Diatom Recording Using Metabarcoding

Thesis submitted for the degree of Doctor of Philosophy by

Mathieu Ramon

School of Natural and Environmental Sciences

Newcastle University, Newcastle-upon-Tyne

United Kingdom

October 2023

ABSTRACT

Freshwater monitoring is crucial to preserve the ecological services these ecosystems provide. Diatoms are known to be reliable indicators of water quality, hence the historical analysis of their community for routine biomonitoring.

This thesis aims to optimise the current diatom biomonitoring method based on the most recent metabarcoding tools, in order to assist freshwater environment surveillance. Light Microscopy (LM) coupled with morphological identification is the traditional approach for diatom surveys. The more recent alternative, metabarcoding, combines barcoding to identify species using DNA variations from short conservative sequences (barcodes) and High Throughput Sequencing (HTS), that allows the analysis of thousands of sequences simultaneously.

Aspects of the methodology, including primers, were tested in a variety of environments, including rivers and mesocosms, enabling the optimisation of the whole process and confirmation of the reliability of the short barcode located in the *rbcL* gene. This experimentation showed the interchangeability of LM and metabarcoding approaches for most routine diatom biomonitoring surveys.

Both Illumina and MinION HTS platforms were compared to the LM method and judged to be a success but no benefit was found using a longer *rbcL* barcode region with MinION. Bioinformatic pipelines were created for each sequencing technology, based on new bioinformatic tools and particularly the denoising/polishing algorithms which generated equal or better results than the current QIIME1 bioinformatic pipeline. As an appropriate reference library is crucial for taxonomic assignment of sequences, the current UK reference library was compared with and updated from the European reference library, diat.barcode. In addition, non-diatom phytoplankton taxa were added to the reference library, improving the species assignment.

The evolutionary history of the barcode region, the *rbcL* gene, was investigated using a phylogenetic approach. This demonstrated the link between *rbcL* evolution and diatom morphology, and its suitability as a 'barcode' were discussed.

This project succeeded in improving diatom biomonitoring via HTS, and further demonstrated the reliability of this approach.

ACKNOWLEDGEMENT

This work was funded by IAFRI, a collaboration between FERA and Newcastle University.

I would like to express my thanks to my supervisors, Eleanor Jones, Edward Haynes, and Maxim Kapralov, for their guidance, help, support, and patience throughout this thesis project. You even literally saved my life, and I do not think this was part of the job you signed up for. I particularly want to acknowledge Martyn Kelly and Frederic Rimet for their guidance in diatom ecology.

I would also like to extend personal thanks to the FERA staff for their help and friendly support, especially Ines, Benito, Hollie, Matias, Charlie, Marco, Valeria, and Aimee. You were all part of the best moments of this rocky road. I would like to thank my real family (Mamiche, Freddie, and Anouk) and my "Nous" family for keeping me motivated and supporting me during this adventure. It would not have been possible without you. Finally, I would like to express massive thanks to Marion for being such a tremendous support and source of motivation every day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IBD - Indice Biologique Diatomées (AFNOR NF T90-354)

IPS – Indice de polluosensibilité spécifique (Coste in CEMAGREF, 1982)

LM – Light Microscopy

ONT - Oxford Nanopore technologies

OTU - Operational taxonomic unit

PCR - Polymerase chain reaction

TDI – Trophic Diatom Index (Kelly et Whitton, 1995)

WFD - Water Framework Directive

DECLARATION

I declare that this thesis was written by myself and that the work within is my own unless explicitly stated otherwise. The work in this thesis has not been submitted for other degrees or qualifications.

PUBLICATIONS ARISING FROM THIS WORK

Diatom DNA Metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization.

Bonnie Bailet, Laure Apothéloz-Perret-Gentil, Ana Baričević, Teofana Chonova, Alain Franc, Jean-Marc Frigerio, Martyn Kelly, Demetrio Mora, Martin Pfannkuchen, Sebastian Proft, Mathieu Ramon, Valentin Vasselon, Jonas Zimmermann, Maria Kahlert

Science of the Total Environment 745, 140948, **2020**

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

Maria Kahlert, Satu Maaria Karjalainen, Francois Keck, Martyn Kelly, Mathieu Ramon, Frederic Rimet, Susanne Schneider, Kálmán Tapolczai, Jonas Zimmermann

Ecological Indicators 141, 109114, **2022**

CHAPTER 1 INTRODUCTION

Biomonitoring is the routine sampling of an environment to analyze the composition of its biocenosis in order to evaluate its ecological status and assess the quality of the environment. The key ecosystems that comprise freshwater bodies are among the most monitored ecosystems due to their relative rarity (3% of the total water on Earth, of which only a quarter is in a liquid state) compared to the ecological niches they provide to both animal and plant species as well as the ecosystem services that they offer to civilization, such as filtration, water withdrawal for agriculture and human consumption, or fisheries. Hence, the monitoring of freshwater bodies is a necessity to protect them from pollution that threatens their ecological values and thereby the quality of the ecosystem services they provide.

Diatoms are unicellular algae that are present in all aquatic ecosystems and responsible for around 20% of global primary production (Mann, 1999). Due to their ecological preferences and rapid growth diatoms are excellent bioindicators of water quality. Therefore, the biomonitoring of diatoms has been chosen as a part of the ecological assessment of UK rivers in the context of the water framework directive (Kallis and Butler, 2001)

The traditional assessment method used in the UK relies on Light Microscopy observations of the biofilm samples found on river rocks to determine the community composition present in the rivers. Although this approach has a high reliability it is also very time consuming, expensive and requires highly skilled experts. Several studies have highlighted the potential of environmental DNA Metabarcoding approach to give fast and accurate composition of the diatom community (Duleba et al., 2021; Vasselon et al., 2017a, 2017c; Zimmermann et al., 2015), and it has been trialed in the UK (Kelly et al., 2018). Notwithstanding the potential of this molecular approach, some improvements and evaluations are needed to allow the method to be used routinely. The aim of this study is to improve the diatom biomonitoring method in order to create a standard methodology for environmental managers, including steps from the sampling to the environment quality index calculation. The study of diatom DNA sequences allows me to reconstruct the evolutionary history of the *rbcL* gene used as a barcode in my method in order to quantify the positive selection effect on the *rbcL* gene, which enables me to evaluate its adequacy as barcode.

Diatom biomonitoring is fueled by the Water Framework Directive (WFD) that set mandatory surveys of the water quality of the rivers to identify the trophic state of each waterbody and the pressure affecting them. Eutrophication was described by European policy as a priority concern for water management (European Commission, 2021).

