UNOISE ($r = 0.705$). Correlations between DADA2, UCLUST and UPARSE and LM TDI5 scores are slightly lower than that for the original HTS pipeline ($r = 0.872$), which is expected as the TDI5HTS metric was tuned to maximise the correlation with TDI5LM using this pipeline. Correlations between TDI scores and measured nutrient pressure are highest for LM ($r=0.77$) and lowest UNOISE3 ($r=0.53$).

Table 7 shows the extent to which species assignments and the resulting TDI scores are influenced by the confidence threshold at which ASVs/OTUS are assigned to species. Relaxing the threshold from 0.97 to 0.93 leads to an additional approximately 20 ASVs/OTUs being assigned to species and increases the correlation between TDI scores slightly from around 0.85 to 0.87.

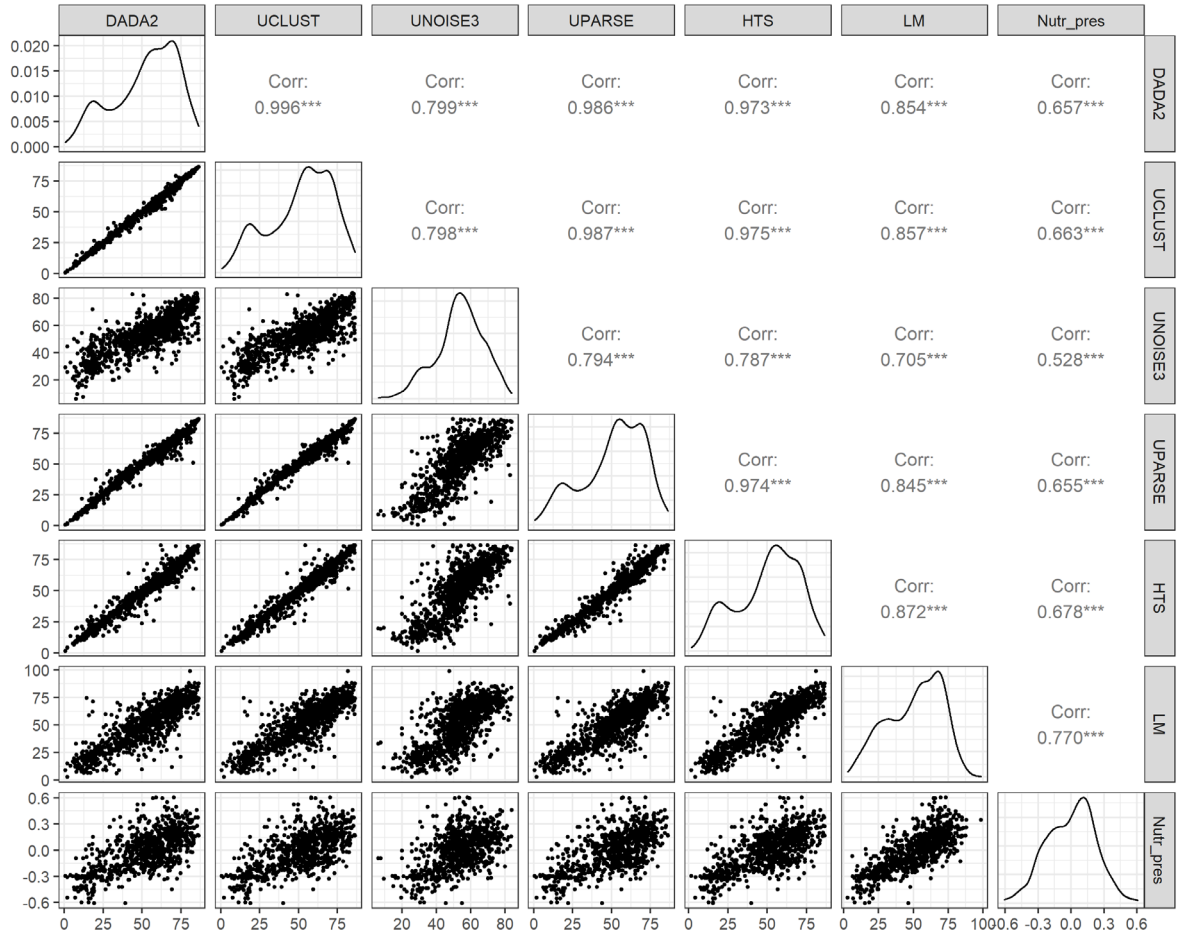


Figure 4: The relationships between TDI5 scores for the different pipelines and LM data sets and a composite N and P nutrient pressure variable. Density plots on the diagonal show the distribution of each group of variables

Table 7: Numbers of assigned species and correlations between HTS and LM TDI5 scores using different species assignment thresholds

Pipeline	Threshold =0.97		Threshold = 0.93	
	No. species	r	No. species	r
DADA2	265	0.854	287	0.867
UCLUST	230	0.857	249	0.871
UNOISE3	271	0.705	294	0.713
UPARSE	223	0.845	242	0.876

Figure 5 compares the site-based WFD status classes for the 5 HTS pipelines against that derived from light microscopy. Converting the TDI values to EQRs using the current UK

reference model to predict 'expected' TDI shows how these differences will convert to changes in ecological status. Except for UNOISE3, the different pipelines have very similar patterns of agreement with LM, with 62 to 66% of sites allocated to the same status class as that derived from LM, and 96 to 98% allocated to the same or neighbouring class (Table 8). UNOISE3 is again the outlier, with only 43% of sites allocated to the same status class (Table 8).

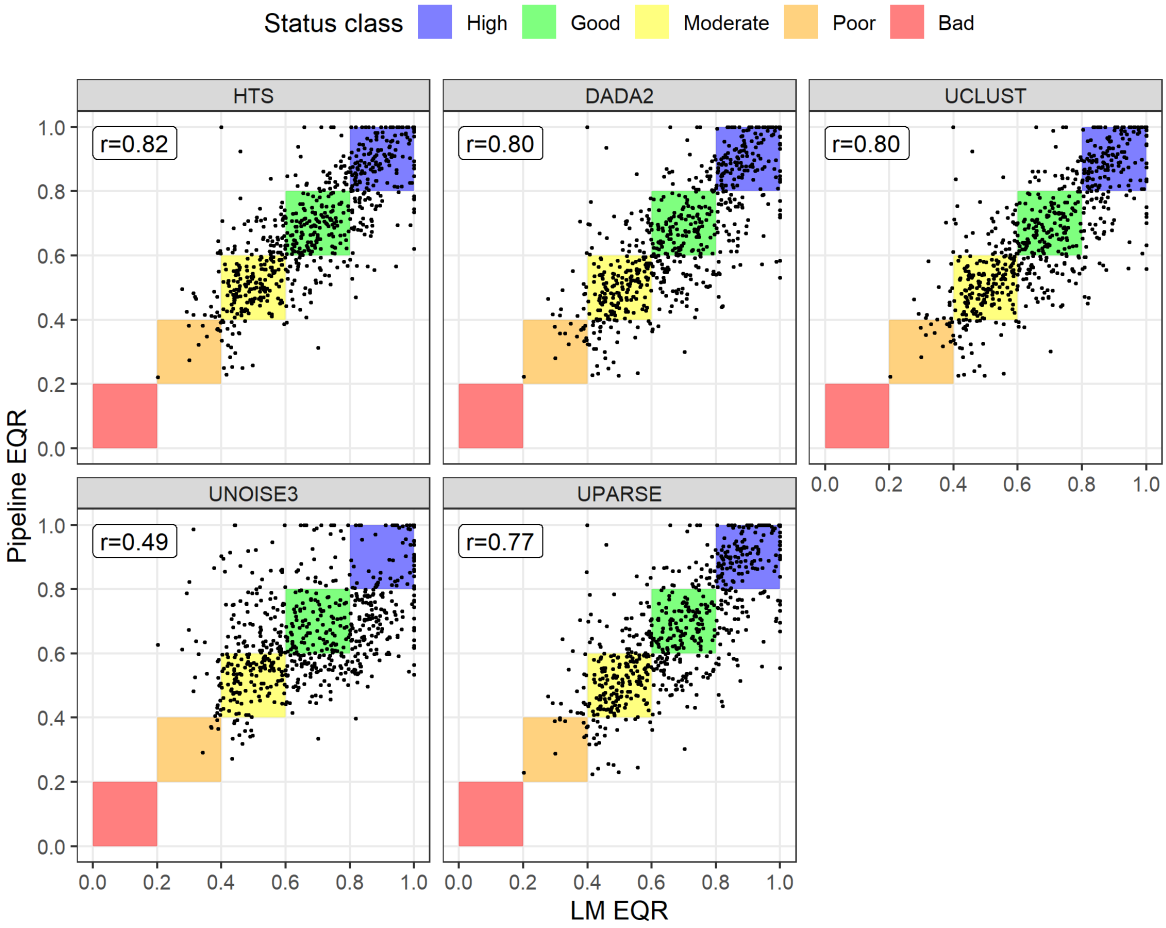
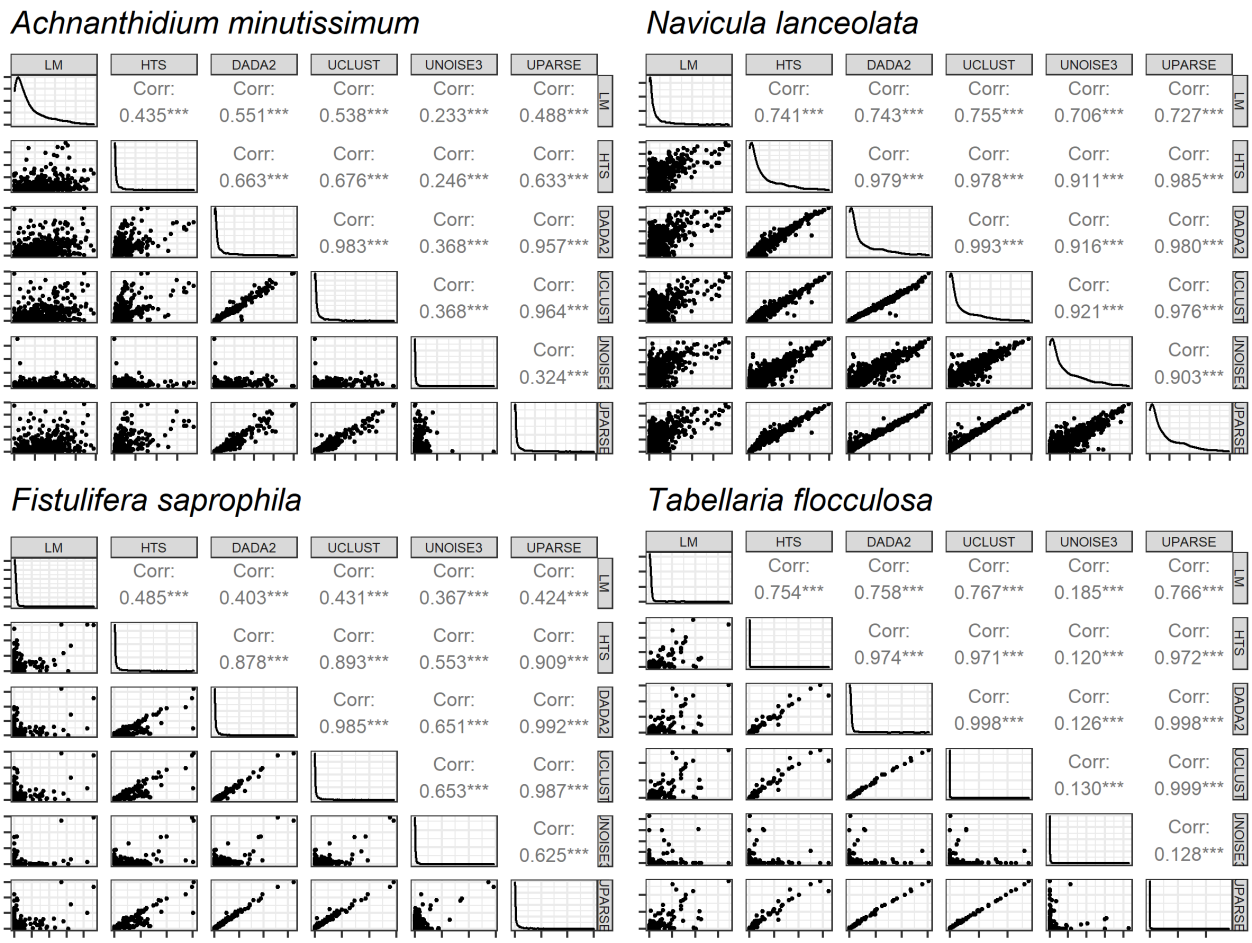


Figure 5: Comparison of site-based EQRs for each HTS pipeline (y-axis) and LM (x-axis) (N=675)

Table 8: Agreement between site-based predictions of WFD status class for each pipeline and status class derived from light microscopy data. Numbers indicate percentage of sites (N=675)

Agreement	HTS	DADA2	UCLUST	UNOISE3	UPARSE
Same class	65.8	64.7	65.5	42.7	62.2
1 class	31.9	32.0	31.3	48.7	34.2
2 classes	2.2	3.1	3.1	7.9	3.3
3 classes	0.1	0.1	0.1	0.7	0.3

The relationships between the representation of a range of common taxa in LM and the different pipelines are shown in Figure 6. The relationship with LM is always poor, due to the fundamentally different nature of LM and HTS data; more surprising is the extent of variation among the pipelines, bearing in mind that they are sorting the same sequence reads using the same reference database. Once again, UNOISE is a clear outlier, with generally low correlations with other pipelines. There are also differences among other pipelines. For example, DADA2 and ULCUST generally give high correlations for most taxa, but there is a low correlation between proportions of Tabellaria flocculosa detected with these pipelines.



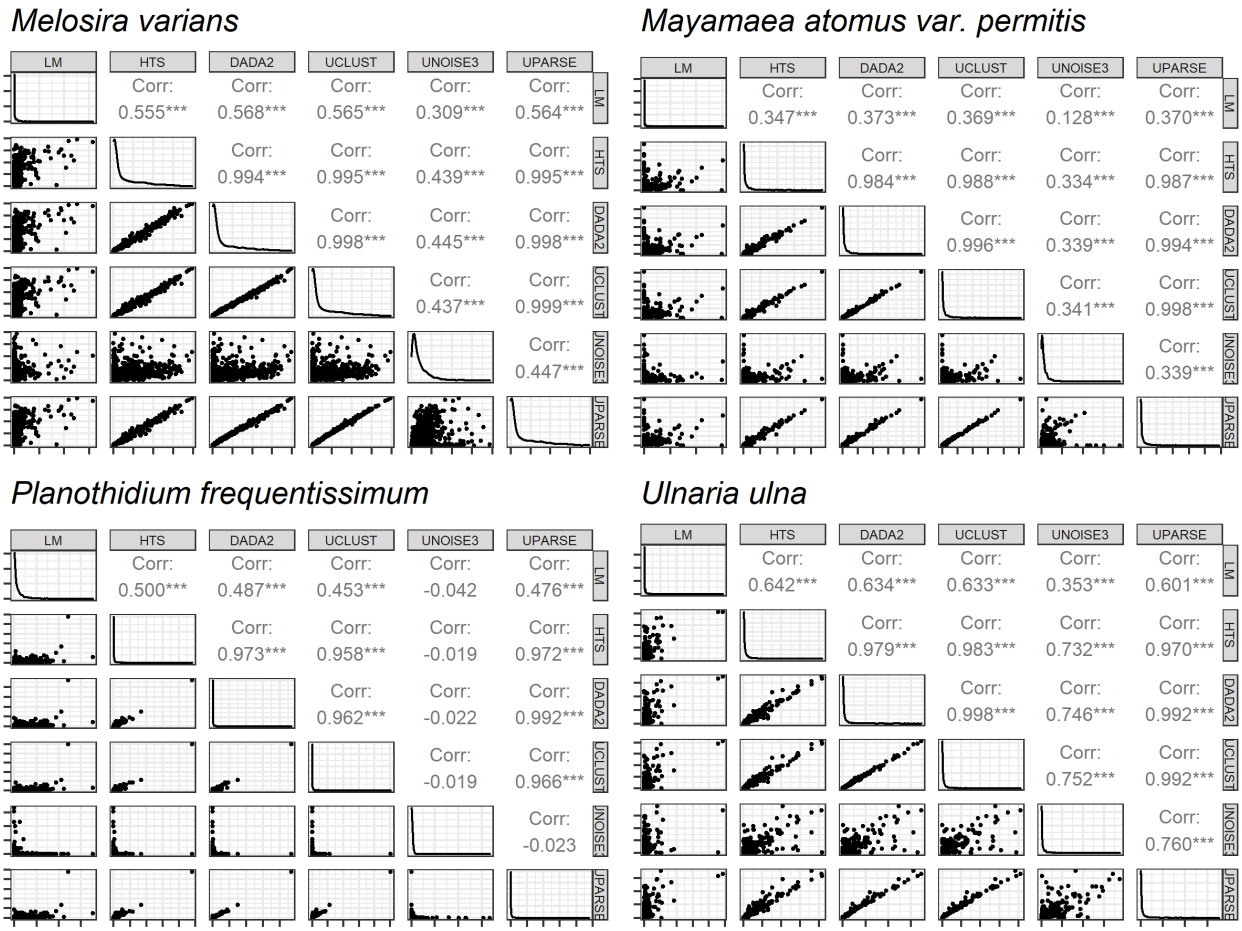


Figure 6: Relationships between the proportions of selected taxa for each pipeline and LM data sets

2.3.4 Discussion

The HTS pipeline used for the original development of the diatom metabarcoding approach (Environment Agency, 2018) was state-of-the-art when it was written, but subsequent developments in bioinformatics, particularly the benefits of ASVs over OTUs (Pérez-Burillo and others, 2021) have made this approach no longer the preferred standard in analysis of metabarcoding data. Furthermore, the OTU clustering algorithm used by QIIME- UCLUST has been consistently outperformed by similar OTU clustering algorithms for accuracy and precision. QIIME 1 has been surpassed by QIIME2 and QIIME1 is therefore no longer supported. Additionally, QIIME2 does not implement many of the options used in the original pipeline in its default analysis. This means that a revision to the pipeline was overdue. This section compares outputs from the original HTS pipeline with 4 newer alternatives, and with LM results.

For regulatory purposes, a seamless transition between old and new methods is desirable, with either no significant change in outcomes, or an ability to explain and account for observed changes. In statistical terms, this means that analyses are treated as ‘type 1’

regressions, where the existing approach ('X') is assumed to be without error and all variability is due to the new approach ('Y'). In reality the situation is better considered as a 'type 2' regression, with error shared between X and Y. Individual plots in Figure 5 and Table 8 showing status class agreement between approaches, are likely to show inflated estimates of the number of sites that have changed class specifically because of the change in method. It is also reasonable to assume that some of those sites that have changed class could have been wrongly classified using the microscopy-based method. This is acknowledged in the use of 'confidence of class' in formal ecological status assessments, but this nuance could not be incorporated into the present study. Because the mode of quantification in LM and HTS (direct counts versus relative abundance of sequence reads) is completely different, it is unrealistic to expect perfect agreement between the 2 approaches. While LM counts valves (= half a frustule/cell wall), HTS counts sequence reads of rbcL which are related to cell volume (Vasselon and others, 2018). This explains at least part of the variability between LM and HTS reported by Kelly and others (2020), Pérez_Burillo and others (2020) and, more recently, Kulaš and others (2022) as well as in Figure 6 in this report. The extreme examples of the impact of cell volume are large-celled species such as Melosira varians and Ulnaria ulna, where a small number of valves reported in an LM analysis may translate into an overwhelming dominance of rbcL sequence reads in HTS. Kelly and others (2020) argued that this meant that HTS, by quantifying a property related to a key photosynthetic enzyme, offers a better indication of the contribution made by each species to primary production than LM. These studies also suggest an underrepresentation of small, weakly-silicified taxa such as Fistulifera saprophila and Mayamaea permitis in LM, most likely due to their dissolution during preparation for light microscopy.

These potential sources of difference between LM and HTS do not invalidate either approach but do need to be considered when comparing outputs. Furthermore, diatom metrics are based on proportional representation of individuals subject to the properties of multinomial distributions. This means that a sharp increase of Melosira varians sequence reads in HTS relative to LM will lead to a decrease in the proportions of many other diatoms in the same sample. This, in turn, will have consequences for the values of metrics calculated from these data. While the link between cell volume and rbcL read number has not yet been tested on field data, we believe this to be a major reason for observed differences greater than one WFD class.

Using this information and other published sources, the variation observed in Figures 3-6 can be partitioned into the following categories:

a) Statistical

- Repeated measurements on the same sample are unlikely to generate exactly the same values ('repeatability' - see Environment Agency, 2020). Nonetheless, this is unlikely to account for any differences greater than one class.
- Variation due to the mode of quantification – likely to contribute to differences both less than and greater than one class.

b) Molecular

- Variation due to gaps in the reference database – Figure 10 from Kelly and others (2020) suggests that this source of difference is, again, too small to account for differences greater than one class and the reference database has grown since this graph was derived, meaning that impacts should have reduced even further.
- There are known issues with molecular work, including biases in DNA extraction or incomplete cell/chloroplast lysis, biases in PCR amplification for certain sequences, biases in sequencing for certain sequences, PCR primer bias, and PCR primer exclusion of certain taxa. Cumulatively, these are likely to have the biggest impact.
- Variation due to some diatom species with larger cells with higher biovolume having more rbcL copies than smaller cells.

c) Laboratory practice

- Variation due to potential contamination, cross-sample mixing during processing and molecular work, or contamination of samples from positive controls or previously processed samples. This is likely to affect only a minimal number of samples and will be controlled by good sampling and laboratory protocols.
- Variation due to sample mislabelling either physically or within the sequence data (multiple steps in sample label transfer presents a high risk of this, for example, FERA ID to sample ID conversion). This is likely to affect only a few samples, but the scale of effect could be large.

All of the above sources of variation, although small individually, will have a cumulative effect on the data set. In brief, considerable variability between LM and HTS is to be expected; what is more of a surprise is the level of difference observed between the performance of individual pipelines. Despite these differences and the fact that all pipelines are analysing the same data set with the same biases and issues, these differences, in most cases (UNOISE3 is the exception), do not translate into major differences in index values (Figure 3). However, these strong correlations conceal some major differences in the composition of the assemblages. Noteworthy too is the relatively good performance of the original HTS pipeline. It should be borne in mind that development of the HTS TDI (TDI5HTS) was 'tuned' to the characteristics of this particular pipeline, giving it an inbuilt advantage. While it may be questioned whether a switch from the original pipeline is justified, it should be noted that the original HTS pipeline and the UCLUST pipeline (which is the same clustering algorithm used in the original pipeline) performed poorly on the mock community analysis, generating thousands of taxonomic units for just 11 species. In the data from environmental samples, the impact of this appears to have been smoothed because the taxonomic units were collapsed to the species level, and because of the high volume of (poor quality) data the HTS pipeline incorporates. However, any extra information that may be obtained from the data (see section 3) is completely lost when using the HTS pipeline, and individual differences in taxonomic units for species between samples will likely be a result of erroneous sequences rather than true biological variation. To our knowledge, UNOISE3 had never been used for diatom rbcL data before, and its performance was poor compared to the other methods. This is contrary to work done for 16S rRNA gene data (Nearing and

others, 2018) which has shown better performance in mock communities and environmental samples for bacteria. The algorithm was designed and tested on 16S analysis and it would appear, perhaps because rbcL is a protein coding region, that UNOISE3 does not perform well in generating biologically accurate ASVs for the rbcL gene.

Overall, the results of this study are similar to those of Baillet and others (2020) who also explored taxonomic composition in detail, demonstrating considerable differences in the allocation of species within the genera Fragilaria and Eunotia*,* despite the same reference database being used with all pipelines. These results caution against simplistic interpretations in terms of good fits with ecological assessment metrics and point to issues in correctly characterising the diatom assemblage itself. More work is needed to decide which pipeline, and which settings within each pipeline, are most appropriate. Meanwhile, a pragmatic way forward is to consider the pipeline to be an integral part of the assessment protocol, not as a distinct preliminary stage.

We recommend using DADA2 over the other pipelines for both scientific and practical reasons. Importantly, the performance of DADA2 on the mock communities was good compared to the other pipelines, generating 52 ASVs in total and classifying 10 ASVs to species level (of an 11 species mock community). It is likely that using ASVs will enable finer scale detection of changes in assemblages in response to environmental pressures. They also make it possible to look at potential population level differences that would be missed by morphological analysis alone (Pérez-Burillo and others, 2021; 2022). The error correction that DADA2 uses on an individual sequence run basis provides more confidence in the sequence data and allows for better comparison between sequence runs. This was particularly important for this project where there was high variability in data quality and quantity between runs. Another important reason for selecting this pipeline is that other researchers in continental Europe have also been using it for diatom metabarcoding data (Apothéloz-Perret-Gentil and others, 2021; Pérez-Burillo and others, 2021; 2022) allowing for pan-European comparisons in diatom responses to pressure gradients. As the research community further develops the pipeline for diatom metabarcoding and develops better metrics and indices for molecular monitoring of ecological status, it will be important to promote a unified approach.

DADA2 is becoming a standard pipeline in microbial ecology and because of this it is highly likely to continue to be maintained and supported. It has been stable since 2016 (no DADA3 released) and although updates have been released, the main features remain unchanged. This pipeline is implemented in the R programming language which can be run in the program Rstudio (Rstudio team 2020) on a standard Windows machine. Since DARLEQ 3 is also written in R, bioinformatics and metric calculation could be merged in the future. This makes it easier to implement than a Linux based pipeline. DADA2 also requires less memory than the other pipelines and is fully open access. While the UPARSE pipeline was comparable in some respects to DADA2, to analyse very large data sets a 64bit version of USEARCH is required; this is a paid-for version which is limited to a single machine. Despite reports in the literature that DADA2 was slower than other pipelines, for this study, total

processing times were similar for DADA2, UPARSE and UNOISE3 (approximately 24 hours each). However, we also emphasise that, just as bioinformatics have moved on since the original QIIME HTS pipeline was written, they are likely to continue to evolve. How to incorporate new scientific developments into assessment approaches without compromising the need for stable methods for regulation needs to be considered.

All the pipelines found many diatom taxa with uncertain taxonomic identity. For example, they were either identifiable with confidence to genus level only or identified to higher taxonomic levels. While at present there is no alternative to using Diat.barcode as the DNA reference database, the ASVs themselves can be converted progressively into a reference data set. Several thousand ASVs are already known and could potentially be added to the reference database once their identities have been established with an acceptable degree of certainty. There is scope for doing this by careful comparative analysis of the matched HTS and LM data, by identifying samples with unusually high abundances of particular species in cell counts (cf. Rimet and others, 2018) or, more generally, by correlating ASV and LM relative abundances. Subsequently, exact matches could be used for identification, rather than the somewhat arbitrary percentage cut-offs currently applied. It is likely that relatively few previously undetected ASVs will be found in future sampling campaigns, unless new ranges of waterbodies are sampled.

2.3.5 Recommendations

- LM should not be treated as a benchmark for testing HTS data unless biovolumes are also available.
- DADA2 is our recommended pipeline. We believe it to be an accurate, stable pipeline that allows for more accurate and detailed analysis of diatom assemblages. DADA2's widespread use is also likely to improve user confidence. This pipeline is unlikely to change within the next 3 to 5 years.
- More work is needed to understand the relationship between outputs from bioinformatic pipelines and the biological communities they represent (improvement to mock community analysis, see section 2.2.4).
- Develop a unified catalogue of identified ASVs to move towards identification by exact matching. We can use ASVs to build an accurate local reference database, particularly for those that don't have a good reference match in diat.barcode.
- Establish protocols for data handling. Regarding sequence data we strongly recommend:
    - a unified sample labelling strategy for sequence data with no variation in the numbering from sample collection through to generating fastq files
    - a standard labelling system for sequence runs
    - a complete data archiving strategy for sequence data (upload to National Centre for Biotechnology Information (NCBI))
    - storing sequence runs/projects as zipped/tarball archives (reduces file size and improves sample transfer and transport time)

- a better catalogue of repeated samples and informative labelling for any repeats
- a better link of sequence data files to meta-data
- a detailed catalogue of associated run meta-data reads per sample, quality information

3. Maximising the potential of rbcL metabarcoding data

3.1 Objectives

Although the previous section recommended going forward with DADA2, there was no improvement in the strength of the relationship with the underlying pressure gradient (Figure 4) for the reasons set out in section 2.2.4. The mismatch between LM and HTS remained. All pipelines examined in phase 1 assigned only about 60% of the total sequence to known diatom taxa. However, it is possible that the unassigned reads, many of which belong to non-diatom algae, will add extra ecological information and, as a result, increase the strength of the relationship between phytobenthos and important environmental variables, or provide new insights regarding ecological function.

The objectives of phase 2 of this project were to explore HTS data to determine the potential to extract wider information on the river phytobenthic community and to examine relationships with environmental variables to assess the potential for new approaches/new metrics for the assessment of river phytobenthos.

3.2 Analysis of non-diatom sequences within rbcL gene fragment metabarcoding libraries

3.2.1 Introduction

The UK rbcL primers had previously been shown to detect other algal groups in addition to diatoms. At the time of the original work there were very few reference sequences for these groups and so further detailed taxonomy was not possible for many of the OTUs. Given the high number of OTUs generated in the original study, it was not feasible to manually investigate the taxonomy of the main groups. The generation of ASVs produces a much more manageable number of sequences to investigate the taxonomy of non-diatoms in the data set. Since the original projects were completed, new rbcL sequences for different algal groups have been added to the NCBI nucleotide database. However, there still is no curated database for other algal groups as we have for the diatoms in the diat.barcode reference database. Given what is known about the rbcL gene and the primer sequences used to generate previous data sets, we expected several taxa may be amplified by the primer set (see Table 1). We therefore sought to investigate the taxonomy

of non-diatom ASVs in a data set generated from the data used in phase 1 of this project. The aims in this section of the project were to:

- taxonomically classify non-diatom ASVs within a large rcbL metabarcoding data set, generated from environmental samples using the UK diatom 'specific' primers
- investigate the feasibility of a non-diatom reference database

3.2.2 Methods

To investigate the representation of non-diatoms in the HTS data set a subsample of 10 sequencing runs was prepared. The data set was analysed with the DADA2 pipeline and taxonomically assigned with the RDP classifier following the same methods as above (section 2.3.2). Non-diatom ASVs were extracted from the total list using the criterion of failing to achieve 100% bootstrap support for classification in Bacillariophyta at phylum level. Over 8,000 'non-diatom' ASVs were detected. Some were artefacts or pseudo-genes, despite the use of the denoising and chimera detecting algorithms in the DADA2-based pipeline. This was shown by inspecting the amino-acid sequences coded by the ASV sequences, which showed that some contained stop codons despite the barcode region being well within the gene, or implausible amino-acid substitutions. To minimise these residual errors, we worked only on ASVs represented in more than 2 samples (out of approximately 1,825 samples in the data set used) and we focused on the most abundant 1,000 of these to examine the representation of different non-diatom groups.

3.2.3 Results and discussion

Non-diatom ASVs were in general much less abundant than diatom ASVs. For example, the 5 most abundant non-diatom ASVs (Ulvella cf. tongshanensis, Heribaudiella fluviatilis*,* Chlioscyphos polyanthos, Oedogonium sp. and Diplosphaera chodatii) were ranked 141, 180, 218, 220 and 258, in order of total read abundance across the 1,825 samples.

The rbcL primers developed by the Environment Agency (2018, 2020) were designed to capture and quantify diatom diversity. However, rbcL is a highly conserved and functionally vital gene, which means that a wide range of other photosynthetic organisms are also amplified by the primer set. Among the 1,000 most abundant ASVs we analysed, there were representatives of almost all the autotrophic phyla known to occur in freshwaters (Table 9), including both eukaryotes and prokaryotes. The extent of this 'contamination' was rather surprising, however, because the groups detected span both of the principal lines of rbcL evolution in eukaryotes. These are both the the line represented by the green plants (Viridiplantae = Plantae) and Euglenophyta, in which RuBisCO was derived from Cyanobacteria, and the 'red lineage' of chloroplasts (Rhodophyta, Haptophyta, Cryptophyta and Ochrophyta), which acquired a proteobacterial rbcL by horizontal gene transfer after the primary endosymbiotic event that created the eukaryotic chloroplast (Delwiche & Palmer

1996). The wide spread of taxa that can be amplified with the ‘UK’ primers must either reflect conservation of ancestral sequences in the primer-binding regions in many distantly related lineages, or convergent evolution in a gene where there are limited possibilities for variation without loss of enzyme function.

Assessing bias for certain taxonomic groups

It is impossible to know, without laboratory tests of amplification success with mock communities, what relation the numbers of reads recorded bears to the abundance of organisms in environmental or other samples. Nevertheless, some non-diatom groups are rather well amplified and probably well quantified (these include Cryptophyta, Eustigmatophyceae and some green algae), whereas others are much less abundant than would be expected in the natural communities sampled. For example, it is unlikely that no Synurophyceae (for example, Mallomonas, Synura) and very few Chrysophyceae were present in any of the samples, and they should have been detected given the number of samples collected. The almost total absence of Synurophyceae and Chrysophyceae is likely to be because of lack of amplification. Indeed, inspection of the primer binding sites shows multiple mismatches in the forward primer for these 2 groups, even though the amino-acid sequence is conserved at this point. In addition, Cyanobacteria are not represented very strongly among the ASVs but will have been present in many of the habitats sampled. On the other hand, some ‘unlikely’ groups were recorded, including Raphidophyceae and various riparian angiosperms, which can only have been present in trace amounts (they will not have been growing on river cobbles and other hard substrata). However, in around 15% of sites, where sampling is not possible from cobbles, macrophytes are sampled, so some of the angiosperm sequences could have come from these.

Underrepresentation of some groups of green algae reflects unusual chloroplast genome organisation, notably in Cladophorales (Del Cortona and others, 2017). Probably the only detectable member of this order will be Chaetomorpha linum, which is relatively uncommon in rivers. In other green algae and euglenophytes, introns interrupt the barcode region, for example, in some Chaetophorales (a search in the NCBI GenBank gives examples in Uronema confervicola MN701586, Draparnaldia mutabilis MN659372) and in Euglena (Koller and others 1984). Nevertheless, there are some Chaetophorales among the ASVs and these species must either not possess introns or the introns are outside the region amplified. The only way to detect algae with rbcL introns would be to use RNA, rather than DNA. Many green algae, however, do not possess introns and are common among the non-diatom ASVs, especially unicellular and coenobial species of the Chlamydomonadales, Scenedesmaceae and Trebouxiophyceae.

Table 9. Numbers of ASVs and reads attributable to the major groups of photosynthetic eukaryotic microalgae and bacteria among the 1,000 most abundant ASVs in the data set analysed (2014 to 2017 reads). The figures are illustrative, to give an impression of the representation and diversity of the groups in the UK metabarcoding data set: no attempt was made to standardise read numbers among samples. Some diatom ASVs were present in the data set, having failed the criterion for exclusion from the non-diatom data set (100% bootstrap support for classification in 'Bacillariophyta' from the naïve classifier, using the Diat.barcode v. 10 reference database); these comprised 27 ASVs and 14,030 reads.

Major group	Family/subgroup	ASVs	Reads
Green plants (Viridiplantae)	Chlorophyta	648	1,761,699
	Streptophyte green algae	19	58,752
	Marchantiophyta+Bryophyta	3	78,749
	Angiosperms	5	3,379
Euglenophyta		19	62,098
Rhodophyta		4	51,554
Haptophyta		4	6,189
Cryptophyta		75	165,297
Ochrophyta	Chrysophyceae	1	360
	Dictyophyceae	5	10,313
	Eustigmatophyceae	75	159,252
	Phaeophyceae	5	128,102
	Phaeosacciophyceae	1	550
	Phaeothamniophyceae	1	301
	Raphidophyceae	2	643
	Xanthophyceae	35	62,793
	Total Ochrophyta	125	362,314
Cyanobacteria		67	111,052
Proteobacteria		5	2,295

Potential to develop a reference database

There is no off-the-shelf rbcL reference database for algae apart from diatoms (in Diat.barcode) and the existing rbcL sequences in GenBank need curation to try to eliminate incorrect identifications and modernise taxonomy. There is circumstantial evidence that some of the identifications in GenBank have been obtained by matching sequences to those already in GenBank, rather than by independent identification from morphology. Consequently, value judgements must be made about the trustworthiness of particular

GenBank accessions. Moreover, availability of reference sequences is patchy across the range of phyla and classes represented among the ASVs, largely reflecting differences in which genes have been chosen for phylogenetic analysis and barcoding by taxonomists. In red and brown algae (Rhodophyta and Phaeophyceae), for example, rbcL is often used and so there are many reference sequences available, but these groups are not strongly represented in freshwaters. There is active research into the Eustigmatophyceae using rbcL as a marker of choice, and so the reference database for this group can be expected to become fuller in the near future, and Cryptophyta are also quite well covered. There are many rbcL sequences for green algae (Chlorophyta and Streptophyta). However, some green algal systematists prefer the internal transcribed spacer (ITS) regions of rDNA for species differentiation. As a result, many green algal ASVs do not have close matches in GenBank. The reference database for Cyanobacteria is poor. While Cyanobacteria were fairly abundant among the non-diatom ASVs relative to other non-diatom groups, very few cyanobacterial ASVs had close matches in the reference database, although some of the assignments were plausible (for example, Chamaesiphon).

To produce a reference database equivalent to that of diat.barcode for non-diatom taxa would require a significant amount of work based on reference sequences from cultures of a number of taxa.