EUTROPHICATION OF FRESHWATER ENVIRONMENT

The monitoring of freshwater environments is motivated by the increase of water quality degradation originating frequently from eutrophication and pollution (e.g., heavy metals, pesticide, organic components, etc.). These degradations have impacts on water ecosystem services, hence the importance of their maintenance for human societies (Grizzetti et al., 2016).

Trophic state		Total Nitrogen ($mg.L^{-1}$)	Total Phosphorus ($mg.L^{-1}$)
Lakes	*Oligotrophic*	<0.35	<0.01
	Mesotrophic	0.35 - 0.65	0.01 - 0.03
	Eutrophic	0.65 - 1.20	0.03 - 0.10
	Hypertrophic	>1.20	>0.10
Streams	*Oligotrophic*	<0.70	<0.025
	Mesotrophic	0.70 - 1.50	0.025 - 0.075
	Eutrophic	>1.50	>0.075

Table 1 Classification of the trophic state of bodies of water (Dodds et al., 1998; Nürnberg, 1996)

Eutrophication is defined as the process by which nutrients accumulate in a habitat and change the state of the ecosystem to another higher trophic state (Figure 1). The nutrients of most concern are Nitrogen and Phosphorus as they limit the growth factor of plant and algae (Rabalais, 2002)Table 1). Eutrophication generates outcomes on both biocenosis and biotope. In aquatic ecosystems, eutrophication created by the high availability of nutrient leads to an excess growth of plant and algae (blooms) that block sunlight and affect the whole ecological

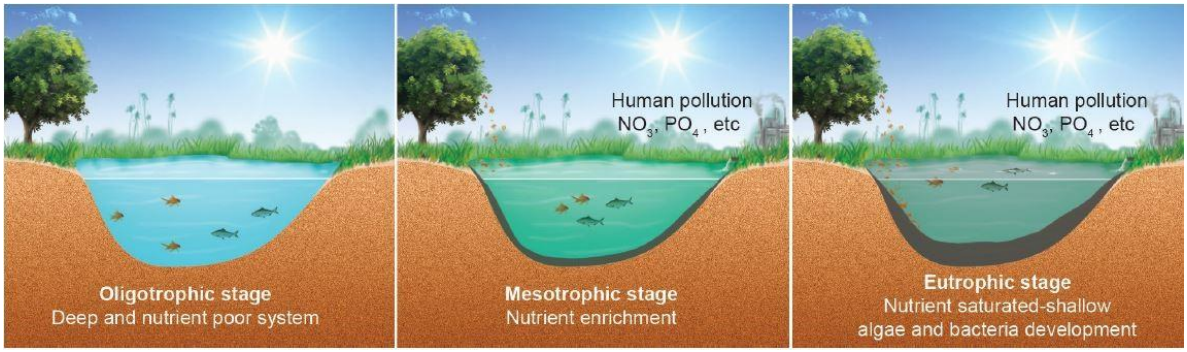


Figure 1 Three different stages of eutrophication, from left to right. From Civilpedia

community. Due to the aerobic decomposition of algae that consumes large quantities of oxygen, hypoxia and anoxia can also be an outcome and lead to death of aquatic species including fish and amphibians. Environmental managers monitor the eutrophication of aquatic ecosystems due to its deleterious consequences on water quality for human use (turbidity, unpleasant odour, toxicity, etc.) and for the ecosystem itself (anoxia, unbalanced communities with overgrowth of hypercompetitive species).

The eutrophication of waterbodies has been a major concern since the 1970s (Lund, 1972), from which arose the need to monitor the chemistry and the ecological states of both running and still water bodies. Although numerous ecosystem services are provided by waterbodies (Table 2) eutrophication can directly decrease the quality of the services provides by the waterbodies (Environment Agency, 2016)

The trophic state of streams and lakes can be assessed directly by measuring nutrient water concentrations or by biomonitoring (e.g., algae, aquatic plants, fish)(Brooks et al., 2001; Danilov and Ekelund, 2001; Kelly et al., 2008; Muscutt and Withers, 1996; Schneider and Lindstrøm, 2011). The advantage of biomonitoring over measuring nutrient water concentration is that the composition of communities is an indicator of the trend of the concentration of the nutrient rather than a snapshot of a particular moment (Li et al., 2010). Waterbodies are very unstable environments due to weather (rain/drought events) and human activities (e.g., one-time point source pollution)(Riley et al., 2018). The biomonitoring approach, by assessing the quality indirectly via the community resulting from all the conditions affecting an ecosystem, provides a result less perturbed by the very short-term nutrient variation created by all the individual events affecting the environment (Moog et al., 2018).

Water Supply	Supply of goods other than water	Non-extractive or instream benefits
Household uses including drinking, cooking, washing. Industrial uses including manufacturing, thermoelectric power generation. Irrigation Aquaculture	Fish Waterfowl Clams and mussels	Flood control Transportation Recreational swimming, boating, etc. Pollution dilution Bioremediation and phytoremediation Water quality protection Hydroelectric generation Wildlife habitat Soil fertilization Enhanced property values Non-user values

Table 2 Ecosystem services provided by freshwater bodies and wetlands. Adapted from Daily, 1999.

Although river biomonitoring is an efficient tool to detect eutrophication, the community that is monitored must be adapted to the aspect of the environment that is the subject of the study. For example, the fish communities analysis would not use the same techniques nor generate the same conclusion as the analysis of phytoplankton communities (Keck et al., 2017). In river biomonitoring the principal communities studied are:

1. **Fish**: Difficult to sample from the environment due to high mobility. The identification of their communities is rather straightforward due to their large body size and the fact that the fish diversity of a single environment is relatively low compared to other taxonomical groups. They are particularly good indicators of long-term pollution and hydromorphology of the waterbodies (Physical barriers such as waterfall or dams)(Cooper et al., 2016; Okwuosa et al., 2019).
2. **Macrophytes**: Their absence of mobility can aid sample collection, but their possibly deep-water location can sometimes complicate this. Their identification is well documented and straightforward. They are good indicator of eutrophication, turbidity (caused by organic and inorganic material), riparian zone integrity and hydromorphology (Bresciani et al., 2012; Carbiener et al., 1990; Tarkowska-Kukuryk and Mieczan, 2017).
3. **Benthic macroinvertebrate** (fixed to the bottom of the body of water): Their intermediate size facilitates their sampling and identification. They are considered

as good indicator for both short- and long-term organic pollution as they have responses to both pollution accumulations and single pollution events. They can be an indicator of habitat loss or heterogeneity (Lepori et al., 2003; Maitland et al., 2020; Mir et al., 2021).