Unusual records of algae

Regardless of whether it is valuable to use non-diatom data for ecological assessment (cf WFD ecological status assessments), the ASVs detected using the UK rbcL metabarcoding protocols provide valuable information about the diversity and distribution of algae present in UK rivers. This is especially true for groups which are difficult to identify, such as the green alga Oedogonium (morphological identifications depend to a considerable extent on resting spore morphology, and resting spores are observed only rarely) and the many coccoid and coenobial representatives of the green algal classes Chlorophyceae and Trebouxiophyceae.

Among the ASVs analysed there are some interesting finds. The most striking and clearly established new record is of the brown alga Bodanella lauterbornii, known previously from only a handful of subalpine lakes in central Europe, including Lake Constance, where it occurred deeply submerged (10 to 30m) on steep limestone cliffs (Schütz and others, 2021). Its occurrence in UK rivers is therefore very unexpected, its distribution and requirements need investigating, and it may deserve a conservation assessment.

Among the red algae, one ASV is a 100% match to Nemalionopsis shawii. This genus appears never to have been recorded before in European freshwaters; the few existing records coming from Asia and North America. The most abundant red algal ASV has no close match in GenBank (the nearest is Madagascaria erythrocladioides). It is therefore possibly an unsequenced member of the Compsopogon group.

3.2.4 Recommendations

- While numerous taxa are amplified by the primers, these are in much lower relative abundance than the diatoms, and there are notable absences in the data set of algal groups that we may expect to be present in the phytobenthos. If some of these groups need to be targeted, the primers either need to be modified, or redesigned entirely, to produce another set of group specific primers (for example, Chlorophyta or Cyanobacteria) to target an algal group of interest.
- Work is required to develop a reference database for non-diatom ASVs, or potentially use the ASVs themselves as reference sequences with a basic taxonomy included.
  .

3.3 Maximising the potential of rbcL metabarcoding ASVs

3.3.1 Introduction

As was both expected and outlined in the section above, the primers used to amplify diatoms also amplify a wide range of other algal groups. Many of the groups, in addition to the diatoms, may show strong responses to environmental pressure gradients. Being able to make use of these data potentially maximises the amount of information we can harvest from rbcL metabarcoding data sets. However, for many of the groups, we know little about their ecology, as differentiation of morphospecies using light microscopy is difficult. Molecular techniques obviously make the differentiation between different species possible. There are also many diatom taxa that are not identifiable to species level in the molecular data set, as there is no reference sequence for these taxa in diat.barcode. These potentially represent undescribed species or cryptic within-morphospecies diversity. For example, in the case of some small Navicula species, distinguishing between different taxa is impossible without scanning electron microscopy. These ‘unidentified’ diatoms may also show important relationships with environmental gradients. In this section, we sought to investigate the relationship of non-diatom ASVs and all diatom ASVs with environmental pressure gradients. The aim was to investigate to what extent both non-diatom ASVs and all diatom ASVs are good predictors of environmental pressure gradients. This analysis was performed on a large sample set of ~900 river phytobenthos samples collected from across the UK that have both rbcL metabarcoding data and nearby environmental water chemistry.

3.3.2 Methods

Sequence processing and taxonomic classification

Samples used included all data from phase 1 of this project (2014, 2015) as well as samples collected and analysed as part of routine monitoring in 2017, 2018 and 2019 phase 2

samples). All samples were analysed using the DADA2 pipeline (see section 2.3.2, and Appendix 1 for details). A total of 45 sequencing runs with over 5,000 individual samples (including PCR positive/negative mock communities and the environmental samples) were analysed. Each run was analysed individually, and the final outputs saved as an rds file before chimera removal and generation of final ASVs on a merged data set of all 45 sequencing runs (this step took over 36 hours to run, on a 64Gb RAM machine, using a maximum of 24Gb of RAM, with 6 CPU cores). Taxonomy was assigned as above (section 2.3.2) The data set was then split for further analysis into diatoms and non-diatoms. In subsequent analyses of these 2 subsets a taxonomy-free approach was adopted; ASVs that showed significant responses were further classified against the full NCBI database.

Data processing and sample matching

Sample metadata for phase 2 HTS samples contained information for 2,556 unique samples from 901 sites. Table 10 summarises the number of samples by waterbody (river or lake) and monitoring purpose. Samples from lakes and those collected from macrophyte or man-made substrates were excluded from subsequent analysis.

Table 10: Numbers of new phase 2 samples by waterbody type, showing those sites where water chemistry monitoring was taken from the same or different waterbodies and sampling purpose

	Number of sites	Number of samples
Total	901	2,556
Lake	51	181
River	850	2,354
Monitoring (same water body)	469	1,313
Monitoring (different water body)	53	138
Investigative	333	924

Water quality data were extracted from the Defra water quality data archive (https://environment.data.gov.uk/water-quality/view/landing). Given the variation in the frequency of water quality determinations, and the range of distances between water chemistry and biology sampling locations, biology and chemistry were linked manually using a custom-written R Shiny (https://shiny.rstudio.com/) map-based application as follows. Firstly, each biology site was plotted together with the closest, and other nearby, chemistry monitoring sites. Locations of potential pollutant discharges were obtained from the list of consented discharges to controlled waters (https://data.gov.uk/dataset/55b8eaa8-60df-48a8-929a-060891b7a109/consented-discharges-to-controlled-waters-with-conditions) and Environment Agency compliance monitoring sites. The remaining biology samples (not lakes, macrophytes or those with no close nutrient data match) were matched to the nearest chemistry monitoring site that had at least 4 total reactive phosphorus determinations in the

sampling year, was less than 1,000m away, and had no obvious sewage or other discharges between biology and chemistry sites. For each of the matched chemistry sites, reactive phosphorus (P), nitrate (NO_3-N), ammonium (NH_4-N), alkalinity, conductivity and pH values were extracted and expressed as annual means (arithmetic for pH, geometric for other variables). Table 11 summarises the number of biology samples by distance to matched chemistry site, and the number of individual reactive P determinations in the sampling year. Reactive P sampling at many locations is infrequent (<6 times per year) and many sites either had no reactive P data or could not be matched to a suitable chemistry site within 1,000m.

Table 11: Number of HTS samples with matching chemistry as a function of distance to matched chemistry site and number of reactive P determinations in the sampling year

Distance to matched chemistry site	1-5	6-11	>= 12	No reactive P data
<300m	524	281	226	344
300 to 1,000m	72	37	45	71
>1,000m or not matched				754

Of the 2,354 river samples, 1,699 could not be matched to a water quality monitoring site within 1,000m distance and with at least 4 reactive P determinations in the sampling year, so were omitted from further analyses. Given the reduced number of samples with matching environmental data, the phase 2 samples were merged with the existing data set from phase 1 to give a total of 1,688 samples. From this total, samples with a total diatom read count of less than 500 and a total non-diatom read count of less than 100 were removed. This screening yielded a final HTS data set of 1,220 samples with matching chemistry data (750 phase 1, 470 phase 2), containing 4,036 diatom ASVs and 3,187 non-diatom ASVs. The data were then split into diatom and non-diatom subsets and each normalised to the total diatom and non-diatom reads respectively. All analyses are based on the screened data set of 1,220 samples unless otherwise stated.

Statistical analysis

Statistical analyses were performed to (1) quantify the magnitude and test the significance of the response of the diatom and non-diatom assemblages to the nutrient pressure gradient, and (2) examine the degree to which these assemblages can be used as indicators of nutrient pressure. In all analyses, nutrient pressure is quantified as a combination of reactive P and NO_3-N (derived from the first axis of a principal components analysis of these 2 variables). The combined pressure gradient explained more variance in the biological communities than reactive P or NO_3-N alone.

The magnitude of compositional turnover or beta-diversity along the nutrient pressure gradient was quantified using detrended canonical correspondence analysis (DCCA), with nutrient pressure as the constraining variable. The gradient length of the first DCCA ordination axis gives a measure of turnover scaled in so-called standard deviation (SD) units, and on average ASVs rise and fall over 4SD. A turnover of 4SD units therefore represents a complete exchange of species (ter Braak and Prentice, 1988). Diatom and non-diatom assemblages were ordinated using non-metric multidimensional scaling and ordinations compared using Procrustes analysis to quantify the overall similarity in assemblage structure between the diatom and non-diatom data sets (Peres-Neto and Jackson 2001). Redundancy analysis (RDA) with the 6 selected water quality variables was used to explore community-wide patterns in assemblage structure. The significance of each variable was assessed using a Monte-Carlo permutation test (Borcard and others, 2011). Variance partitioning based on RDA was used to partition the total variance in the ASV assemblages explained by each water quality variable into shared and unique components (Peres-Neto and others, 2006). Diatom relative abundance data were square root transformed prior to RDA to yield ordinations based on Hellinger's distance, which is more appropriate for ecological community data (Legendre and Gallagher, 2001).

In addition to distance-based methods (RDA/DCCA), we also tested the diatom and non-diatom data sets for nutrient response using a model-based approach to fit separate generalised linear models (GLMs) to each ASV, with nutrient pressure as the explanatory variable, and using a negative binomial link to model the proportional abundance of each taxon (Wang and others, 2012). Significance of trends were assessed as for RDA. This model-based multivariate-GLM (mGLM) provides an alternative to RDA to test community-level trends and has the advantage that it also provides a significance test of the response for each individual ASV (Wang and others, 2012).

Predictive models were developed using weighted averaging (WA) and supervised machine learning (SML). Weighted averaging underlies the TDI and many other diatom-based indices and provides a benchmark for existing approaches. SML methods allow the development of predictive models using the information contained in a large training data set. Here, we develop SML models using random forests (RF) (Brieuc and others, 2018) and boosted regression trees (BRT) (Elith and others, 2008), as these are 2 of the most commonly used machine learning algorithms and can model non-linear relationships and high-dimensional data. Both methods are based on decision trees but take differing approaches to reducing errors in prediction or classification. Prediction error is a function of model bias and variance. A shallow tree with few splits will have high bias (poor accuracy) but low variance (small changes in the data won't change the outcome much). A large tree with many nodes will have low bias (fits the data well) but high variance (small changes will give a different outcome, that is, the model is 'overfitted'). A single tree is a weak learner and will usually perform poorly. RF and BRT are both ensemble methods that combine a large number (potentially thousands) of slightly different trees to reduce bias and variance. RF generates a large number of independent trees built using a random sample of the data, and the prediction is based on the consensus or mean of all trees. BRT builds the trees sequentially,

with each new tree trying to correct errors made by previous trees. BRTs and RFs have different advantages and disadvantages, so we evaluated both. BRTs can be more accurate but are prone to overfitting with noisy data and RFs are less prone to overfitting but can exhibit shrinkage of the predicted value towards the mean.

RF and BRT models were fitted to a subset of the diatom and non-diatom data sets containing only those ASVs with a maximum relative abundance greater than 1% and at least 10 occurrences. Hyper-parameters used to tune the tree models (number of potential variables for each split, number of trees for RF, depth of tree, learning rate for BRT) were optimised with a grid search using 2-fold cross-validation, in which the data were randomly split in 2 and the model trained on one half and tested on the other. 2-fold cross-validation was also used to compare the performance of WA and tree models as it is a useful guide to the likely error when the model is applied to new data. Each tree in a RF or BRT ensemble is trained on a slightly different random selection of variables. The relative importance of each explanatory variable (ASV) for predicting nutrient pressure can be estimated as a function of the number of times each variable is selected to split individual trees (weighted by the improvement in prediction error for BRTs).

All statistical analyses were performed using R software for statistical computing (R Core Team, 2020) with the following additional packages: vegan (PCA and RDA: Oksanen and others, 2020), mvabund (mGLM: Wang and others, 2012), ranger (RF: Wright and Zeigler 2017), gbm (BRT: Greenwell and others, 2020); rioja (WA: Juggins 2020), caret (tree optimisation, Kuhn 2021).

3.3.3 Results

Water quality data

Figure 7 summarises the distribution of selected water quality variables in the paired HTS-chemistry data set. As is common in landscapes shaped by human activity several variables were strongly correlated with each other (for example, reactive P with both NO_3-N and NH_4-N, alkalinity with conductivity). pH was relatively weakly associated with nutrients, but there are also relatively few records with pH values low enough to expect significant impacts on the biota.

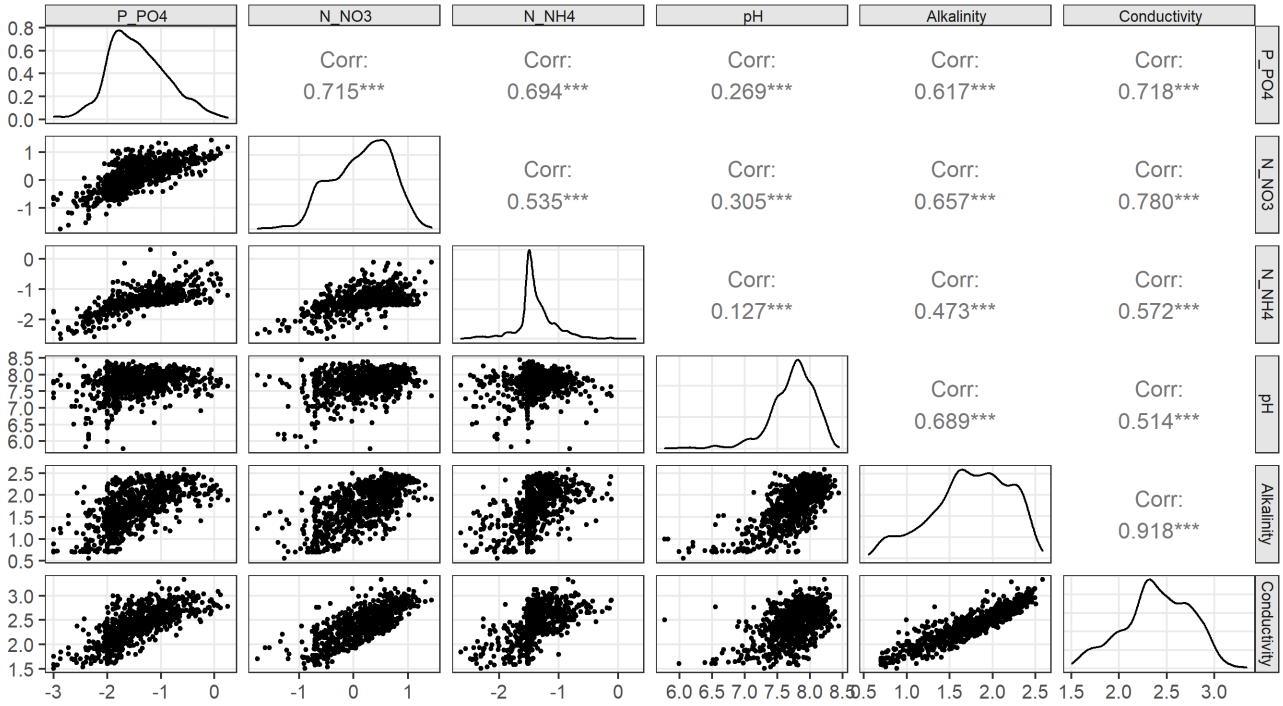


Figure 7: Distribution and relationship between selected water quality variables in the paired HTS-chemistry data set (N=1,220). Units are log_{10} μgL^{-1} (PO_4-P); log_{10} mgL^{-1} (NO_3-N, NH_4-N); log_{10} mgL^{-1} (alkalinity); log_{10} $\mu S\ cm^{-1}$ (conductivity)

HTS data characteristics

Figure 8 shows the distribution of total diatom and non-diatom reads and the proportion of the total reads assigned to non-diatom ASVs. The majority of samples are dominated by diatom ASVs, with only 36% of samples having more than 5% non-diatom reads. 98.5% of samples have at least 500 diatom reads but only 36% have more than 500 non-diatom reads.

The HTS data set contains 4,036 diatom and 3,187 non-diatom ASVs. The majority of these have low occurrence and/or low maximum relative abundance: only 1,152 diatom ASVs (28%) have a maximum abundance of greater than 1% in any single sample and only 398 (9.9%) are recorded in more than 50 samples. Non-diatom ASVs exhibit a different pattern, with 2,252 (71%) having a maximum abundance of greater than 1%, but only 66 (2.1%) recorded in more than 50 samples. Figure 9 (left) shows this apparent difference in the abundance/occupancy relationship in more detail. There is a strong positive relationship between abundance and occupancy for both taxonomic groups (diatoms, r=0.81; non-diatoms, r=0.78, both $p < 0.001$): more abundant taxa are generally more widespread. However, the pattern of taxon abundance is visibly different for the 2 groups: However, the pattern of taxon abundance is visibly different for the two groups: non-diatom ASVs tend to be more abundant for any given level of occupancy, and non-diatoms ASVs tend to be less widespread but more locally abundant. The patchy or disjunctive distribution of many non-

diatom ASVs is especially apparent in the left-hand side of the figure that shows that many non-diatom ASVs with only one or two occurrences can be dominant in some samples.

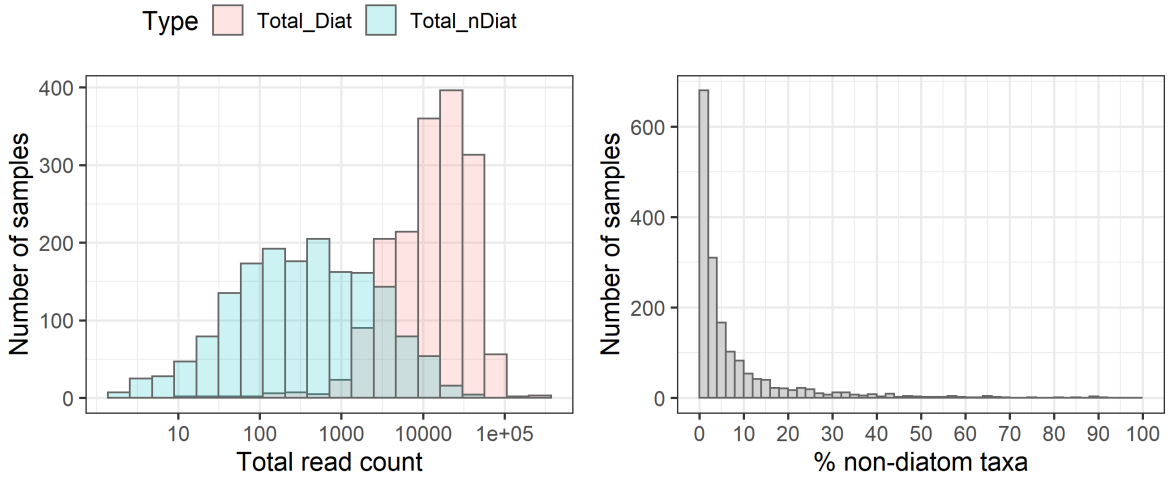


Figure 8: Distribution of total diatom and non-diatom read counts (left) and proportion of the total read count represented by non-diatom ASVs (right). Data represents paired HTS-chemistry data before screening for total read count (N=1,688)

An alternative way to examine the abundance/occurrence characteristics of organisms is to compare their N_2 number of occurrences. Whereas N_0, or the total number of samples in which an ASV is recorded, takes no account of abundance, N_2 is a measure of the effective number of occurrences, that is, the number of N_0 occurrences with equal abundance that would be needed to give the same N_2 value (Hill 1973). Figure 9 (right) shows the distribution of Hill's N_2 occurrences and is again markedly different between the 2 groups: 298 diatom ASVs (7.4%), but only 53 (1.7%) non-diatom ASVs have more than 20 N_2 occurrences, again indicating that, in general, non-diatom ASVs have a patchier distribution than diatoms and are generally recorded in fewer samples.