4. **Benthic Diatoms** (fixed to the bottom of the body of water): Their small size and their particular location, fixed to the sediment and rock of the shallow river margins, make sampling straightforward. Their size is a disadvantage for identification, as well as the high average diversity present in rivers. They are known to be excellent indicators of eutrophication, pollution and hydromorphology of the river. They are more sensitive to long and midterm condition than one-time pollution events (Kelly et al., 2008; M. G. Kelly et al., 1998; Prygiel et al., 2002).
5. **Phytoplankton**: As this group can be sampled directly from the water it is convenient to collect. Nevertheless, the smallness of their size coupled with the high diversity of organisms to consider leads to arduous identification. However, they are good indicators of eutrophication (Danilov and Ekelund, 2001; Emiliani, 1997; Jacquet et al., 2005).

Thus, the taxonomic group of interest during a river biomonitoring study directly impacts the difficulty of the different steps (e.g., sampling, identification, etc.) and the possibility to assess particular ecological aspects. Due to their ease of collection and suitability as indicators, diatoms have been considered and used as indicators of choice for river biomonitoring to evaluate the pollution and eutrophication affecting aquatic environments (M. Kelly et al., 1998; Pandey et al., 2017). Thence, numerous ecological indexes based on diatom communities have been developed and used for routine water quality assessment. In the UK the standard is the Trophic Diatom Index (TDI) (Kelly, 1998; Kelly and Whitton, 1995), other European indexes exist, including the “Diatom Biologic Index” (Indice Biologique Diatomée, IBD)(Prygiel and Coste, 2000).

DIATOM CHARACTERISTICS

Diatoms are a class of single cell algae and phytoplankton. They phylogenetically form the phylum Bacillariophyta which is part of the subkingdom Heterokont (Round et al., 1990). Diatoms are present in most (if not all) aquatic ecosystems of any size and salinity (Mann, 1999). They are frequently described as ubiquitous due to this extremely large distribution (Mann and Vanormelingen, 2013). Although they are numerous, diatoms are among the more diversified groups with around 12 000 described species and an estimated 100 000 species in total (Guiry, 2012; Mann and Vanormelingen, 2013). Fossil traces of diatom silica envelopes

have been dated back to 185 million years ago (Gross, 2012). The diversity and ubiquity of diatoms are some of the reasons that make them biomonitoring indicators of choice for the water environment.

As part of the single cell algae, diatoms are photosynthetic Eukaryota. As such, they contain organelles such as chloroplasts, vacuoles, Golgi complexes, nucleus and mitochondria (Figure 2) (Herringer et al., 2019). Their chloroplasts originate from a secondary endosymbiosis with a red alga (Figure 3) (Nonoyama et al., 2019).

Either the host or the endosymbionts may have possessed genes retained from a cryptic endosymbiont of green algal origin, although this remains debated (Moustafa et al., 2009). Overall, there is a hypothesis that chloroplast-targeted proteins from contemporary diatoms have a bacterial origin, either in the host or symbiont, and have evolved from the event of endosymbiosis (Prihoda et al., 2012). This particularity results in the genome of diatoms containing genes of two eukaryotic nuclei and two prokaryotic genomes, one from the mitochondria and the one from the chloroplast (Konur, 2020). Moreover, the chloroplasts present in diatoms, as well as in Stramenopiles, Haptophytes, and Cryptomonads, are surrounded by four membranes instead of the common two membranes. (Bedoshvili et al., 2009).

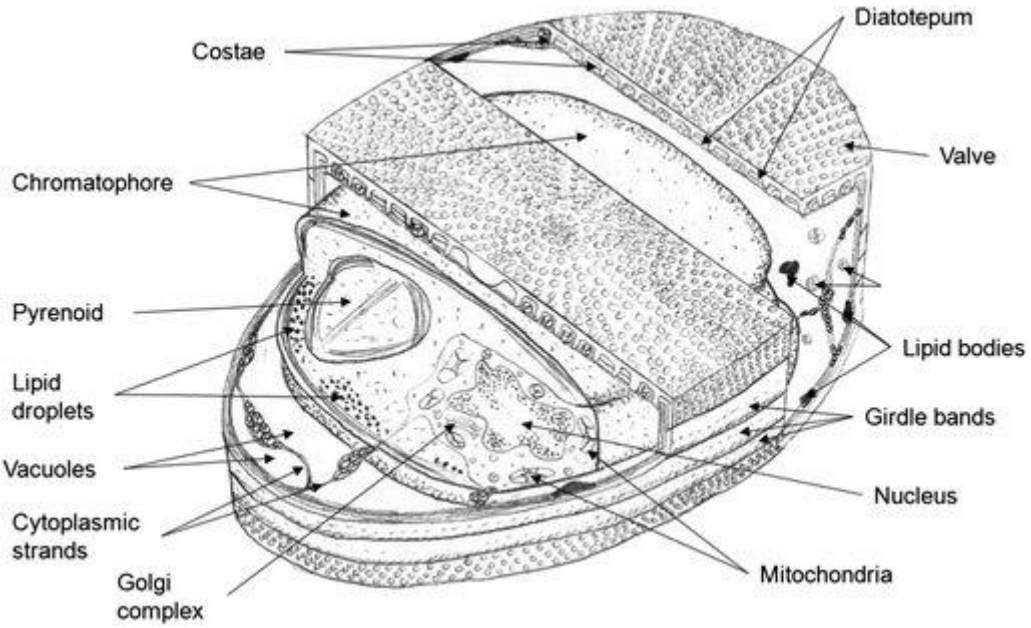


Figure 2 Schematic overview of internal and external structures of Diatom. From Herringer et al., 2019

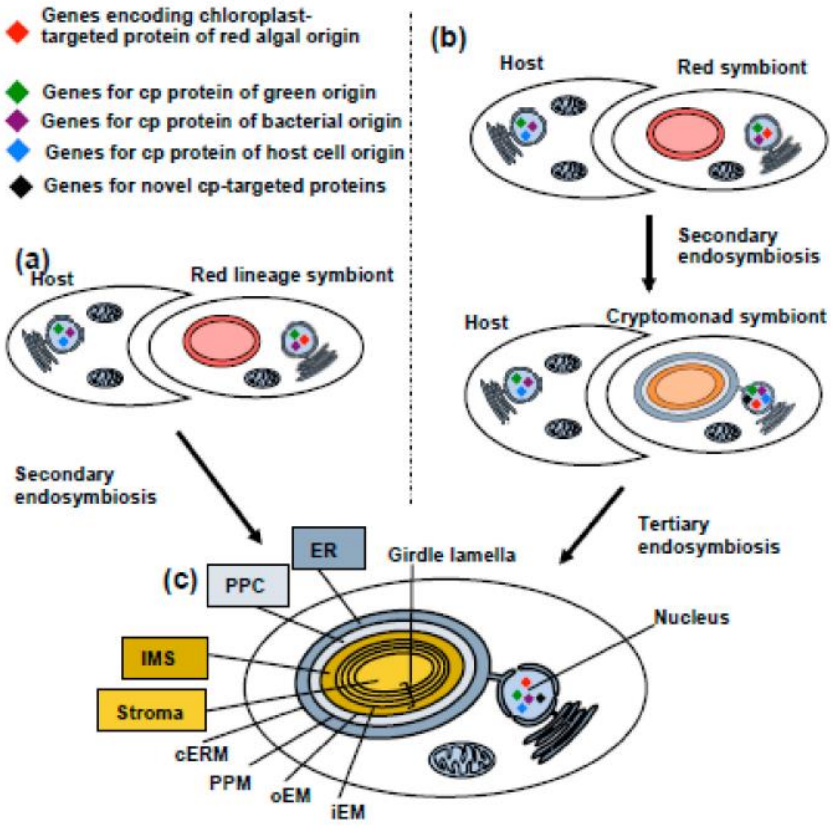


Figure 3 Origins and structure of the diatom chloroplast. This schematic Figure shows two alternative hypotheses for the origins of the diatom chloroplast: (a) secondary endosymbiosis of a red alga by a common ancestor of photosynthetic stramenopiles or (b) tertiary endosymbiosis of a cryptomonad-like organism, itself harbouring a chloroplast of secondary, red algal endosymbiotic origin. (c) shows a schematic diagram of the four membranes surrounding the diatom chloroplast. Abbreviations are as follows: cERM; chloroplast endoplasmic reticular membrane; ER, endoplasmic reticulum; iEM, inner envelope membrane; IMS, intermembrane space; oEM, outer envelope membrane; PPC, periplastid compartment; PPM, periplastid membrane. From Nonoyama et al., 2019.

Diatoms have an exoskeleton, called a frustule, made of hydrated silica that lets the light pass through this layer which enable the diatoms to perform photosynthesis and benefit from the protection of a shell (Figure 4) (Aguirre et al., 2018). Diatom species have a high variety of pigments; chlorophyll a and c (replacing the chlorophyll b from higher plants), fucoxanthin, carotenoids such as β-carotene, diadinoxanthin and diatoxanthin, violaxanthin, antheraxanthin, zeaxanthin or even marennine (Kuczynska et al., 2015). This mix of pigment is the cause of their golden-brown color which leads to frequently referring to them as brown microalgae. Moreover, their photosynthetic machinery is especially efficient at absorbing light, not only in the red and blue wavelengths but also in the green wavelengths, which is

notable among other photosynthetic groups(Goss et al., 2020). This efficiency contributes to making them one of the biggest contributors to global primary production. In fact, diatoms are part of phytoplankton, the latter representing as little as 1% of Earth's photosynthetic biomass but contributing to around 45% of the global primary production (Field et al., 1998). More specifically, marine diatoms contribute to half of the primary production of phytoplankton, which make them responsible of around 20% of the atmospheric oxygen production (Mann, 1999).

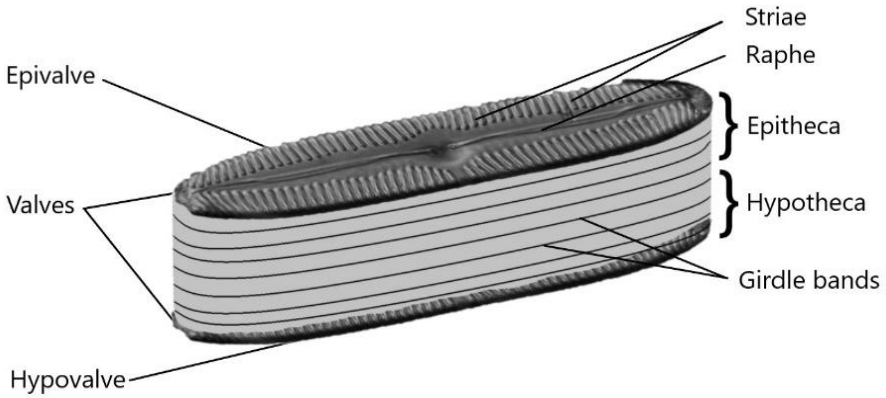


Figure 4 Diagram of a Diatom frustule structure, *Pinnularia sp*.

Research suggests that the silica wall is among the reasons why diatoms have been so successful in term of their widespread distribution and high contribution to primary production (Vasselon et al., 2017c). Frustule production is energy efficient and the silica is an advantage over Carbonate biomineral shells (present in other phytoplankton group such as Coccolithophores) (Bach et al., 2015) of other taxa group as it is not sensitive to ocean pH (Cermeño et al., 2015). The diatom frustule is the basis of morphological identification and sometimes complemented with organelles observation for living diatoms (Aguirre et al., 2018; Jones et al., 2005; Meyer et al., 2012). The durability of the diatom after its death may take account for the mismatch between microscopy-based identification and diatom nucleic acid-based identification (as discussed in Chapters 3, 4 and 5).

The key enzyme of photosynthesis is Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), which catalyzes the Carbon fixation of Carbon dioxide to form glucose or other energy-rich molecules. RuBisCO is the most abundant protein on earth and studies (Kapralov and Filatov, 2007; Young et al., 2012) have found that algal RuBisCO gene evolution was under Darwinian selection. Nevertheless, diatom RuBisCO evolution has received very little

attention even though the analysis of this gene could increase our understanding of the reason of the success of the diatom group in term of primary production (Chapter 4).

Diatoms are divided in two distinct groups based on their shape (Figure 5): centric diatoms that present a radial symmetry and pennate diatom that are bilaterally symmetric. Pennate diatoms are further divided in groups according to the presence of a raphid, a slit within the silica cell wall, which can be axial, eccentric, circumferential or absent (Araphidae) (Lange-Bertalot et al., 2017).