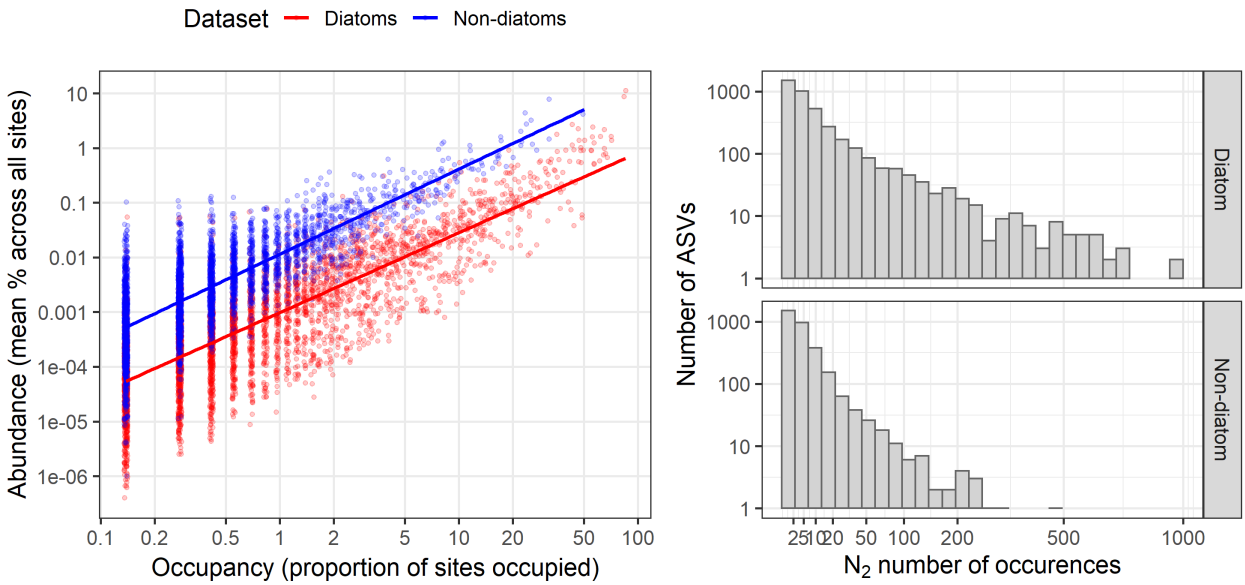


Figure 9: Relationship between the frequency of occurrence (occupancy) and mean relative abundance (left) and the distribution of Hill's N_2 number of occurrences for diatom and non-diatom ASVs (right). Note the overplotting of diatom ASVs on the left of the abundance/occupancy plot

Figure 10 summarises patterns in assemblage diversity between the 2 groups. Figure 10 (left) shows the distribution of N_2 effective number of ASVs per sample: for diatoms, 38% of samples have a N_2 diversity of 10 or more. For non-diatoms, the figure is 15%, meaning that non-diatom assemblages are, in general, less diverse than diatom assemblages. Figure 10 (right) shows the proportion of the assemblage that is contained in the top 100 most abundant diatom or non-diatom ASVs. For diatoms, at least 50% of the assemblage is comprised of the top 100 ASVs in 92% of samples. For non-diatoms, the figure is 71%, again emphasising the patchy distribution of ASVs in this data set, where there is a greater tendency for assemblages to be dominated by taxa that are rare or have low overall abundance.

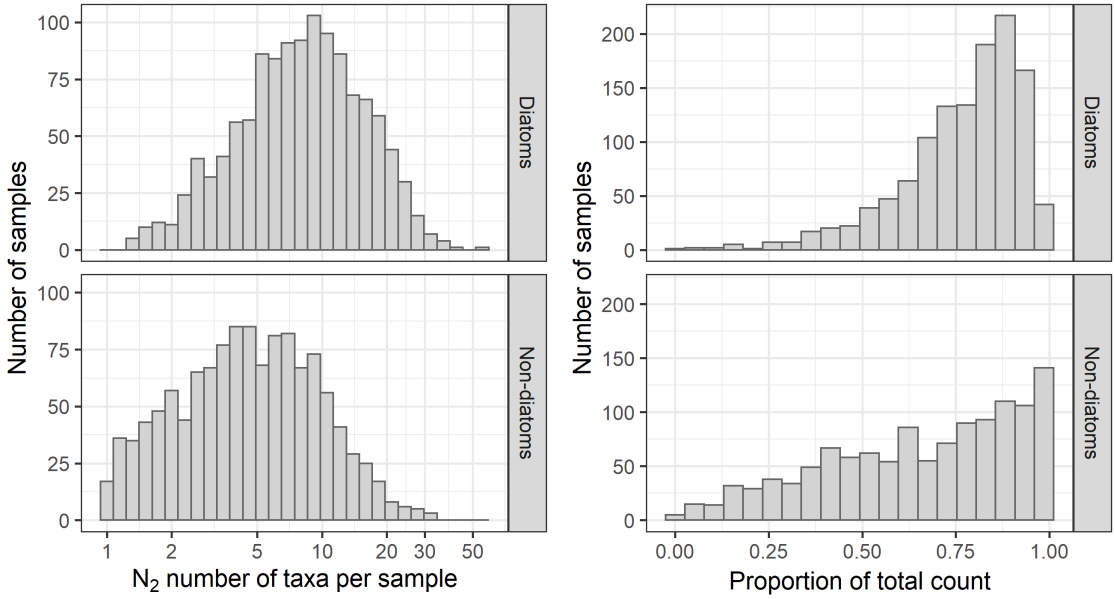


Figure 10: N_2 effective number of ASVs per sample (left) and proportion of the total assemblage accounted for by the 100 most abundant ASVs (right)

HTS community characteristics and species/environment relationships

The gradient length of the diatom ASV data set was 4.46 and non-diatom ASV data set 4.70, derived using detrended canonical correspondence analysis constrained to the nutrient pressure gradient. On average, each ASV rises and falls over 4 SDs. The gradient lengths of both data sets are similar and greater than 4, indicating a complete turnover of taxa – samples from sites with low nutrient pressure would be expected to have no species in common with those subjected to high nutrient pressure. The similarity in gradient length suggests that the 2 data sets have similar beta diversity, or patterns of turnover along the nutrient pressure gradient. That is, diatoms and non-diatoms respond in a similar way, and should exhibit similar sensitivity of species' response to nutrient pressure.

The species/environment patterns in the 2 data sets are further explored using ordination analysis. Procrustes correlations between unconstrained ordinations of the 2 data sets (using non-metric multidimensional scaling, nMDS) and constrained ordinations (using redundancy analysis, RDA) were performed. The first compares the main patterns of overall species distributions, making no assumptions about the relationship to pressure or other environmental gradients. The second quantifies the similarity of the explicitly modelled species environment relationship between the data sets. Results of the RDAs are shown in Figure 11. Procrustes correlation for the unconstrained ordination is 0.49, suggesting that approximately 25% of the variance in ASV distribution is common to both data sets. The corresponding correlation for the constrained ordination is 0.72, suggesting that about 50% of the species-sample-environment relationship is common to both data sets. Taken together, the results suggest that both data sets reflect a strong, common signal in the

gradients represented by the environmental data, but that there are also other patterns of variation not related to the measured environmental data which may differ between the diatom and non-diatom data sets.

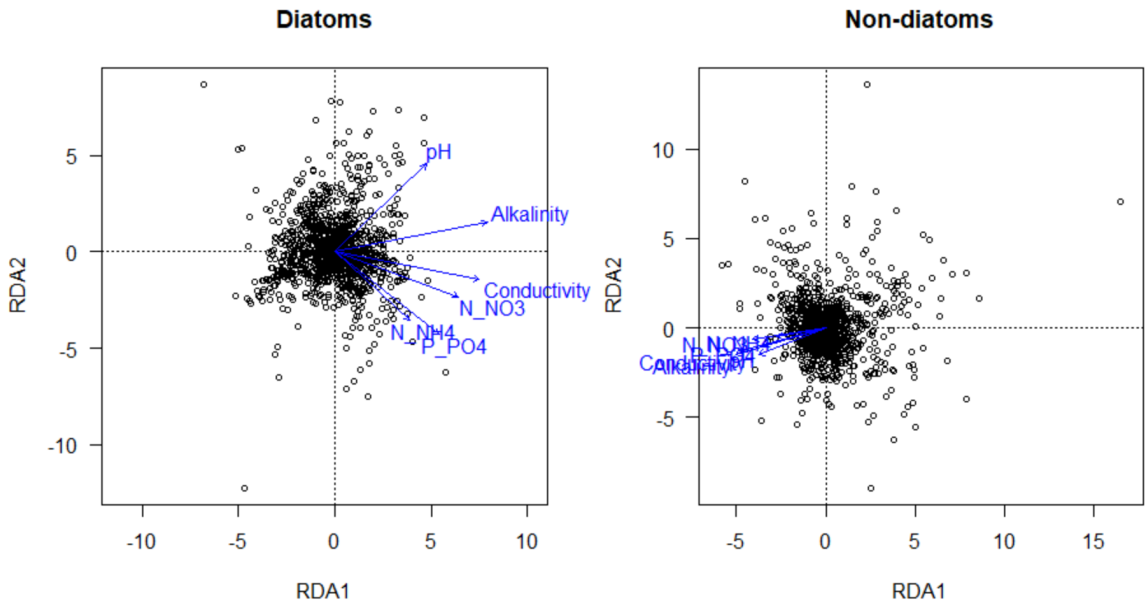


Figure 11: Redundancy analysis biplots of the HTS data sets showing sites (circles) and environmental variables (arrows). The marginal effects of all environmental variables are significant (p <0.01; 999 permutations)

Results of the constrained ordination suggest that the main species/environment response is related to the alkalinity/conductivity gradient. Nutrient-related variables (reactive P, NO_3-N, NH_4-N) also have a statistically significant effect, albeit confounded with the alkalinity gradient in both data sets, somewhat more so for non-diatoms.

Figure 12 shows the variance partitioning results for diatoms and non-diatoms. For both data sets the explained variance by each of the 4 selected environmental variables is low (approximately 4 to 6% for diatoms, 1 to 2% for non-diatoms). Such low values are usual for large, highly diverse data sets, and all the shared and unique components of variation are significant (p <0.01). The variance partitioning results mirror the conclusions from the RDA ordinations, that alkalinity and conductivity explain slightly more variation than the nutrient variables, and that a large part of the explained variance is shared or confounded between nutrient and alkalinity gradients. The pattern of shared and unique variance explained is remarkably similar between diatoms and non-diatoms, suggesting they are responding in the same way, at least to these gradients, except that there might be a slightly stronger NO_3-N response in the non-diatom ASVs.

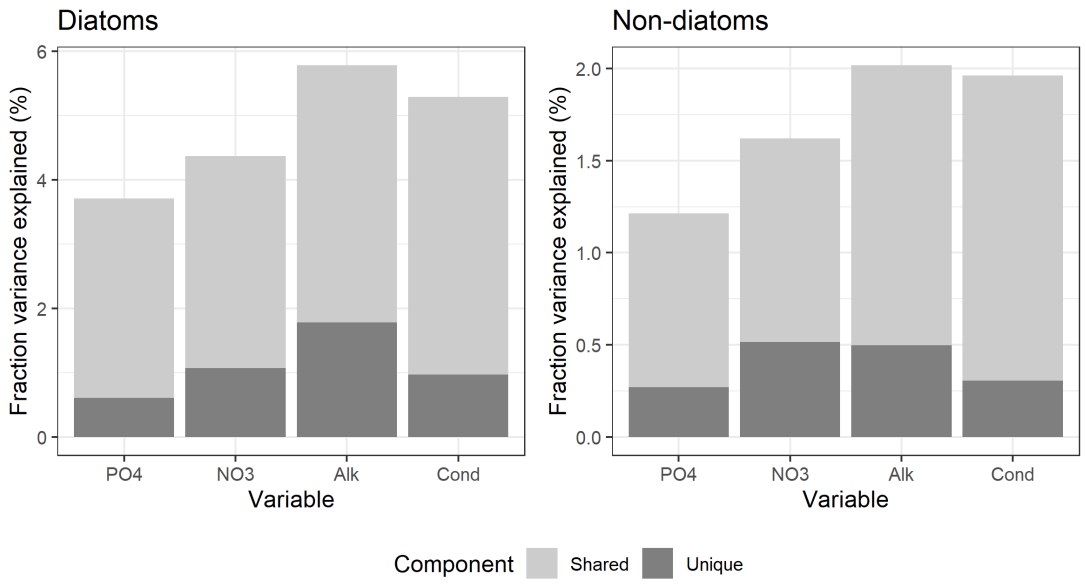


Figure 12: Variance partitioning results, showing the total variance explained by each environmental variable partitioned into unique and shared components. All components of variation are significant ($p < 0.01$)

Results of the ordination analyses indicate a statistically significant effect of nutrient-related variables on diatom and non-diatom assemblage composition at the assemblage level. Figure 13 shows the results of ASV-level analysis using generalised linear modelling (GLM) to model the response of each ASV to the pressure gradient. This analysis was performed on a subset of ASVs that have a maximum relative abundance of at least 1% and at least 10 occurrences, yielding reduced data sets of 765 and 449 ASVs for diatom and non-diatoms respectively. Overall, a total of 244 diatom ASVs (32%) and 108 non-diatom ASVs (24%) exhibited a significant response to the nutrient pressure gradient ($p < 0.01$). The majority of dominant or highly abundant diatom ASVs (with maximum relative abundance greater than 10% and present in more than 100 samples) have a significant response, although significant responses are also recorded for many less frequent and less abundant ASVs. Non-diatom ASVs tend to have fewer occurrences and significant responses are recorded predominantly for those ASVs with maximum relative abundance greater than 10%.

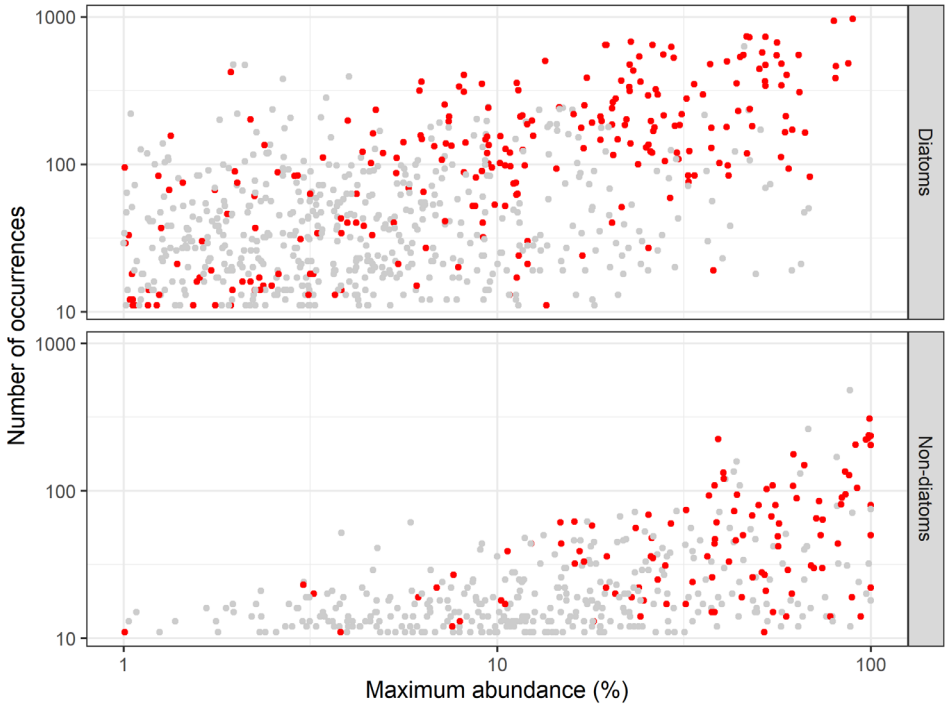


Figure 13: Abundance–occurrence plot for the subset of ASVs used in species response modelling. ASVs with a significant response to the nutrient pressure gradient are shown as red circles

Figure 14 plots the distribution of the 36 most abundant diatom and non-diatom ASVs that show a significant response along the nutrient pressure gradient, along with their fitted response models. While these models are an oversimplification of the distributions shown by many ASVs, they do capture the broad trends in species turnover along the pressure gradient and highlight differences between ASVs. Note too that several species' names (Achnanthidium minutissimum, Navicula gregaria, Amphora pediculus) recur, as ASVs will be picking up variation within 'species' as identified by light microscopy. A. minutissimum is a catch-all for a complex that is known to be more diverse than the current UK recording convention allows, but variation within N. gregaria does not map neatly onto known morphospecies. Noteworthy within the non-diatoms is the large number of 'cf.' (= conferret, literally, 'compare', used to indicate difficulties in assigning a binomial) as well as 3 'NAs', where it was impossible to make a link with any known species. Non-diatoms in this list are drawn from the Chlorophyta, Streptophyta, Xanthophyceae, Cyanobacteria, Euglenophyta and Cryptophyta, not all of which would necessarily be expected to amplify readily using our primers (see section 3.2.3). Interestingly, the Phaeophyceae, Eustigmatophyceae and several Ochrophyta that might be expected to amplify well were absent.

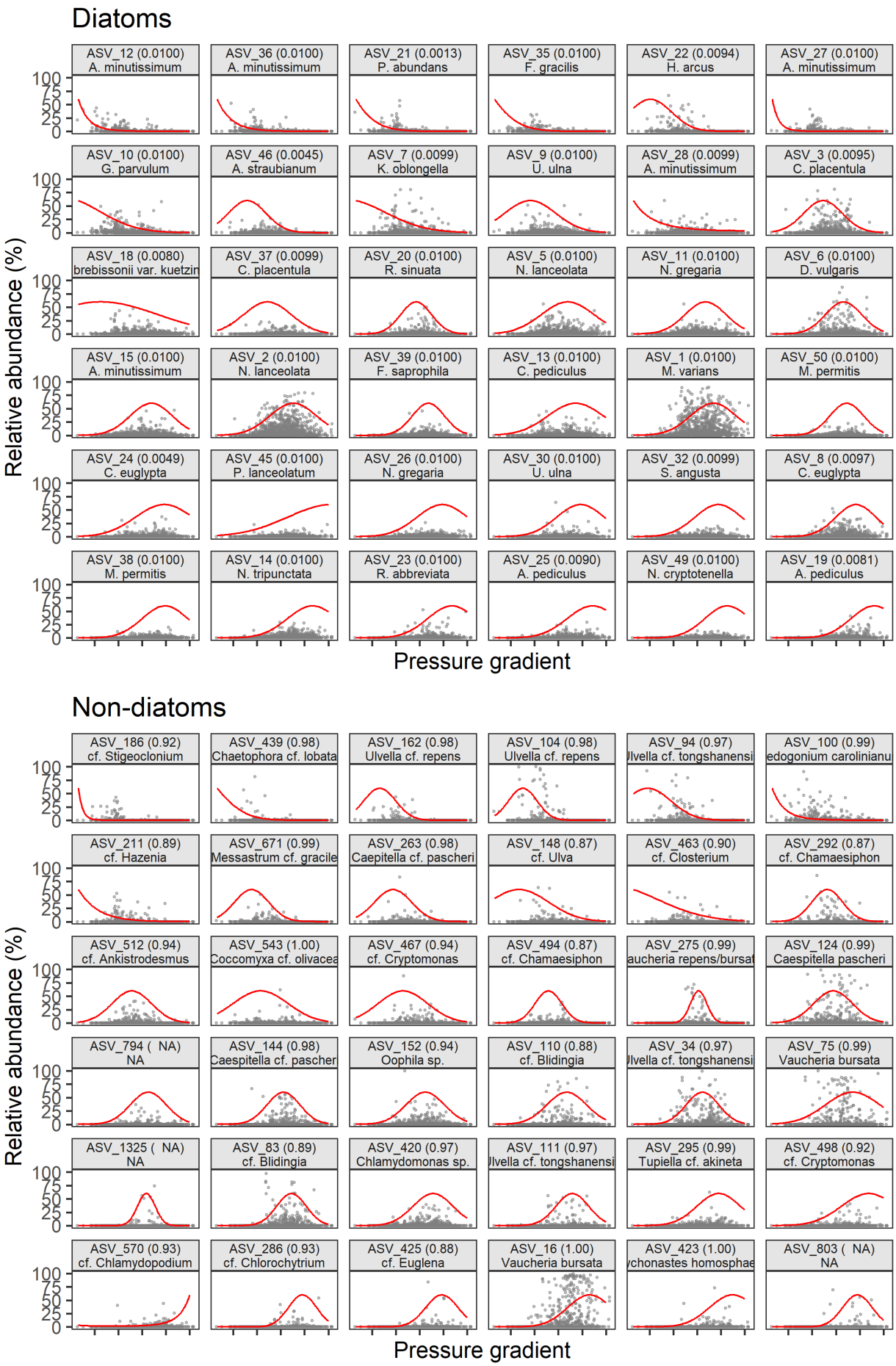

Diatoms
ASV_12 (0.0100) A. minutissimum
ASV_36 (0.0100) A. minutissimum
ASV_21 (0.0013) P. abundans
ASV_35 (0.0100) F. gracilis
ASV_22 (0.0094) H. arcus
ASV_27 (0.0100) A. minutissimum
ASV_10 (0.0100) G. parvulum
ASV_46 (0.0045) A. straubianum
ASV_7 (0.0099) K. oblongella
ASV_9 (0.0100) U. ulna
ASV_28 (0.0099) A. minutissimum
ASV_3 (0.0095) C. placentula
ASV_18 (0.0080) brebissonii var. kuetzin
ASV_37 (0.0099) C. placentula
ASV_20 (0.0100) R. sinuata
ASV_5 (0.0100) N. lanceolata
ASV_11 (0.0100) N. gregaria
ASV_6 (0.0100) D. vulgaris
ASV_15 (0.0100) A. minutissimum
ASV_2 (0.0100) N. lanceolata
ASV_39 (0.0100) F. saprophila
ASV_13 (0.0100) C. pediculus
ASV_1 (0.0100) M. varians
ASV_50 (0.0100) M. permitis
ASV_24 (0.0049) C. euglypta
ASV_45 (0.0100) P. lanceolatum
ASV_26 (0.0100) N. gregaria
ASV_30 (0.0100) U. ulna
ASV_32 (0.0099) S. angusta
ASV_8 (0.0097) C. euglypta
ASV_38 (0.0100) M. permitis
ASV_14 (0.0100) N. tripunctata
ASV_23 (0.0100) R. abbreviata
ASV_25 (0.0090) A. pediculus
ASV_49 (0.0100) N. cryptotenella
ASV_19 (0.0081) A. pediculus
Relative abundance (%)
100
75
50
25
0
Pressure gradient
Non-diatoms
ASV_186 (0.92) cf. Stigeoclonium
ASV_439 (0.98) Chaetophora cf. lobata
ASV_162 (0.98) Ulvella cf. repens
ASV_104 (0.98) Ulvella cf. repens
ASV_94 (0.97) Ulvella cf. tongshanensi
ASV_100 (0.99) edogonium carolinianu
ASV_211 (0.89) cf. Hazenia
ASV_671 (0.99) Messastrum cf. gracile
ASV_263 (0.98) Caepitella cf. pascheri
ASV_148 (0.87) cf. Ulva
ASV_463 (0.90) cf. Closterium
ASV_292 (0.87) cf. Chamaesiphon
ASV_512 (0.94) cf. Ankistrodesmus
ASV_543 (1.00) Coccomyxa cf. olivacea
ASV_467 (0.94) cf. Cryptomonas
ASV_494 (0.87) cf. Chamaesiphon
ASV_275 (0.99) aucheria repens/bursat
ASV_124 (0.99) Caespitella pascheri
ASV_794 (NA) NA
ASV_144 (0.98) Caespitella cf. pascher
ASV_152 (0.94) Oophila sp.
ASV_110 (0.88) cf. Blidingia
ASV_34 (0.97) Ulvella cf. tongshanensi
ASV_75 (0.99) Vaucheria bursata
ASV_1325 (NA) NA
ASV_83 (0.89) cf. Blidingia
ASV_420 (0.97) Chlamydomonas sp.
ASV_111 (0.97) Ulvella cf. tongshanensi
ASV_295 (0.99) Tupiella cf. akineta
ASV_498 (0.92) cf. Cryptomonas
ASV_570 (0.93) cf. Chlamydopodium
ASV_286 (0.93) cf. Chlorochytrium
ASV_425 (0.88) cf. Euglena
ASV_16 (1.00) Vaucheria bursata
ASV_423 (1.00) chonastes homosphae
ASV_803 (NA) NA
Relative abundance (%)
100
75
50
25
0
Pressure gradient

Figure 14: Distributions of the 36 most abundant diatom (top) and non-diatom (bottom) AVSs that show a significant response to the nutrient pressure gradient ($p < 0.01$). Pressure gradient scaled from low (left) to high (right). Fitted species response curves shown as red lines. Numbers in brackets are the confidence of species assignment.

Quantifying predictive pressure-response relationships

Results of the ordination analyses indicate a statistically significant effect of nutrient-related variables on diatom and non-diatom assemblage composition. We now reverse the direction of the modelling and ask how well can nutrient pressure be predicted using biology? The degree to which taxonomy-free HTS data can be used as an indicator of nutrient pressure was explored by fitting a series of predictive models, starting with a simple weighted-averaging (WA) model as this is the basis of the TDI and many other biological water quality metrics.

Results of the WA models are shown in Figure 15. Diatoms, non-diatoms and the combined data set ('both') exhibit a similar and strong correlation between measures of nutrient pressure and model fits (r=0.83 to 0.85). This compares well with the correlation of 0.80 between TDI5LM scores and nutrient pressure, calculated from phase 1 data. Corresponding values under 2-fold cross-validation, which gives a more realistic idea of model performance when applied to samples, are only slightly lower for diatoms and 'both' (r=0.80) but more so for non-diatoms (r=0.76). The poorer performance of the non-diatom model under cross-validation is the result of the greater heterogeneity and patchy/disjunctive distribution of many ASVs in this data set. The scatter plots and density plots also both show that predicted values shrink towards the mean of the pressure gradient, especially at high nutrient pressure.

Figure 16 shows similar results for random forest and boosted regression tree models. For diatoms and the combined diatom and non-diatom data set, both methods have the same correlation with nutrient pressure and have similar patterns of predictions. Although there is some bias towards the mean (that is, high values are under-predicted, low values are over-predicted, this is less so than for WA models (compare Figure 15 and 16 density plots). Performance of the non-diatom tree models is also similar for the 2 methods and uniformly worse than the diatom models (r=0.70 and 0.73).

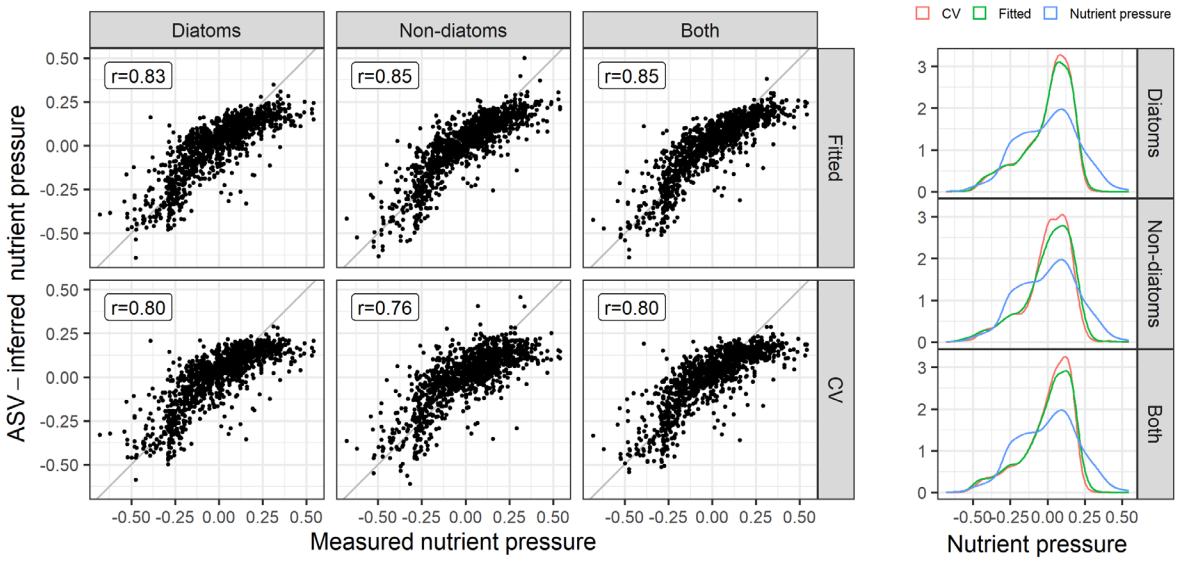


Figure 15: Results of weighted-average (WA) predictive models, showing relationship between observed and ASV-inferred nutrient pressure for diatoms and non-diatoms, and both (combined) (left). Upper plots show relationship for model fits, lower plots show results after 2-fold leave-out cross validation. Right-hand figures show the distribution of original and predicted nutrient pressure

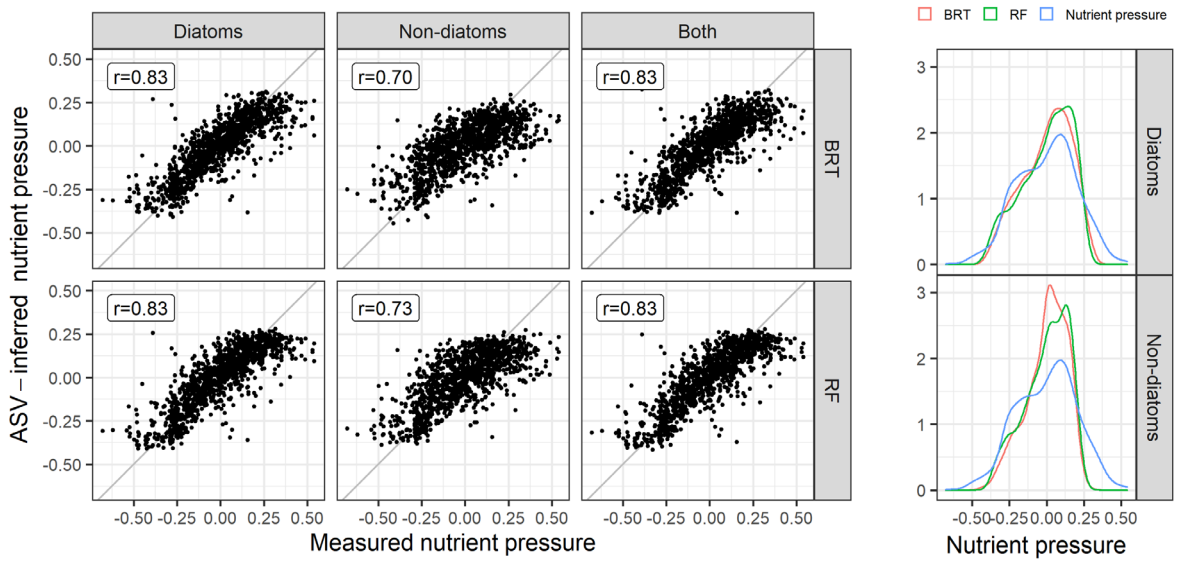


Figure 16: Results of random forest (RF) and boosted regression tree (BRT) predictive models, showing relationship between observed and ASV-inferred nutrient pressure for diatoms and non-diatoms (left). All plots show cross-validation predictions based on 2-fold leave-out. Right-hand figures show the distribution of original and predicted nutrient pressure

Variable importance statistics, derived from the random forest model trained using the combined diatom and non-diatom data, indicate that the model is primarily based on diatom ASVs, especially those assigned to the orders Achnanthales, Bacillariales, Naviculales, Fragilariales and Tabellariales (Figure 17). However, ASVs belonging to several Chlorophyta orders and several Vaucheriales also contribute. Variable importance derived from the BRT models is very similar to that from RF and is not shown.

Figure 17: Boxplot summarising importance of ASVs for predicting nutrient pressure, grouped by taxonomic order, for random forest model using combined diatom and non-diatom data

Figures 18 and 19 summarise the response of the 80 most important ASVs in the diatom (Figure 18) and non-diatom (Figure 19) RF predictive models to nutrient pressure. ASVs have been assigned names using diatom and non-diatom barcode libraries as described in

section 3.2. There are marked differences in distribution for most ASVs listed and whether these ASVs can be assigned to a morphological taxon with some confidence. There are also marked intra-species differences: for example, differences in distribution for ASVs assigned to Tabellaria flocculosa and Achnanthidium minutissimum.

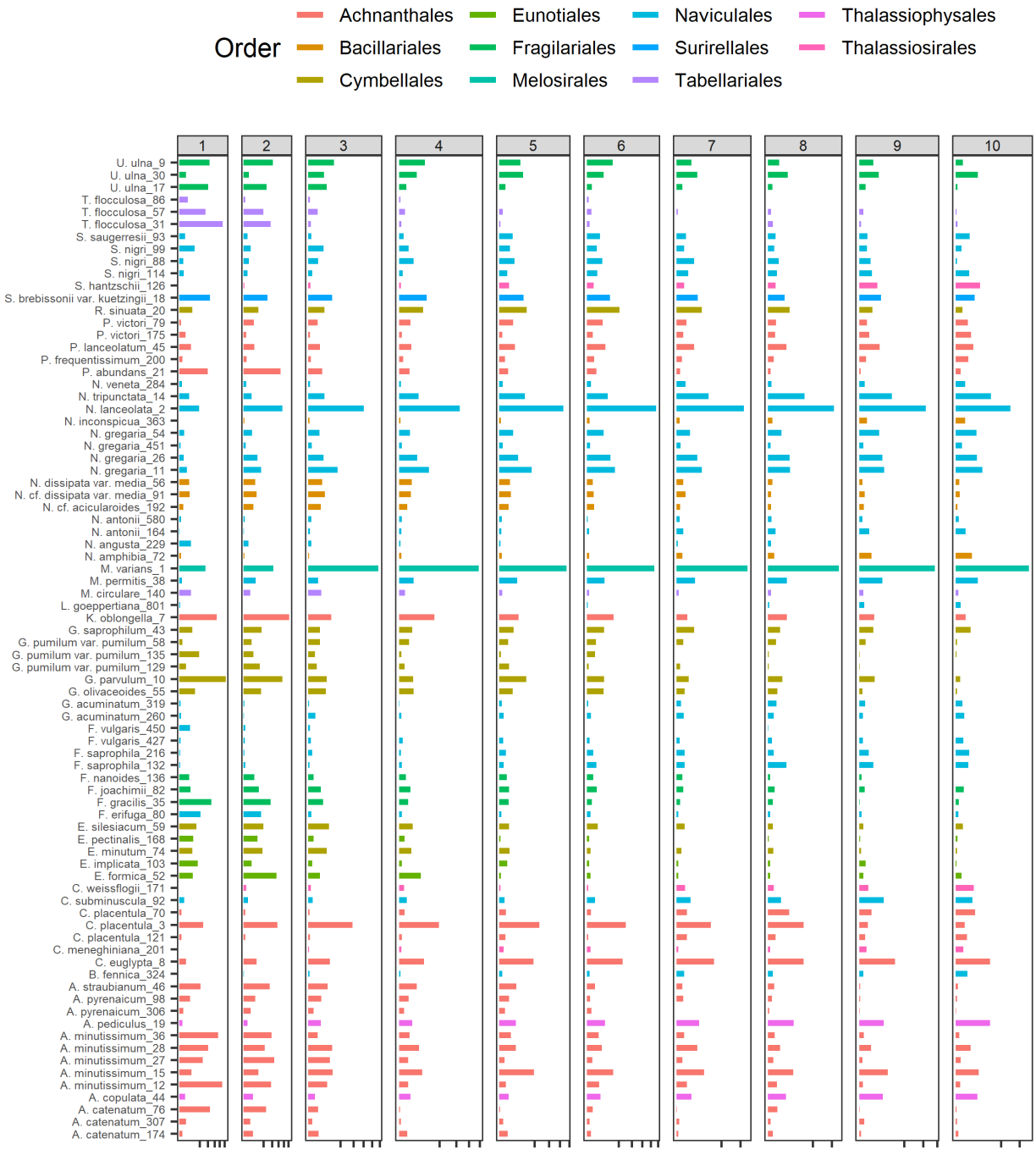


Figure 18: Distribution of the 80 most important diatom taxa for predicting nutrient pressure using tree-based models. Nutrient pressure gradient is divided into 10 equal sample-size bands (1=low). Bars show the mean relative abundance of ASVs in that nutrient pressure band. X-axis is square root scaled