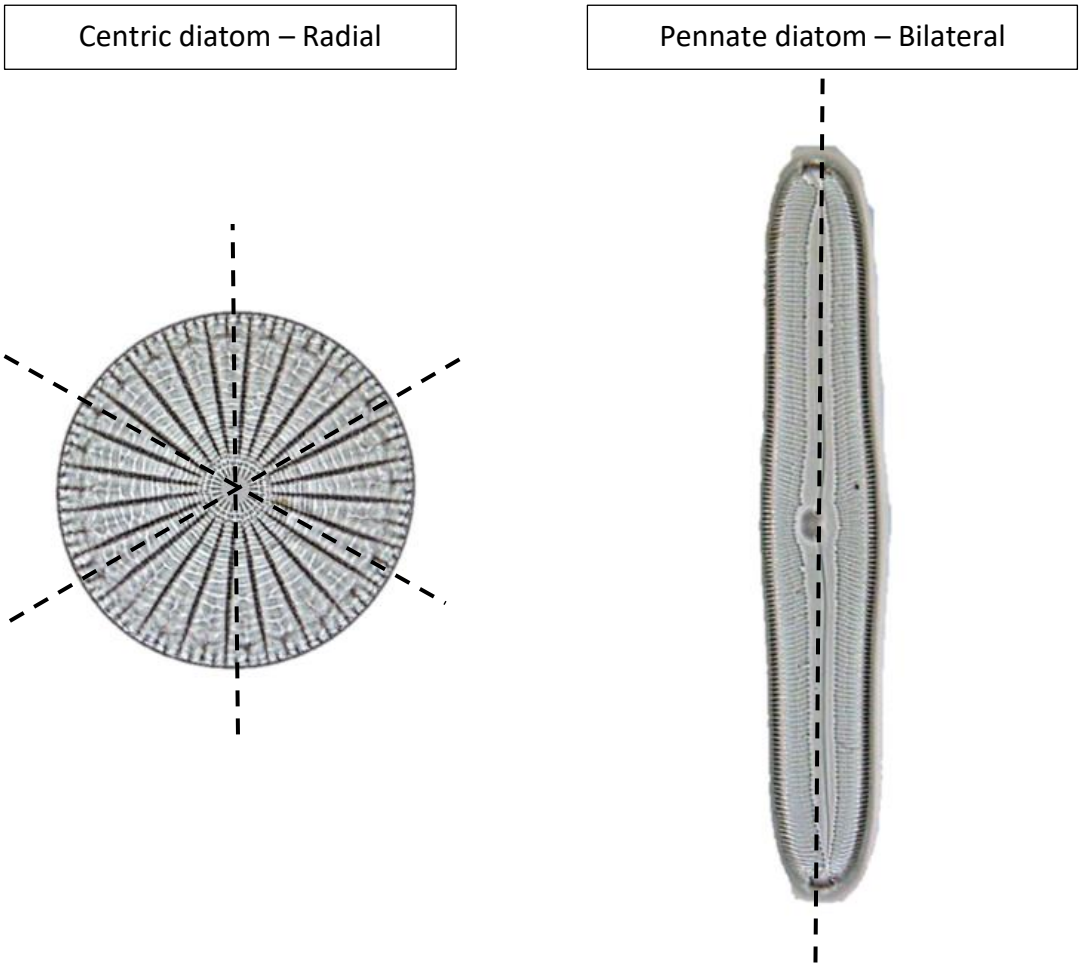


Figure 5 Symmetry comparison of the two main groups of Diatom: Centric (left) and Pennate (right)

DIATOM REPRODUCTION

Diatom reproduction can be sexual and asexual.

The primary form of reproduction is the asexual fission, a mitosis that generates two new smaller diatoms from the binary fission of one bigger diatom(Chepurnov et al., 2004). As at mitosis, it is preceded by a replication of the DNA which leads to the division of each

chromosome in two halves. The DNA is then separated in two, with one half of each chromosome present in each part. The formation of two frustules starts around these two parts, with each daughter cell receiving a theca (half of a frustule, see Figure 6; Montresor et al., 2016) from the initial frustule. Each daughter cell builds a new theca to complete the one given by the parent cell. The new theca is always smaller than the one given by the parent cell (the new one is always the Hypotheca and the old one the Epitheca). This kind of reproduction leads to the production of smaller and smaller cells, since each reproduction produces one smaller cell and one with the same dimension of the parent cell (Sánchez et al., 2019). This is not a viable long-term reproduction method and as such, the sexual reproduction is present in order to, but not exclusively to, manage the shrinkage in the size of cell generation after generation.

As the diatom vegetative phase is in a diploid form, they need to undergo a meiosis phase to produce gametes(Chepurnov et al., 2004). Males produce flagellate sperms and female produce eggs that will form a zygote after they meet and fuse. In order to do that the female cell creates an opening in its cell wall. The fertilized egg creates an envelope with its own cell wall and nucleus. This will cause the new diatom to grow to its full size and to form with the parent diatom an auxospore which could be set to a dormant stage during which it is called a "resting spore" (Pelusi et al., 2020). This dormant cell type is able to survive under unfavorable conditions during extended periods of time and awaken when the conditions are more optimal.

In all diatom life cycles there is an auxospore/resting spore phase (Tréguer et al., 2017). It is very noticeable that auxospore can be present in the environment whereas no mature diatom is present. In such cases the DNA of the diatom species would be present in the environment although the mature diatom frustule cannot be found in the environment (Sanyal et al., 2022). This mismatch may account for some of the differences between microscopy and Metabarcoding results, as discussed in Chapters 3 and 4.

As discussed before, diatoms present a silicate exoskeleton called a frustule that persists in the environment and are uncomplicated to isolate from biofilm or sediment. Therefore, diatom identification using Light Microscopy is reliable and does not require elaborate tools (Kelly, 1998). Hence the historical use of diatoms for biomonitoring due to the accessibility of the frustule and the identification based on the shape of the frustule (Round et al., 1990). Although the presence of the frustule is an advantage for the Light Microscopy approach, the frustule could be an obstacle for a DNA-based approach because this extra silicate layer makes the DNA less accessible compared to other algal taxa (Mora et al., 2019). Thus, it is sensible to measure the proportion of each diatom and non-diatom DNA obtained from a known sample and to compare the morphological count in order to reveal overrepresentation and underrepresentation of diatom and non-diatom taxon during Metabarcoding survey

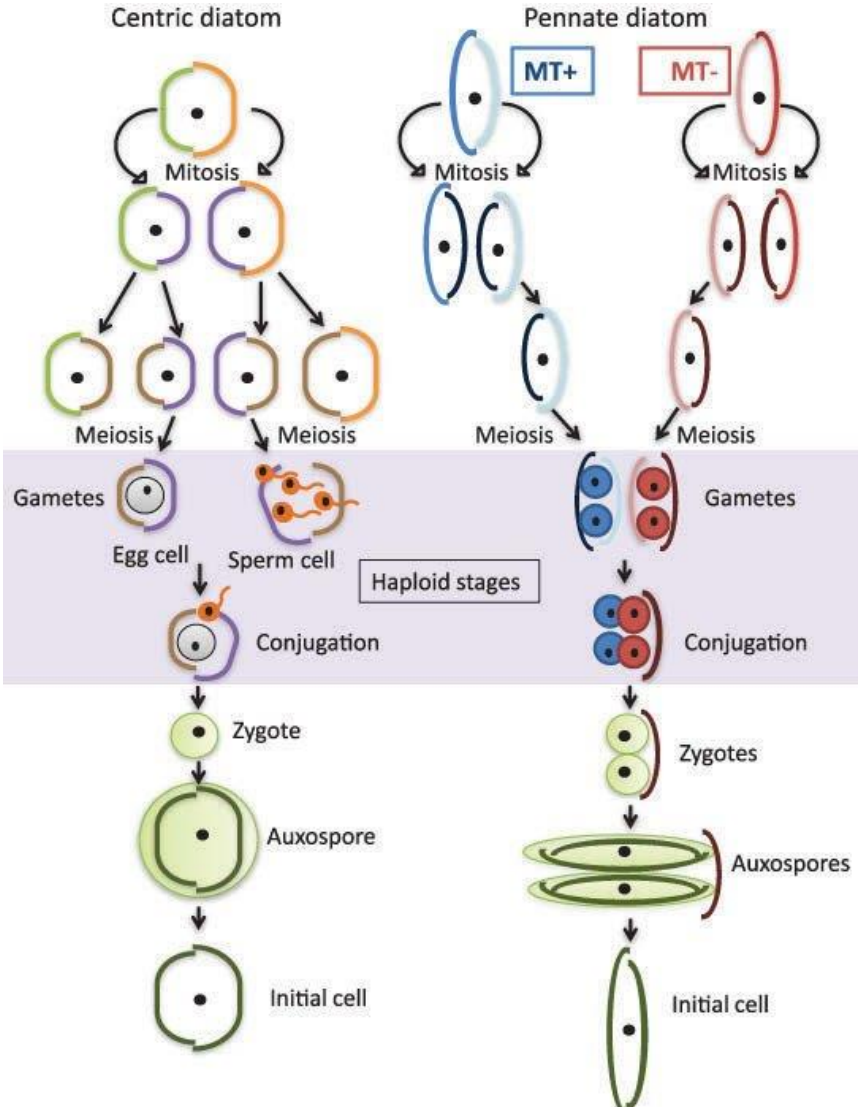


Figure 6 Schematic drawing of the life cycle of a centric and a pennate diatom. From Montresor et al. 2016.

(Vasselon et al., 2018). This is useful to apply potential correction factors before calculating ecological indexes. It is further discussed in the Chapter 3.

METABARCODING

DNA barcoding is a method of species discovery and identification based on a short sequence of DNA from a predetermined area of the genome. That DNA sequences, called a “barcode”, can be used to identify an organism to taxon. Specific regions in the genome have been found to be informative for distinguishing taxa from each other. The first of these “standardised regions” was the mitochondrial gene CO1 (Hebert et al., 2003).

The increasing use of High Throughput Sequencing (HTS) (see next section) makes possible the combination of barcoding with HTS to identify simultaneously different taxa from a single sample, for example environment samples such as water, biofilm, or soil. This kind of samples contains environmental DNA (eDNA) composed of the DNA from individual organism present in the sample along with the DNA released and accumulated from the surrounding organisms or previously present organisms. This method is called Metabarcoding and permits the analysis of the whole community of the samples rather than target single species (Taberlet et al., 2012).

Discovery of new species as well as identification of species in an environment can be performed by DNA Metabarcoding. Nevertheless, a preliminary database creation is the backbone of a good barcoding identification (Kelly et al., 2018). This is done by sequencing the standardised regions of morphologically verified voucher specimens to create a reliable database. In order to assign identities to DNA sequences a reference library is needed, the largest is the Barcode of Life Data Systems (BOLD) (Hebert and Ratnasingham, 2007). Alternatively, specified reference library can be created to be optimized to the organism and barcode targeted and provide a better taxonomic assignment.

The traditional DNA Metabarcoding process is composed of an extraction of the DNA of an environmental sample, including DNA from living organism and/or DNA in the environment present in the sample, followed by a PCR amplification of a targeted barcode region. Then a sequencing run generates the DNA sequences of the whole community that can be identified bioinformatically using a reference library. Additional bioinformatic steps can be done to

correct errors generated during sequencing (incorrect or low-quality nucleotide reads) and PCR (chimeras). This results in an Operational Taxonomic Unit (OTU, grouping of individuals based on genetic similarities without relying on taxonomic rank) abundance table that are combined with the taxonomic assignment to generate a taxonomic list of the taxa present in each sample. Such abundance tables are the input needed for calculation of ecological indexes such as TDI.

HTS PLATFORMS FOR DIATOM BIOMONITORING

Different technologies of sequencing are available, but they are not all optimized for diatom biomonitoring. In the case of a Metabarcoding study both read length and quality are determining factors for good taxonomic assignment (Pearman et al., 2020).

Many of the most recent diatom Metabarcoding studies use the MiSeq Illumina Sequencing by synthesis (SBS; Figure 7 & 8). An advantage of this platform is that the length of the diatom *rbcL* (coding for the large subunit of RuBisCO) amplicons mostly used (312 or 340 base pairs (bp)) matches the capacity of this technology to provide high quality reads up to this length (Kelly et al., 2018), and that the number of sequences reads generated by this technology is sizeable (several Gigabytes of data for a MiSeq run which represent up to 25 million reads) comparing to older technologies. This depth of read coverage allows reliable biomonitoring assessment because we can investigate abundant and rare species present in the ecosystem. Besides this method, the Ion Torrent™ PGM has also been used (Kermarrec et al., 2014; Vasselon et al., 2017b) but the higher errors rate and the lower coverage of sequences lead to a large proportion of discarded reads. The MiSeq was one of the platforms used in this study, for example in Chapters 3, 4, and 5.