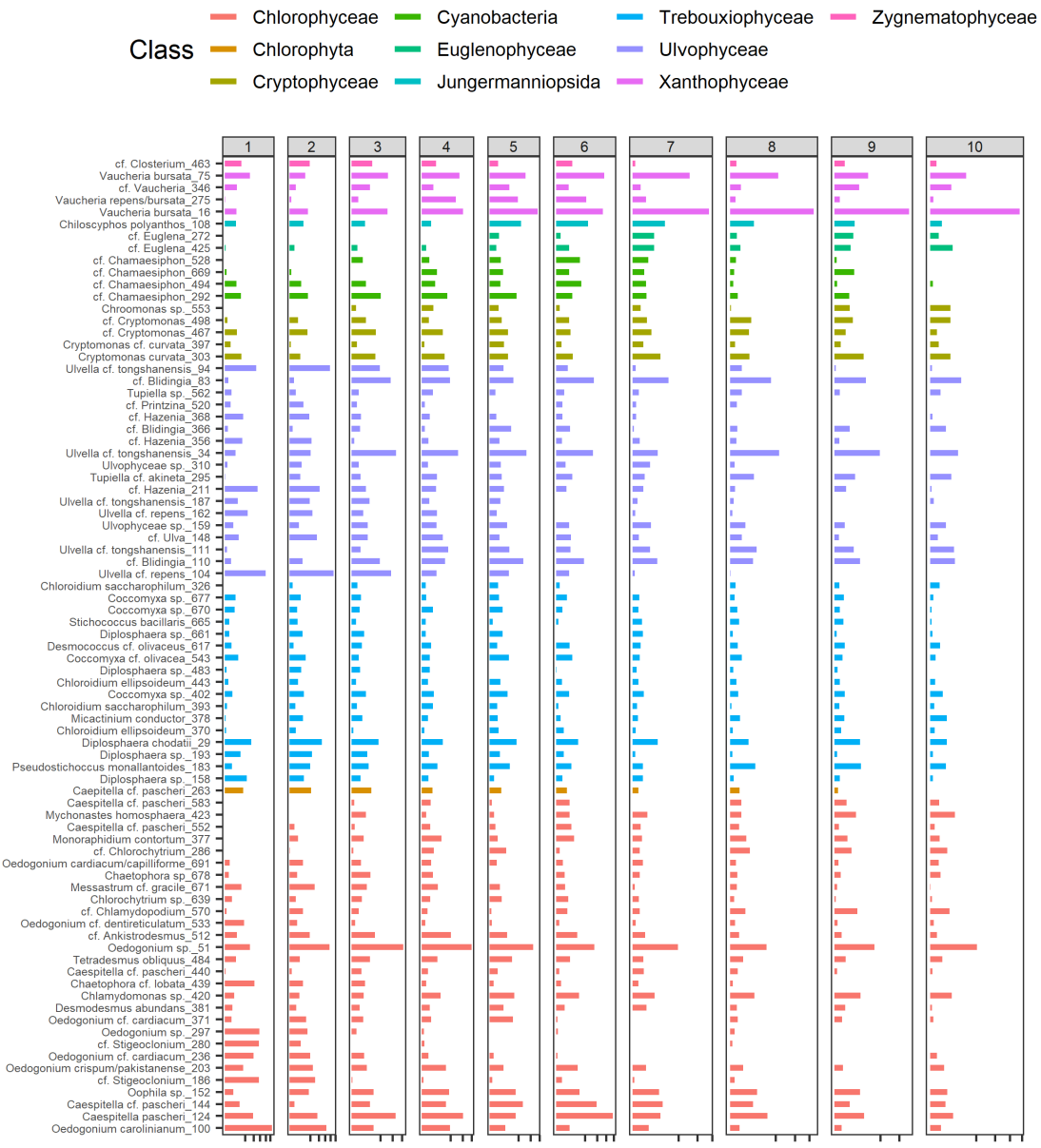


Figure 19: Distribution of the 80 most important non-diatom taxa for predicting nutrient pressure using tree-based models. Nutrient pressure gradient is divided into 10 equal sample-size bands (1=low). Bars show the mean relative abundance of ASVs in that nutrient pressure band. X-axis is square root scaled

3.3.3 Discussion

Limitations in the water quality data greatly reduced the number of HTS samples that could be used in this analysis, but by combining the phase 2 (2017 to 2019) HTS samples with existing data from 2014 to 2016 analysed in phase 1 we were able to produce a large, paired HTS and water quality data set. These data were generated using a consistent approach and can be used to explore the relationships between taxonomy-free diatom and non-diatom ASVs and nutrient pressure.

Examination of the abundancy, occupancy and diversity plots reveal differences in the frequency/abundance relationships between diatom and non-diatoms ASVs. While non-diatoms often have a patchier distribution within reaches (M. Kelly, unpublished data), these differences are likely to also be a result of primer bias as the primers were designed for diatoms. We have no way of assessing how the primers react to other groups with and without the presence of a large diatom biomass. At certain times of the year when diatom biomass may be lower, there may be greater amplification of non-diatoms. Furthermore, some of the other algal groups are multicellular and are classed as macroalgae rather than microalgae. This will have a distinct impact on their relative abundance in the sequence libraries as a single individual of a multicellular macroalga could dominate a sample, but it does not mean that it is more abundant at one site than another or more abundant than a biofilm of diatoms. As a result, the non-diatom ASV data must be treated with caution, both in terms of the total non-diatom reads, and the community-wide patterns of non-diatom composition. It should also be noted that both the sampling protocol and DNA extraction has been optimised for diatoms, and other methods may well be more appropriate for non-diatom taxa.

Despite problems accurately identifying and quantifying non-diatom ASVs, the species/environment patterns in the diatom and non-diatom data sets are remarkably similar, and both data sets exhibit similar patterns and strength of response to the nutrient and other water quality variables. This is perhaps in part because both groups are from the same sample but also because, as algae, both groups are likely to respond in the same way to environmental variables required for a photosynthetic life strategy.

The potential for developing new metrics based on diatom and/or non-diatom taxonomy-free HTS data using weighted-averaging and decision tree models was evaluated. The target variable was a composite nutrient pressure metric derived from combining P and N variables, as this relationship was stronger than for P or N alone. The aim was not to predict nutrient pressure - this can be measured directly with water chemistry - but the models were used to quantify the likely strength and relative performance of any future metrics developed using this approach.

We measured the strength of this relationship using the correlation between observed and predicted nutrient pressure under 2-fold cross-validation. When evaluating these models, it should be noted that the unexplained error emanates both from the **model error** (how well mean annual chemistry encapsulates an ecologically meaningful driver, and how the numerical method accurately models the potentially complex and non-linear response to this variable) and the **measure error** in the target variable (how accurate the estimate of mean annual chemistry is). The measurement error of the composite nutrient variable can be estimated from the within-year variance in standardised N and P measurements at each site and is 14.8 for PO_4 and 12.4 for NO_3. Therefore, the maximum variance the models could explain is around 1 to 13.6, or 86.4%, equivalent to a correlation between measured and predicted nutrient pressure of 0.93.

Comparing phase 2 results with those from phase 1 indicates that both diatom and non-diatom ASV-based models outperform the equivalent taxonomy-based HTS models (compare Figures 15 and 16 with Figure 4). For example, the DADA2 TDI5 model has a correlation of 0.66 with the pressure gradient, whereas the cross-validated DADA2 ASV diatom models have correlations of 0.80 to 0.83. In addition, results suggest that diatom ASV models may perform as well as, or better than, the equivalent light-microscopy model. Part of this improvement may be the finer taxonomic resolution offered by HTS. For example, a total of 47 ASVs are assigned to the taxon Achnanthidium minutissimum. Figure 20 shows the distribution of the most frequently occurring 12 (with N >40). Some ASVs have similar distribution (for example, 15 and 237), but there is also considerable variation in the apparent response of these ASVs to nutrient pressure. This potentially useful information would be lost using a taxonomy-based approach. Achnanthidium minutissimum is a complex of species that have been 'lumped' for convenience in the TDI but where there is abundant evidence of many species, albeit difficult to separate reliably using LM.

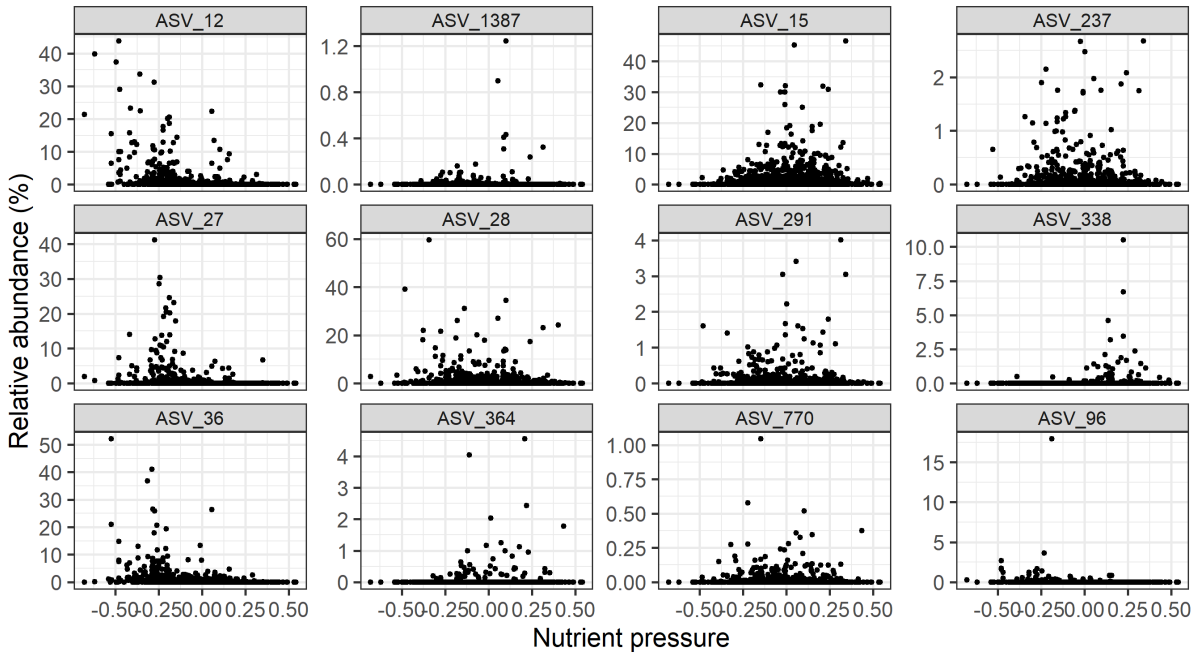


Figure 20: Distribution against nutrient pressure of the most frequently occurring ASVs assigned to Achnanthidium minutissimum

Given the problems with the enumeration of the non-diatom phytobenthos, it is difficult to compare the relative predictive power of diatom and non-diatom components and to quantify the improvement in predictive ability of any new metrics by including non-diatoms. That said, the performance of the non-diatom tree-based models, the importance of some non-diatom ASVs in these models, and the statistically significant relationships of many non-diatom AVS with nutrient pressure suggest that there is the potential for improvement.

A major problem in deriving a nutrient pressure metric using observation data is the correlation between alkalinity and nutrients in many data sets, making it difficult to separate a nutrient from an alkalinity response (see Figures 7 and 12; Baattrup-Pedersen and others,

2022). Species response modelling using multivariate GLMs may help here. The models shown in Figure 14 have a single explanatory variable (nutrient pressure) and are based on a simple symmetric unimodal response curve. More complex distribution models could be fitted to multiple environmental variables using generalised additive models (GAMs) to test the differential response to nutrient and alkalinity gradients at the individual ASV level. Many ASVs have extremely low abundance or occurrence (or both) and do not make a meaningful contribution to prediction. Extracting useful information from these rare ASVs is an additional challenge that could be addressed by using a 2-step GAM to model presence and abundance separately (Barry and Welsh 2002). From a practical point of view, however, alkalinity is used to predict the reference value in EQR calculations and Kelly and others (2020) extended this to explicitly account for interactions between nutrients and alkalinity.

Although taxonomy-free approaches have been suggested by others (Apothéloz-Perret-Gentil and others, 2017; Feio and others, 2020), we stress the importance of retaining the link between HTS output and Linnaean nomenclature, the latter acting as a bridge to the wide knowledge base on freshwater functioning. The approach of Apothéloz-Perret-Gentil and others (2017) aimed at little more than replicating the correlations with water quality achieved using LM indices: we argue that the 'added value' for ecological assessment depends on being able to understand how phytobenthos assemblages interact with physical, chemical and biological components of ecosystems. This is exemplified by Willby (2011) and Poikane and others' (2018) use of 'guiding images' to encapsulate 'good ecological status' for lake macrophytes, but the concepts extend to phytobenthos too (Kelly, 2012; Kelly and others, 2019), with the 3-dimensional arrangement within biofilms influencing (and influenced by) interactions with other trophic levels. Matching OTUs to Linnaean nomenclature before deriving metrics was a sensible approach when negotiating the transition from LM to HTS, helping us to understand how the 2 types of data compared (section 2.2.4). We now recommend basing metrics on ASVs and fitting outputs to Linnaean nomenclature as a final step to produce taxa lists to help interpretation. This will enable metrics to make full use of information, including cryptic species (Kahlert and others , 2019; Pérez-Burillo and others, 2021) and species not yet included in the barcode database. Although this does not result in a significant increase in the strength of the relationship between the diatom assemblage and pressures, it is a conceptual shift that will create more opportunities for developing the phytobenthos metabarcoding approach in the future.

Further development of the phytobenthos metabarcoding approach does not have to be confined to diatoms (see section 1.2), although in this study we saw no increase in sensitivity to nutrient pressure when non-diatoms were included in models. Section 3.1 outlines some of the challenges involved in detecting non-diatoms with a barcode optimised for diatoms. Extending the reach of the approach also brings some new challenges, such as distorted signals when the biofilm is sampled from macrophytes rather than cobbles, and contamination from terrestrial and exotic vegetation (for example, bananas). It is clear some groups of non-diatoms should amplify well with the present barcode (Cryptophyta, Eustigmatophyceae, Phaeophyceae), but others that would be expected to be abundant in rivers are under-represented. Reasons are discussed in section 3.1, and Figures 11 and 12

also highlight these issues. The present barcode library is not suitable for evaluating the condition of the whole phytobenthos community. It is suitable for diatoms and the groups listed above, likely to be bycatch, dependent upon competition from diatoms at the PCR stage, rather than as a representative ecological signal.

Detecting the whole phytobenthos community would mean either the UK rbcL primers would need to be redesigned, an optimised rbcL primer set added for other major algal lineages found in rivers, or a different marker (for example, 18S) used, which would still miss Cyanobacteria – a major component of riverine biofilms. There would also be a need to develop well-curated reference libraries similar to diat.barcode (Rimet and others, 2019) for these groups (we caution against indiscriminate use of GenBank as a taxonomic resource for this purpose). We have, however, identified ASVs from non-diatoms that do have ecological signals along pressure gradients of interest (Figure 20), and which could be used in future metric development. Caution is needed when using these, as there is no independent validation of the distribution of the organisms that they represent. Nonetheless, a subset, where there is a priori reason to assume reliable amplification, could be included in future metrics. Once again, however, we emphasise that the overall increase in sensitivity to principal pressure gradients from including these will be relatively small, largely, we suspect, because they contribute relatively little unique information to a gradient already well captured by the diatoms.

Finally, we would like to highlight that traditional microscopic analysis of diatoms for the purpose of monitoring is built on several decades (perhaps even a century) of research and analysis of specific species. We know a great deal about their ecology and responses to pressure gradients. The same is not true for molecular analysis of these taxa. As already mentioned, there is potential to reveal within-morphospecies diversity. While there is good agreement in some respects between LM and molecular data for monitoring diatoms, there is still a learning curve regarding responses of ASVs to environmental pressures. We have also highlighted many potential caveats within the molecular analysis of diatoms. With this in mind, an exclusive focus on diatoms may not be the best way forward for monitoring rivers. Several studies have highlighted the utility of other microbial groups, such as bacteria (Stoeck and others, 2018; Aylagas and others, 2021; Pearman and others, 2022), protists (Ai and others, 2021; Kulaš and others, 2021) or multi-taxa assemblages using multiple phylogenetic markers (Keeley and others, 2018; Clark and others, 2020) in classifying the status of aquatic habitats, based on the assemblages of these taxa in response to environmental pressures (see Sagova-Mareckoba 2021 for a review). As we have already highlighted, the rbcL primers would need to be redesigned to target other photosynthetic organisms to ensure consistency of taxa recovery.

Currently, there is a legislative requirement to evaluate phytobenthos (WFD Regulations 2017 and WFD Annex V), and diatoms have proved valuable in meeting this need. While the regulatory landscape is determined by the WFD, there is sense in continuing with an approach that minimises disruption (even if this cannot be eliminated entirely). However, as policy evolves, there may be advantages in looking at other microbial groups entirely or

using a combined multi-taxa data set to build a new index equivalent to the TDI but based on molecular monitoring of numerous microbial groups.

3.3.4 Recommendations

- The link between HTS outputs and Linnaean taxonomy should be retained for the information this provides when interpreting outputs. However, future developments should use ASVs to calculate metrics, with links to reference libraries made as a final step in order to generate taxa lists for interpretation.
- Studies using mock communities of known biomass are needed to better understand the relationship between sequence biomass of individual taxa and sequence reads in a metabarcoding library; this work would also underpin better QA/QC within workflows. The mock communities should include a range of benthic species, including non-diatoms, where possible.
- There is limited benefit in including non-diatom reads in models using the current barcode. There are also some problems, such as distorted signals when the biofilm is sampled from macrophytes rather than cobbles, and contamination from terrestrial and exotic vegetation. Therefore, we do not recommend using ASVs without some prior filtering.
- It may be possible to include a few other groups known to amplify well, but this will, at best, only give a partial overview of the phytobenthic community. If a coarse 'total phytobenthos' assessment is required, a more general approach (general eukaryote primers for sequencing) should be used. If a finer resolution is required, a number of specific taxa should be targeted (using a number of different group-specific PCR primer sets).
- The present study indicates that there is relatively little scope for improvement in the current approach (deriving EQRs from measures of community turnover); therefore, the possibility of exploring alternative metrics (for example, incorporating diversity) or bypassing EQR calculation and predicting status class directly should be explored.
- Any further exploration of the potential of non-diatoms will benefit from access to a well-curated reference database, similar to diat.barcode. Such a database does not yet exist; we caution against the indiscriminate use of GenBank as a taxonomic resource.
- This study has identified considerable diversity in Eustigmatophyceae (previously poorly known) and a wider distribution than previously thought for the freshwater Phaeophyceae. Although beyond the formal remit of this project, these results offer a strong case study for the benefits of metabarcoding for expanding knowledge of aquatic biodiversity in the UK. It is therefore recommended that metabarcoding is more widely incorporated into biodiversity monitoring.