New sequencing platform can be interesting alternatives, such as the Oxford Nanopore Technology (ONT) sequencing platform which enables longer reads but with a relatively lower quality. This sequencing technology is based on identification of the nucleotide by analysing the change in the electric current voltage density induced across a nanopore when a fragment of DNA passes through this micropore (Figure 7). The MinION device is the most used ONT platform for biomonitoring as it presents a good compromise between price (both the device itself and reagents) and number of reads. Moreover, the small dimensions of the MinION device itself make it very portable.

However, any new platform should be tested against the existing widely used platform (in this case, the MiSeq) to assess its performance characteristics, cost and speed of deployment. The Nanopore technology (and especially the MinION device) produce an average of 5% to 20 % error rate (Weirather et al., 2017; Jain et al., 2018; Tedersoo et al., 2018). The error rate is quite constant but remains high for biomonitoring which prefers low error rates to ensure the reliability of the taxonomic assessment. Nanopore and Illumina technology were compared with simulated reads of a large range of length (Pearman et al., 2019), and it appeared that for markers longer than 1500 bps the ONT platform will be more accurate for identification than the Miseq platform because the long marker (thousands of kilobases produced by the MinION compare to 350bp for MiSeq) are more reliable even if the quality is low. Comparison of ONT MinION and Illumina MiSeq during diatom Metabarcoding studies have been made (Glover, 2019) and results are mitigated by the use of an incomplete reference library coupled with BLAST assignment which lead to numerous unassigned sequences. Hence my motivation to design the experiment in the Chapter 3, which compared between MinION and MiSeq result in artificially created algal communities, after curation of the reference library and test of different amplicon.

Sequencing platforms still have limitations such as read length and read error rate. Therefore, the choice of the barcode region is crucial and must be optimised for the targeted organism(s) and the type of survey.

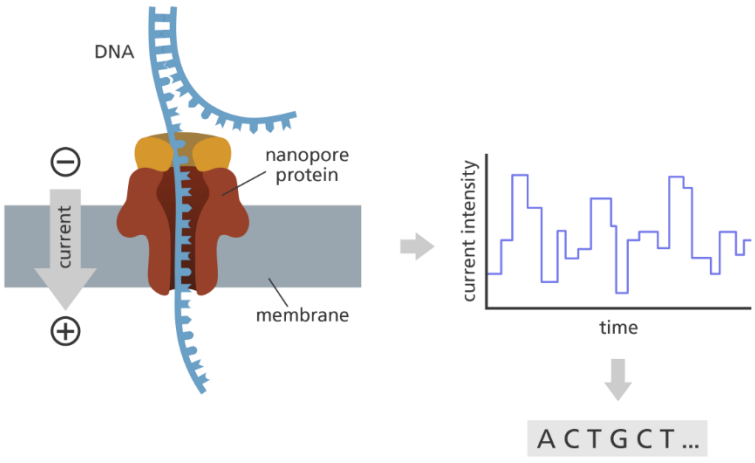


Figure 8 Nanopore Oxford Nanopore Technology (ONT) sequencing. From Genome Research Limited.

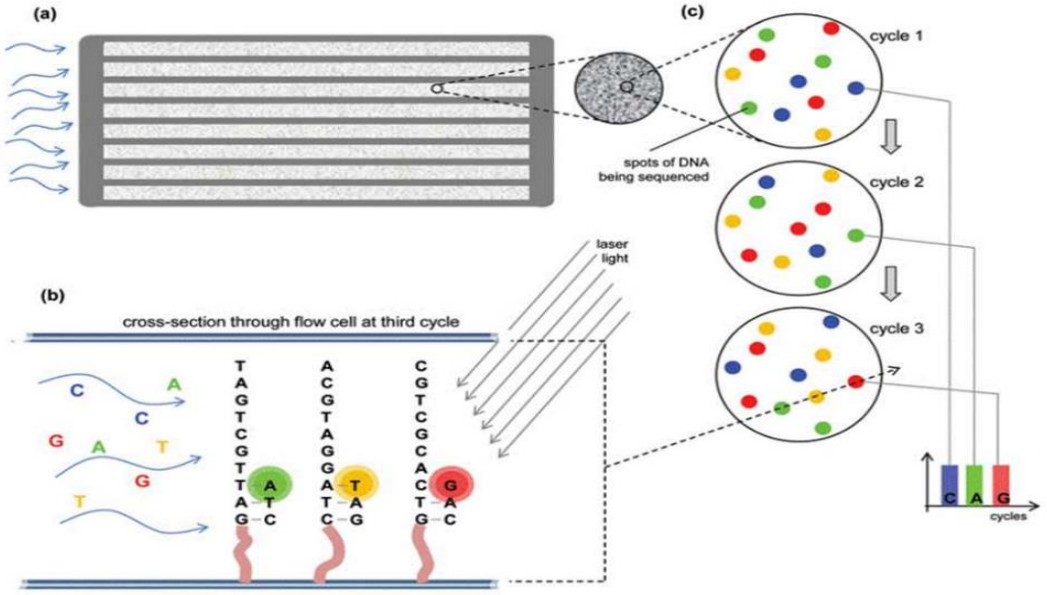


Figure 9 Principle of Illumina Sequencing (Sequencing by Synthesis). (a) flow cell overview; (b) incorporation of nucleotides results in release of fluorescence; (c) zoomed in the flow cell –different nucleotides with their specific fluorescents colour (modified by Untergasser after Genomics 2019).

TAXONOMIC REFERENCE LIBRARY

DNA-based identification efficiency is directly dependent on the reference library created beforehand. This is due to identification being based on similarities between the reads from the environment and the sequences present in the reference library. A percentage similarity threshold is present, meaning that an error in the reference library read (erroneous bases, gaps, confused naming, etc) can directly lead to unassigned sequences or incorrectly assigned sequences. A reference library is composed of sequences from securely identified species, usually extracted from pure culture.

A well curated and diverse reference library is the backbone of reliable studies based on diatom biomonitoring using Metabarcoding (Kelly et al., 2018). Although diatoms are among the most used bioindicator, the largest missing gaps in barcode reference libraries for freshwater biomonitoring are associated to diatoms and invertebrates (Weigand et al., 2019).

The UK Diatoms reference library, created for in Kelly et al 2018, suffers from a low coverage of species: only 176 diatoms species have their sequences present in the “Gold-Standard” reference library, which is less than 10% of the species present in the UK and Ireland. A large proportion of the most abundant and most ecologically informative species are included.