4.Moving forward

4.1 Phase 1 – diatom metabarcoding methods

4.1.1 Options

1. Continue with current pipeline (QIIME 1) with methodology as it is for the analysis of metabarcoding data.
   a. Advantages:
      i. Will not require updates to software or hardware
      ii. Has been used on previous data sets, although these could be rerun quickly with newer pipelines
      iii. Appears to be meeting the needs of stakeholders and end users
   b. Disadvantages:
      i. Generates many spurious OTUs
      ii. Low accuracy with potential for false positives/negatives
      iii. Limited utility beyond aims of original study
      iv. QIIME 1 no longer supported
      v. Potentially can't update linux operating system or servers as QIIME has several dependencies where the programs need to be specific versions(i.e these cannot be updated)
      vi. Potentially difficult to publish data using UCLUST estimates of OTUs

2. Switch to a UPARSE OTU based pipeline
   a. Advantages:
      i. Showed similar correlation to light microscopy as original pipeline
      ii. Better accuracy than the original pipeline
      iii. Easy to install and run
   b. Disadvantages:
      i. Does not denoise data, will still contain spurious OTUs
      ii. Larger datasets will require 64bit version, cost over £1000 per machine
      iii. Does not have some of the advantages of ASV generation - no within species/population variation
      iv. Requires update of code and methodologies
      v. Cost involved in setting up new protocols

3. Switch to DADA2 pipeline for all future diatom metabarcoding work.
   a. Advantages:
      i. Showed similar correlation to light microscopy as original pipeline
      ii. Runs on Windows or linux in R. Can be wrapped with DARLEQ scripts

iii. Stable release continually supported
iv. Extra taxonomic information can be gained from non-diatoms and total diatoms

b. Disadvantages:
i. Requires update of code and methodologies
ii. Produces larger data sets – that is, many ASVs where just species names needed, would require additional scripts
iii. Cost involved in setting up new protocols

4.1.2 Recommendation

Option 3, upgrading the current pipeline to DADA2 is recommended. This will allow detailed analysis of all rbcL metabarcoding data sets and comparison/validation with European monitoring (and potentially globally, depending on species assemblages), where metabarcoding is used. It will improve accuracy and reduce the potential for false negatives/positives. The potential extra utility gained from this method would be significant.

4.2 Phase 2 – inclusion of non-diatom taxa

4.2.1 Options

1. Go forward with current rbcL primers.
   a. Advantages:
      i. Good representation of diatoms
      ii. Improved reference model already available
      iii. Good (and expanding) reference database available at diat.barcode
   b. Disadvantages:
      i. Limited representation of other algae
      ii. Some important filamentous algal genera missed entirely
      iii. Limited potential for improving predictions of ecological status along principal pressure gradient
   c. Other:
      i. Some potential for moving beyond weighted average-type models and developing new statistical approaches for predicting status
      ii. Scope for developing additional metrics using existing data and/or easy-to-collect field information (visual assessment of filamentous algae) to increase ecological insight

2. Add extra rbcL primers to improve representation of other groups.
   a. Advantage:

i. Better representation of other algal groups

b. Disadvantages:
   i. Well-curated reference libraries will need to be developed
   ii. May not overcome reach-level ‘patchiness’ of many algae
   iii. rbcL is not a standard taxonomic marker for some algal groups
   iv. Expensive

3. Shift to alternative marker or taxonomic group(s) (for example, 16S,18S).
   a. Advantages:
      i. Better representation of other algal groups
      ii. Expand coverage to include other eukaryotic organisms
   b. Disadvantages:
      i. Well-curated reference libraries will need to be developed
      ii. May not overcome reach-level ‘patchiness’ of many algae
      iii. Quantification less well understood than for algae using the rbc*L* marker
      iv. There is an advantage in using a chloroplast marker, because it restricts the data set to photosynthetic organisms (in some communities, these could be swamped by heterotrophic protists and fungi in rDNA HTS outputs) and excludes problems caused by intraindividual rDNA variation (for example, Behnke and others, 2004), which could exaggerate and bias perceived organismal diversity
      v. Where tested, 18S was a less effective barcode for algae than rbcL (for example, Mann and others, 2010)
      vi. Expensive

4.2.2 Evaluation

Phytobenthos metrics, to date, have focused on community turnover along stressor gradients, and the present study suggests limited improvement from simply adding more (non-diatom) taxa to these metrics. Option 1, staying with the current primers, is the least risky option, as the other options assume that there is a significant untapped signal in the other algae. However, the general direction of travel (away from an exclusive focus on diatoms) is more likely to meet longer term goals of monitoring aquatic habitats with molecular techniques, allowing. assessment of ecological functioning. Using these data to develop new approaches may be as fruitful as adding a wider range of taxa. Based on results for phase 2, it can be assumed that all options will result in considerable change in the assessment of ecological status, so significant further work would be required to embed any novel approach within the future statutory framework.

4.2 Future work

The potential studies listed below, particularly further mock community validation, would help to progress development of the methodologies, with the aim of improving end-user confidence and uptake of the methods. Although it would increase the cost, developing methodologies for analysis of other microbial communities alongside the diatoms is likely to provide a great deal of information about nutrient and other pressures, habitat quality and ecological functioning of rivers.

1. Experiments could be conducted with mock communities of different taxa, both diatoms and non-diatoms, to better assess how accurately the assay performs in recovering relative abundance information from mixed communities. This would provide validation of the DNA extraction methodology, by adding known numbers of cells to extraction tubes and looking at recovery rates in relation to other taxa.
2. Reference library: assigning identities to ASVs to generate reliable taxa lists; use phylogenetic placement as a tool to ensure correct identifications at the lowest taxonomic level possible; include non-diatoms for groups that are known to amplify with the present primers. At this stage, it is more important to ensure that the higher level taxonomy is correct within the reference library, rather than that coverage of freshwater diatom species is complete. This will boost the efficiency of bioinformatics algorithms and limit the number of misclassifications of ASVs.
3. Explore new metrics and new approaches to predicting ecological status using supervised machine learning algorithms. Detailed development and testing of metabarcoding with different marker genes 16S, 18S, internal transcribed spacer (ITS) region to look at a range of different microbial communities and how they respond to environmental stressors. A multi-taxa assessment of the river benthos is likely to give better assessment of the impact of water quality/stressors than a single taxon alone. Other microbial groups may well perform better than the diatoms. Assessment of bacterial communities may also identify factors such as point source pollution from agriculture or sewage treatment (such as detection of faecal coliforms).

References

ADL, S., BASS, D., LANE, C.E., LUKEŠ, J. AND OTHERS., 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. Journal of Eukaryotic Microbiology 66: 4-119.

AI, M.A., XUE, Y., XIAO, P., CHEN, H., ZHANG, C., DUAN, M. AND YANG, J., 2021. DNA metabarcoding reveals the significant influence of anthropogenic effects on microeukaryotic communities in urban waterbodies. Environmental Pollution, 285, p.117336.

APOTHÉLOZ-PERRET-GENTIL, L., BOUCHEZ, A., CORDIER, T., CORDONIER, A., GUÉGUEN, J., RIMET, F., VASSELON, V. AND 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.

APOTHÉLOZ-PERRET-GENTIL, L., CORDONIER, A., STRAUB, F., ISELI, J., ESLING, P. AND PAWLOWSKI, J., 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Molecular Ecology Resources. https://doi.org/10.1111/1755-0998.12668 [accessed 19 October 2022].

AYLAGAS, E., ATALAH, J., SÁNCHEZ-JEREZ, P., PEARMAN, J.K., CASADO, N., ASENSI, J., TOLEDO-GUEDES, K. AND CARVALHO, S., 2021. A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. Molecular Ecology Resources, 21(6), pp.1889-1903.

BAATTRUP-PEDERSEN, A., JOHNSEN, T.J., LARSEN, S.E. AND RIIS, T., 2022. Alkalinity and diatom assemblages in lowland streams: How to separate alkalinity from inorganic phosphorus in ecological assessments? Sci Total Environ, 153829.

BAILET, B., APOTHÉLOZ-PERRET-GENTIL, L., BARIČEVIĆ, A., CHONOVA, T., FRANC, A., FRIGERIO, J.M., KELLY, M.G., MORA, D., PFANNKUCHEN, M., PROFT, S., 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. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.140948 [accessed 19 October 2022].

BAKSAY, S., PORNON, A., BURRUS, M., MARIETTE, J., ANDALO, C. AND ESCARAVAGE, N., 2020. Experimental quantification of pollen with DNA metabarcoding using ITS1 and trnL. Scientific Reports, 10(1), pp.1-9.

BARRY, S.C. AND WELSH, A.H., 2002. Generalized additive modelling and zero inflated count data. Ecol. Modelling 157, 179-188.

BEHNKE, A., FRIEDL, T., CHEPURNOV, V.A. AND MANN, D.G., 2004. Reproductive compatibility and rDNA sequence analyses in the Sellaphora pupula species complex (Bacillariophyta). – Journal of Phycology 40: 193-208.

BENG, K.C., TOMLINSON, K.W., SHEN, X.H., SURGET-GROBA, Y., HUGHES, A.C., CORLETT, R.T. AND SLIK, J.W., 2016. The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. Scientific Reports, 6(1), pp.1-13.

BISTA, I., CARVALHO, G.R., WALSH, K., SEYMOUR, M., HAJIBABAEI, M., LALLIAS, D., CHRISTMAS, M. AND CREER, S., 2017. Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. Nature Communications, 8(1), pp.1-11.

BOMBIN, S., WYSOR, B. AND LOPEZ-BAUTISTA, J.M., 2021. Assessment of littoral algal diversity from the northern Gulf of Mexico using environmental DNA metabarcoding. Journal of Phycology, 57(1), pp.269-278.

BORCARD, D., GILLET, F., LEGENDRE, P., 2011. Numerical Ecology with R. Springer, Dordrecht.

BRAUKMANN, T.W., IVANOVA, N.V., PROSSER, S.W., ELBRECHT, V., STEINKE, D., RATNASINGHAM, S., DE WAARD, J.R., SONES, J.E., ZAKHAROV, E.V. AND HEBERT, P.D., 2019. Metabarcoding a diverse arthropod mock community. Molecular ecology resources, 19(3), pp.711-727.

BRIEUC, M.S.O., WATERS, C.D., DRINAN, D.P., NAISH, K.A., 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Mol Ecol Resour 18, 755-766.

BUKIN, Y.S., GALACHYANTS, Y.P., MOROZOV, I.V., BUKIN, S.V., ZAKHARENKO, A.S. AND ZEMSKAYA, T.I., 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Scientific Data, 6(1), pp.1-14.

CALLAHAN, B.J., MCMURDIE, P.J., ROSEN, M.J., HAN, A.W., JOHNSON, A.J.A. AND HOLMES, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583.

CAPORASO, J.G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F.D., COSTELLO, E.K., FIERER, N., PEÑA, A.G., GOODRICH, J.K., GORDON, J.I. AND HUTTLEY, G.A., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), pp.335-336.

CHARLES, D.F., KELLY, M.G., STEVENSON, R.J., POIKANE, S., THEROUX, S., ZGRUNDO, A. AND CANTONATI, M., 2021. Benthic algae assessments in the EU and

the US: Striving for consistency in the face of great ecological diversity. Ecological Indicators. https://doi.org/10.1016/j.ecolind.2020.107082 [accessed 19 October 2022].

CLARK, D.E., PILDITCH, C.A., PEARMAN, J.K., ELLIS, J.I. AND ZAIKO, A., 2020. Environmental DNA metabarcoding reveals estuarine benthic community response to nutrient enrichment–Evidence from an in-situ experiment. Environmental Pollution, 267, p.115472.

DEL CORTONA, A., LELIAERT, F., BOGAERT, K.A., TURMEL, M., BOEDEKER, C., JANOUŠKOVEC, J., LOPEZ-BAUTISTA, J.M., VERBRUGGEN, H., VANDEPOELE, K. AND DE CLERCK, O., 2017. The plastid genome in Cladophorales green algae is encoded by hairpin chromosomes. Current Biology 27: 3771-3782.

DELWICHE, C.F. AND PALMER, J.D., 1996. Rampant horizontal transfer and duplication of Rubisco genes in Eubacteria and plastids. Mol. Biol. Evol. 13: 873–882.

DOPHEIDE, A., TOOMAN, L.K., GROSSER, S., AGABITI, B., RHODE, B., XIE, D., STEVENS, M.I., NELSON, N., BUCKLEY, T.R., DRUMMOND, A.J. AND NEWCOMB, R.D., 2019. Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. Ecological Applications, 29(4), p.e01877.

DULEBA, M., FÖLDI, A., MICSINAI, A., VÁRBÍRÓ, G., MOHR, A., SIPOS, R., … ÁCS, É., 2021. Applicability of diatom metabarcoding in the ecological status assessment of Hungarian lotic and soda pan habitats. Ecological Indicators, 130, 108105.

EDGAR, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), pp.2460-2461.

EDGAR, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods, 10(10), 996–998.

EDGAR, R.C., 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, 081257.

EDGAR, R.C. AND FLYVBJERG, H., 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics, 31(21), pp.3476-3482.

EDGAR, R.C., HAAS, B.J., CLEMENTE, J.C., QUINCE, C. AND KNIGHT, R., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27(16), pp.2194-2200.

EGAN, C.P., RUMMEL, A., KOKKORIS, V., KLIRONOMOS, J., LEKBERG, Y. AND HART, M., 2018. Using mock communities of arbuscular mycorrhizal fungi to evaluate fidelity associated with Illumina sequencing. Fungal Ecology, 33, pp.52-64.

ELITH, J., LEATHWICK, J.R., HASTIE, T., 2008. A working guide to boosted regression trees. The Journal of Animal Ecology 77, 802-813.

ENVIRONMENT AGENCY, 2018. A DNA based metabarcoding approach to assess diatom communities in rivers. https://www.gov.uk/government/publications/a-dna-based-metabarcoding-approach-to-assess-diatom-communities-in-rivers [accessed 19 October 2022].

ENVIRONMENT AGENCY, 2020. Assessing river nutrients using diatom DNA: further development of an operational method. https://www.gov.uk/government/publications/assessing-river-nutrients-using-diatom-dna-further-development-of-an-operational-method [accessed 19 October 2022].

FEIO, M.J., SERRA, S.R., MORTÁGUA, A., BOUCHEZ, A., RIMET, F., VASSELON, V. AND ALMEIDA, S.F., 2020. A taxonomy-free approach based on machine learning to assess the quality of rivers with diatoms. Science of the Total Environment, 722, p.137900.

FLYNN, J.M., BROWN, E.A., CHAIN, F.J.J., MACISAAC, H.J. AND CRISTESCU, M.E., 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecology and Evolution, 5(11), 2252–2266.

GREENWELL, B., BOEHMKE, B., CUNNINGHAM, J.E., GBM DEVELOPERS., 2020. gbm: Generalized Boosted Regression Models. R

GUIRY, M.D. & GUIRY, G.M. 2022. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. www.algaebase.org [accessed 19 October 2022]

HAJIBABAEI, M., SHOKRALLA, S., ZHOU, X., SINGER, G.A. AND BAIRD, D.J., 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS one, 6(4), p.e17497.

HILL, M.O., 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology 54, 427-432.

HINZ, S., COSTON-GUARINI, J., MARNANE, M. AND GUARINI, J.M., 2022. Evaluating eDNA for Use within Marine Environmental Impact Assessments. Journal of Marine Science and Engineering, 10(3), p.375.

HLEAP, J.S., LITTLEFAIR, J.E., STEINKE, D., HEBERT, P.D. AND CRISTESCU, M.E., 2021. Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources, 21(7), pp.2190-2203.

JOHN, D.M., WHITTON, B.A. AND BROOK, A.J., 2011. The Freshwater Algal Flora of the British Isles. 2nd edition. Cambridge University Press, London.

JOSHI N.A., FASS J.N., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

JUGGINS, S., 2020. rioja: Analysis of Quaternary Science Data. R package version 0.9-31. http://cran.r-project.org/package=rioja, 0.9-31.

KAHLERT, M., KELLY, M.G., MANN, D.G., RIMET, F., SATO, S., BOUCHEZ, A. AND KECK, F., 2019. Connecting the morphological and molecular species concepts to facilitate species identification within the genus Fragilaria (Bacillariophyta). Journal of Phycology, 55(4), pp.948-970.

KANG, W., ANSLAN, S., BÖRNER, N., SCHWARZ, A., SCHMIDT, R., KÜNZEL, S., RIOUAL, P., ECHEVERRÍA-GALINDO, P., VENCES, M., WANG, J. AND SCHWALB, 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. Ecological Indicators, 121, p.107070.

KARNKOWSKA, A., BENNETT, M.S. AND TRIEMER, R.E., 2018. Dynamic evolution of inverted repeats in Euglenophyta plastid genomes. Scientific Reports 8:16071.

KECK, F., BLACKMAN, R.C., BOSSART, R., BRANTSCHEN, J., COUTON, M., HÜRLEMANN, S., KIRSCHNER, D., LOCHER, N., ZHANG, H. AND ALTERMATT, F., 2022. Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. Molecular Ecology, 31(6), pp.1820-1835.

KEELEY, N., WOOD, S.A. AND POCHON, X., 2018. Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. Ecological Indicators, 85, pp.1044-1057.

KELLY, M., 2012. The Semiotics of Slime: Visual Representation of Phytobenthos as an aid to Understanding Ecological Status. Freshwater Reviews. https://doi.org/10.1608/frj-5.2.511 [accessed 19 October 2022].

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. Ecological Indicators. https://doi.org/10.1016/j.ecolind.2020.106725 [accessed 19 October 2022].

KELLY, M.G., KING, L. AND YALLOP, M.L., 2019. As trees walking: the pros and cons of partial sight in the analysis of stream biofilms. Plant Ecology and Evolution. https://doi.org/10.5091/plecevo.2019.1620 [accessed 19 October 2022].

KELLY, M.G., PHILLIPS, G. JUGGINS, S. AND WILLBY, N.J., 2020. Re-evaluating expectations for river phytobenthos assessment and understanding the relationship with macrophytes. Ecological Indicators 107: 106582.

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. Freshwater Biology 53, 403-422.

KELLY, M.G. and WHITTON, B.A., 1995. The trophic diatom index: a new index for monitoring eutrophication in rivers. Journal of Applied Phycology, 7(4), pp.433-444.

KOLLER, B., GINGRICH, J.C., STIEGLER, G.L., FARLEY, M.A., DELIUS, H. AND HALLICK, R.B., 1984. Nine introns with conserved boundary sequences in the Euglena gracilis chloroplast ribulose-1,5-bisphosphate carboxylase gene. Cell 36: 545–553.

KUHN, M., 2021. caret: Classification and Regression Training. R package version 6.0-90. (https://CRAN.R-project.org/package=caret). [accessed 19 October 2022].

KULAŠ, A., GULIN, V., KEPČIJA, R.M., ŽUTINIĆ, P., PERIĆ, M.S., ORLIĆ, S., KAJAN, K., STOECK, T., LENTENDU, G., ČANJEVAC, I. AND MARTINIĆ, I., 2021. Ciliates (Alveolata, Ciliophora) as bioindicators of environmental pressure: A karstic river case. Ecological Indicators, 124, p.107430.

KULAŠ, A., UDOVIČ, M.G., TAPOLCZAI, K., ŽUTINIĆ, P., ORLIĆ, S. AND LEVKOV, Z., 2022. Diatom eDNA metabarcoding and morphological methods for bioassessment of karstic river. Science of The Total Environment, p.154536.

KUNTKE, F., DE JONGE, N., HESSELSØE, M. AND NIELSEN, J.L., 2020. Stream water quality assessment by metabarcoding of invertebrates. Ecological Indicators, 111, p.105982.

LEGENDRE, P., GALLAGHER, E., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271-280.

MAJANEVA, M., HYYTIÄINEN, K., VARVIO, S.L., NAGAI, S., AND BLOMSTER, J., 2015. Bioinformatic Amplicon Read Processing Strategies Strongly Affect Eukaryotic Diversity and the Taxonomic Composition of Communities. PLOS ONE, 10(6), e0130035.

MANN, D.G., SATO, S., TROBAJO, R., VANORMELINGEN, P., SOUFFREAU, C., 2010. DNA barcoding for species identification and discovery in diatoms. Cryptogamie, Algologie 31: 557-577.

MARTIN, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp.10-12.

MCKNIGHT, D.T., HUERLIMANN, R., BOWER, D.S., SCHWARZKOPF, L., ALFORD, R.A., ZENGER, K.R., JARMAN, S., 2019. Methods for normalizing microbiome data: An ecological perspective. Methods in Ecology and Evolution 10, 389-400.

NEARING, J.T., DOUGLAS, G.M., COMEAU, A.M. AND LANGILLE, M.G.I., 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ, 6, e5364.

OKSANEN, J., BLANCHET, F.G., FRIENDLY, M., KINDT, R., LEGENDRE, P., McGLINN, D., MINCHIN, P.R., O'HARA, R.B., SIMPSON, G.L., SOLYMOS, P. and STEVENS, M.H.H., 2020. Vegan: Community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. R package version 2.5 (2019). R Package Version. Available online: https://CRAN. R-project. org/package= vegan (accessed on 13 December 2021).

PEARMAN, J.K., WOOD, S.A., VANDERGOES, M.J., ATALAH, J., WATERS, S., ADAMSON, J., THOMSON-LAING, G., THOMPSON, L., HOWARTH, J.D., HAMILTON, D.P. AND POCHON, X., 2022. A bacterial index to estimate lake trophic level: National scale validation. Science of The Total Environment, 812, p.152385.

PERES-NETO, P.R. AND JACKSON, D.A., 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia, 129(2), 169–178.

PERES-NETO, P., LEGENDRE, P., DRAY, S., BORCARD, D., 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology 87(10), 2614-2625.

PÉREZ-BURILLO, J., TROBAJO, R., VASSELON, V., RIMET, F., BOUCHEZ, A. AND MANN, D.G., 2020. Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.138445 [accessed 11 October 2022].

PÉREZ-BURILLO, J., TROBAJO, R., LEIRA, M., KECK, F., RIMET, F., SIGRÓ, J. AND MANN, D.G., 2021. DNA metabarcoding reveals differences in distribution patterns and ecological preferences among genetic variants within some key freshwater diatom species. Science of The Total Environment, 798, 149029.

PÉREZ-BURILLO, J., VALOTI, G., WITKOWSKI, A., PRADO, P., MANN, D.G. AND TROBAJO, R., 2022. Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters. Marine Pollution Bulletin, 174, p.113183.

POIKANE, S., PORTIELJE, R., DENYS, L., ELFERTS, D., KELLY, M., KOLADA, A., … VAN DEN BERG, M.S., 2018. Macrophyte assessment in European lakes: Diverse approaches but convergent views of 'good' ecological status. Ecological Indicators, 94. https://doi.org/10.1016/j.ecolind.2018.06.056 [accessed 19 October 2022].

PORNON, A., ESCARAVAGE, N., BURRUS, M., HOLOTA, H., KHIMOUN, A., MARIETTE, J., PELLIZZARI, C., IRIBAR, A., ETIENNE, R., TABERLET, P. AND VIDAL,

M., 2017 Using metabarcoding to reveal and quantify plant-pollinator interactions. Scientific Reports, 6(1), pp.1-12.

PRODAN, A., TREMAROLI, V., BROLIN, H., ZWINDERMAN, A. H., NIEUWDORP, M. AND LEVIN, E., 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLOS ONE, 15(1), e0227434.

RIMET, F., ABARCA, N., BOUCHEZ, A., KUSBER, W.-H., JAHN, R., KAHLERT, M., KECK, F., KELLY, M.G., MANN, D.G., PIUZ, A., TROBAJO, R., TAPOLCZAI, K., VASSELON, V. AND ZIMMERMANN, J., 2018. 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.

RIMET, F., GUSEV, E., KAHLERT, M., KELLY, M.G., KULIKOVSKIY, M., MALTSEV, Y., … BOUCHEZ, A., 2019. Diat.barcode, an open-access curated barcode library for diatoms. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-51500-6.

RIVERA, S.F., VASSELON, V., BOUCHEZ, A. AND RIMET, F., 2021. Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software. Ecological Indicators, 109, 105775. doi: 10.1016/j.ecolind.2019.105775.

R CORE TEAM 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project. org/.

RSTUDIO TEAM (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

SAGOVA-MARECKOVA, M., BOENIGK, J., BOUCHEZ, A., CERMAKOVA, K., CHONOVA, T., CORDIER, T., EISENDLE, U., ELERSEK, T., FAZI, S., FLEITUCH, T. AND FRÜHE, L., 2021. Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. Water Research, 191, p.116767.

SCHAUMBURG, J., SCHRANZ, C., FOERSTER, J., GUTOWSKI, A., HOFMANN, G., MEILINGER, P., … SCHMEDTJE, U., 2004. Ecological classification of macrophytes and phytobenthos for rivers in Germany according to the Water Framework Directive. Limnologica. https://doi.org/10.1016/S0075-9511(04)80002-1.

SCHLOSS, P.D., 2020. Reintroducing mothur: 10 Years Later. Applied and Environmental Microbiology, 86(2).

SCHNEIDER, S.C. AND LINDSTRØM, E.A., 2011. The periphyton index of trophic status PIT: A new eutrophication metric based on non-diatomaceous benthic algae in Nordic rivers. Hydrobiologia. https://doi.org/10.1007/s10750-011-0614-7 [accessed 19 October 2022].

SCHROEDER, A., STANKOVIĆ, D., PALLAVICINI, A., GIONECHETTI, F., PANSERA, M. AND CAMATTI, E., 2020. DNA metabarcoding and morphological analysis-Assessment of zooplankton biodiversity in transitional waters. Marine Environmental Research, 160, p.104946.

SCHÜTZ, W., KELLY, M.G., KING, L. AND CANTONATI, M., 2021. Did zebra mussel fill the type habitat of a worldwide-rare freshwater brown macroalga? Aquatic Conservation: Marine and Freshwater Ecosystems. https://doi.org/10.1002/aqc.3731 [accessed 19 October 2022].

SEPA, 2018. Evaluation of benthic diatom classification in UK rivers using LM and NGS methods. Report No. E18-56.

SMITH, K.F., KOHLI, G.S., MURRAY, S.A. AND RHODES, L.L., 2017. Assessment of the metabarcoding approach for community analysis of benthic-epiphytic dinoflagellates using mock communities. New Zealand Journal of Marine and Freshwater Research, 51(4), pp.555-576.

STOECK, T., FRÜHE, L., FORSTER, D., CORDIER, T., MARTINS, C.I. AND PAWLOWSKI, J., 2018. Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Marine Pollution Bulletin, 127, pp.139-149.

TABITA, F.R., SATAGOPAN, S., HANSON, T.E., KREEL, N.E., SCOTT, S.S., 2008. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships, Journal of Experimental Botany 59: 1515–1524.

TAPOLCZAI, K., SELMECZY, G.B., SZABÓ, B., B-BÉRES, V., KECK, F., BOUCHEZ, A., … 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.

TER BRAAK, C.J.F., PRENTICE, I.C., 1988. A theory of gradient analysis. Advances in Ecological Research 18, 271-313.

Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., … Domaizon, I., 2018. Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. Methods in Ecology and Evolution. https://doi.org/10.1111/2041-210X.12960 [accessed 19 October 2022].

WANG, Q., GARRITY, G.M., TIEDJE, J.M. AND COLE, J.R., 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy ▿ Applied and Environmental Microbiology, 73(16), 5261–5267.

WANG, Y., NAUMANN, U., WRIGHT, S.T. AND WARTON, D.I., 2012. mvabund– an R package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution, 3(3), 471–474.

WILLBY, N.J., 2011. From metrics to Monet: The need for an ecologically meaningful guiding image. Aquatic Conservation: Marine and Freshwater Ecosystems. https://doi.org/10.1002/aqc.1233

WRIGHT, M.N., AND ZIEGLER, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 77.

YANG, C., WANG, X., MILLER, J.A., DE BLÉCOURT, M., JI, Y., YANG, C., HARRISON, R.D. AND DOUGLAS, W.Y., 2014. Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. Ecological Indicators, 46, pp.379-389.

ZHANG, J., KOBERT, K., FLOURI, T. AND STAMATAKIS, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5), pp.614-620.

ZHANG, G.K., CHAIN, F.J., ABBOTT, C.L. AND CRISTESCU, M.E., 2018. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. Evolutionary applications, 11(10), pp.1901-1914.

Appendix I

The following scripts were run to analyse the data using the USEARCH pipelines, the scripts run in the Linux operating system such as ubuntu.

#assumes cutadapt and USEARCHv11 are both installed.

#merge sequence pairs and relabel sequences with sample name

usearch11 -fastq_mergepairs *R1*.fastq -fastqout merged.fastq -relabel @

#trim the primers

cutadapt -a ATGCGTTGGAGAGARCGTTTC...GATCACCTTCTAATTTACCWACAACTG -o trimmed.fastq merged.fastq --discard-untrimmed

#filter and trim

usearch11 -fastq_filter trimmed.fastq -fastq_maxee 0.5 -fastq_minlen 200 -fastaout reads1.fa

#relabel sequences

sed 's/\./\;/g' reads1.fa > reads2.fa

sed 's/>/>barcodelabel=/g' reads2.fa > reads.fa

#dereplicate and sort

usearch11 -fastx_uniques reads.fa -fastaout derep.fa -sizeout

usearch11 -sortbysize derep.fa -fastaout sorted.fa

UPARSE pipeline

usearch11 -cluster_otus sorted.fa -otus otus1.fa

python $HOME/drive5_py/fasta_number.py otus2.fa OTU_ > otus.fa

usearch8 -uchime_ref otus1.fa -db USEDIAT.BARCODE database fasta here -strand plus -nonchimeras otus2.fa

#Relabel OTU names

python $HOME/drive5_py/fasta_number.py otus2.fa OTU_ > otus.fa

#usearch to map the otus back to the original seqs

usearch11 -usearch_global reads.fa -db otus.fa -strand plus -id 0.97 -uc map.uc

#construct otutable (a species abundance matrix)

python $HOME/drive5_py/uc2otutab.py map.uc > otu_table.txt

UNOISE3

takes “sorted.fa” file from above and makes amplicon sequence variants (100% OTUs)

usearch11 -unoise3 sorted.fa -zotus zotus.fa

makes OTU table can use reads.fa or “trimmed.fastq”

usearch11 -otutab reads.fa -otus zotus.fa -otutabout otutab_raw.tx

#assigning taxonomy using RDP https://sourceforge.net/projects/rdp-classifier/

#database here http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/Diat.barcode/

rdp_classifier -Xmx8g classify -t frogs/diat_barcode_v10_all_rbcl_frogs.fasta.properties -o rdp.outputuclust -q otus.fa

Appendix II

The following scripts were run to process diatom metabarcoding data using the DADA2 pipeline.

Cutadapt can be installed on Windows, alternatively in Linux can be run as a loop where your sequences are:

```
for r1 in *_R1_*.fastq.gz; do

    # replace _R1_ in the file name with _R2_

    r2=${r1/_R1_/_R2_}

    out1=$(basename ${r1})

    out2=${out1/_R1_/_R2_}

    cutadapt -a
ATGCGTTGGAGAGARCGTTTC...CAGTTGTWGGTAAATTAGAAGGTGATC -A
AYGGTATCTRATCRTCTTYG...GAGCTGGAATTACCGCRG  --discard-untrimmed --
minimum-length 200 -o trimmed/${out1} -p trimmed/${out2} ${r1} ${r2}

done
```

DADA2 sequencing pipeline modified from https://github.com/fkeck/DADA2_diatoms_pipeline/blob/master/pipeline.R

```
# Ran in R studio. R version 4.0


library(dada2)

# Your path to your sequences here

path <- ("D:/Diatom data all/2018_raw/Run10/trimmed")

path_results <- file.path(path, "results")

if(!dir.exists(path_results)) dir.create(path_results)


# Set patterns to discriminate your forward and reverse read files

#Note for Environment agency data this coding was different for some runs
```

```
file_pattern <- c("F" = "_R1_001.fastq", "R" = "_R2_001.fastq")


fas_Fs_raw <- sort(list.files(path, pattern = file_pattern["F"], full.names = TRUE))

fas_Rs_raw <- sort(list.files(path, pattern = file_pattern["R"], full.names = TRUE))


path_process <- ("D:/Diatom data all/2018_raw/Run10/trimmed") # If you skipped primers removal, provide the path to your sequences here


fas_Fs_process <- sort(list.files(path_process, pattern = file_pattern["F"], full.names = TRUE))

fas_Rs_process <- sort(list.files(path_process, pattern = file_pattern["R"], full.names = TRUE))

#Note sample names cannot be identical

sample_names <- sapply(strsplit(basename(fas_Fs_process), "_"), function(x) x[1])


#### Inspect read quality profiles ####

plotQualityProfile(fas_Fs_process[2])

plotQualityProfile(fas_Rs_process[2])


pdf(file.path(path_results, "Read_quality_profile_aggregated.pdf"))

  p <- plotQualityProfile(sample(fas_Fs_process, replace = FALSE,

                         size = ifelse(length(fas_Fs_process) < 100, length(fas_Fs_process), 100)),

                  aggregate = TRUE)

  p + ggplot2::labs(title = "Forward")

  p <- plotQualityProfile(sample(fas_Rs_process, replace = FALSE,
```

```
                    size = ifelse(length(fas_Rs_process) < 100, length(fas_Rs_process),
100)),
                aggregate = TRUE)
  p + ggplot2::labs(title = "Reverse")
dev.off()



#### FILTER AND TRIM ####
fas_Fs_filtered <- file.path(path, "filtered", basename(fas_Fs_process))
fas_Rs_filtered <- file.path(path, "filtered", basename(fas_Rs_process))
all.equal(basename(fas_Fs_raw), basename(fas_Fs_filtered))

names(fas_Fs_filtered) <- sample_names
names(fas_Rs_filtered) <- sample_names

# Settings below are conservative for poor quality data
out_2 <- filterAndTrim(fas_Fs_process, fas_Fs_filtered, fas_Rs_process, fas_Rs_filtered,
                truncLen = c(200, 170), maxN = 0, maxEE = c(2, 2), truncQ = 2,
                rm.phix = TRUE, compress = TRUE, multithread = TRUE)
head(out_2)



#### LEARN THE ERROR RATES ####
error_F <- learnErrors(fas_Fs_filtered, multithread = TRUE, randomize = TRUE)
error_R <- learnErrors(fas_Rs_filtered, multithread = TRUE, randomize = TRUE)
```

```
pdf(file.path(path_results, "Error_rates_learning.pdf"))
  p <- plotErrors(error_F, nominalQ = TRUE)
  p + ggplot2::labs(title = "Error Forward")
  p <- plotErrors(error_R, nominalQ = TRUE)
  p + ggplot2::labs(title = "Error Reverse")
dev.off()



#### DEREPLICATION, SAMPLE INFERENCE & MERGE PAIRED READS ####
merged_list <- vector("list", length(sample_names))
names(merged_list) <- sample_names

for(i in sample_names){
  cat("Processing -------", which(sample_names == i), "/", length(sample_names), "-------", i,
"\n")
  derep_Fs <- derepFastq(fas_Fs_filtered[[i]], verbose = TRUE)
  derep_Rs <- derepFastq(fas_Rs_filtered[[i]], verbose = TRUE)
  dds_Fs <- dada(derep_Fs, err = error_F, multithread = TRUE, verbose = TRUE)
  dds_Rs <- dada(derep_Rs, err = error_R, multithread = TRUE, verbose = TRUE)
  merged_list[[i]] <- mergePairs(dds_Fs, derep_Fs, dds_Rs, derep_Rs, verbose = TRUE)
}

#### CONSTRUCT SEQUENCE TABLE ####
seqtab <- makeSequenceTable(merged_list)
dim(seqtab)
table(nchar(getSequences(seqtab)))
```

```
saveRDS(seqtab, file="seqtab.rds")
```

The easiest way is to use R's native facilities for saving and reloading objects:

```
#### REMOVE CHIMERA ####

seqtab_nochim <- removeBimeraDenovo(seqtab, method = "consensus", multithread =
TRUE, verbose = TRUE)

dim(seqtab_nochim)

table(nchar(getSequences(seqtab_nochim)))


#### ASSIGN TAXONOMY ####

# Here we download and use the Diat.barcode (last version) pre-processed for DADA2.

# You can use your own local database if needed.

tax_fas <- diatbarcode::download_diatbarcode(flavor = "rbcl312_dada2_tax")

tax <- assignTaxonomy(seqtab_nochim, tax_fas$path, minBoot = 75,

                      taxLevels = c("Empire", "Kingdom", "Subkingdom", "Phylum", "Class",
"Order", "Family", "Genus", "Species"),

                      outputBootstraps = TRUE, verbose = TRUE, multithread = TRUE)


spe_fas <- diatbarcode::download_diatbarcode(flavor = "rbcl312_dada2_spe")

exact_sp <- assignSpecies(seqtab_nochim, spe_fas$path)


#### CLEAN AND SAVE EVERYTHING ####

prep_cdm <- function(x){

  x <- t(x)
```

```
  x <- as.data.frame(x)
  x <- cbind(rownames(x), x)
  colnames(x)[1] <- "DNA_SEQ"
  return(x)
}

write.csv(prep_cdm(st.all), file.path(path_results, "sequence_table.csv"), row.names =
FALSE)
write.csv(prep_cdm(seqtab), file.path(path_results, "sequence_table_nochim.csv"),
row.names = FALSE)
tax <- as.data.frame(tax)
tax <- cbind(rownames(tax), tax)
colnames(tax)[1] <- "DNA_SEQ"
colnames(tax) <- sub("^tax\\.", "", colnames(tax))
colnames(tax) <- sub("^boot\\.", "BOOT_", colnames(tax))
write.csv(tax, file.path(path_results, "seq_nochim_tax.csv"), row.names = FALSE)

exact_sp <- as.data.frame(exact_sp)
exact_sp <- cbind(rownames(exact_sp), exact_sp)
colnames(exact_sp)[1] <- "DNA_SEQ"
write.csv(exact_sp, file.path(path_results, "seq_nochim_exact_sp.csv"), row.names =
FALSE)
```

Would you like to find out more about us or your environment?

Then call us on

03708 506 506 (Monday to Friday, 8am to 6pm)

Email: enquiries@environment-agency.gov.uk

Or visit our website

www.gov.uk/environment-agency

incident hotline

0800 807060 **(24 hours)**

floodline

0345 988 1188 **(24 hours)**

Find out about call charges (https://www.gov.uk/call-charges)

Environment first

Are you viewing this onscreen? Please consider the environment and only print if absolutely necessary. If you are reading a paper copy, please don't forget to reuse and recycle.