A second library is available which derives from a European project called diat.barcode. It uses an open-source reference library within 1401 sequences of diatoms including the Kelly et al. (2018) reference library. Furthermore, the sequences are well curated because of a regular curation of the library before each release of a new version. This European project diat.barcode/Rsyst, initially based in the French laboratory INRA – CARRTEL Thonon-les-Bains (Rimet et al., 2019, 2016) also provides methodology from sampling to bioinformatic analysis.
A second library is available which derives from a European project called diat.barcode. It uses an open-source reference library within 1401 sequences of diatoms including the Kelly et al. (2018) reference library. Furthermore, the sequences are well curated because of a regular curation of the library before each release of a new version. This European project diat.barcode/Rsyst, initially based in the French laboratory INRA – CARRTEL Thonon-les-Bains (Rimet et al., 2019, 2016) also provides methodology from sampling to bioinformatic analysis.

The performances of the two diatom reference libraries are compared in Chapter 1 as well as an improvement of the diat.barcode reference library by adding non-diatom phytoplankton sequences in order reduce the proportion of unassigned sequences. Limitations of the library from Kelly et al. (2018) are discussed in Chapter 3.

Assignment method

In metabarcoding analysis, two assignment methods are most commonly used. Firstly, the naive Bayesian classifier method developed by Wang et al. (2007), which assigns taxonomy across multiple phylogenetic ranks (e.g. to phylum, genus, species, clone, etc.) and secondly, the BLAST assignment method (Altschul et al., 1990) which is an heuristic method based on a similarity matrix to assign the tested sequences to the closest reference sequence. The BLAST assignment cannot assign to a hierarchical taxonomic placement, which is why Naïve Bayesian assignment is implemented in some of the most recent bioinformatic pipelines (Bolyen et al., 2019; Callahan et al., 2016; Schloss et al., 2009). Other alternative methods exist such as Kraken2 (Salzberg and Wood, 2014) with rely on exact k-mer matches in order to classify the sequences again a reference library.

An alternative approach is ‘taxonomy-free’ biomonitoring. An example of this is a Swiss study which tested the approach for diatoms biomonitoring in rivers (Apothéloz-Perret-Gentil et al., 2017). The OTUs were assigned directly to ecological preferences after training of the dataset

with some reference environments with a large range of ecological conditions. This method has the advantage of using more than 95% of the reads instead of 36% with their reference library. This proportion of assigned taxa seems low, potentially as this study uses a very small reference library and the 18S V4 region which is easier to amplify but the assignment is not as effective as the diatom assignments using the *rbcL* barcode. A disadvantage of this approach is that it ignores data on diatom ecological preferences that have been collected during the last century. Moreover, this method does not have any way to control for contamination of the samples, especially because the 18S V4 marker is also present in plants, animals, and fungi. This barcode is frequently used in eukaryote Metabarcoding studies which do not target a specific taxon present in a sample (Pawlowski et al., 2016).

Therefore, within this thesis I decided to focus on comparing the BLAST and the Naïve Bayesian approach. I explore the different assignment methods in Chapter 3.

Although the taxonomic assignment methods rely on well curated and complete reference libraries, sequence quality and integrity need to be as high as possible. As the sequencing steps induces errors in the sequences, new bioinformatic algorithms have been created to correct the sequences before the taxonomic assignment.

NEW DENOISING ALGORITHM-BASED BIOINFORMATIC PIPELINE APPROACH

The sequence reads generated by HTS need to be processed to give the composition of the community present in the environment. Different bioinformatic pipelines exist. Due to the large size of the sequencer output (Gigabytes of data = hundreds of thousands of reads), those processes are likely to require a considerable amount of computational power. The earliest method used to reduce this requirement was to cluster together very similar sequence reads in Operational Taxonomic Units (OTUs), which is also useful to reduce the sequencing errors that could generate erroneous taxonomic units. There are several clustering methods such as Opticlust (Westcott and Schloss, 2017), furthest neighborhood or near neighborhood (Chen et al., 2013), but all of them unavoidably cause a loss of genetic variation information (Callahan et al., 2016).

Alternatively, a new method has been created based on denoising algorithms that oversees the correction of sequencing errors in order to obtain the genuine sequence variations with

a single nucleotide resolution, called amplicon sequence variants (ASVs). A recent paper (Nearing et al., 2018) described a benchmarking study of the principal algorithms used to denoise the raw sequences: Deblur, DADA2 and UNOISE3. It showed that very similar community compositions were obtained with each of the three methods. Only the alpha diversity, the number of different taxonomic units (OTUs when clustered, ASVs when corrected) in the community, seemed to diverge among the methods. This is not a problem for biomonitoring which focuses on the main dynamic of community change and is more tolerant to missing very rare taxa (unlike studies investigating the presence of invasive species species of conservation concern) (Lavoie et al., 2009). Although the results seemed quite similar for biomonitoring studies, the time required to run the different pipelines were significantly different, UNOISE3 (4.6 min) was ~1.3 times faster than DADA2 (5.8 min) and 15 times faster than Deblur (69 min). The OTU clustering method required a large amount of computer memory because all the sequences clustered in a single OTU need to be stored in the computer memory before this cluster is identified against the reference library. Because denoising methods use ASVs instead of OTUs clustering, they can analyse each read independently, resulting in linear scaling with sample number and possible parallelization (multithreading) which reduces the time and computational power requirements (Nearing et al., 2018). Hence the amount of time required is not problematic for processed MiSeq sequencing (but it is not impossible in the future that the datasets will become bigger). Moreover, the DADA2 R pipeline has the advantage of handling everything from input to graphical representations with a lot of flexibility, an important consideration when creating a straightforward method that aims to be routinely used by non-bioinformaticians for biomonitoring.

As I explored the limitation of the technology, the barcode choice is important because it raises its own limitations and specificities.

GENETIC REGIONS FOR DIATOMS BARCODING

While the choice of technology used for Metabarcoding survey is essential, the choice of the genetic region is also essential as it is directly linked to the DNA that will serve as input in the sequencing platform.

Several markers have been used for diatoms Metabarcoding including the mitochondrial genes : *18S*, *28S*, *5.8S*, *SSU*, *cox1*, *ITS* and chloroplast genes such as *rbcL* (Guo et al., 2015; Trobajo et al., 2010; Zimmermann et al., 2011). The *cox1*, *ITS* genes have been preferentially used during diatom taxonomic studies because of their high variability that allow differentiation even to subspecies level, which make them better for phylogenetic studies (Trobajo et al., 2010). For diatoms, the *18S* and *rbcL* are more conserved regions because they are coding regions (the integrity and the efficiency of the resulting protein should be preserved among throughout evolution), hence they are the preferred markers for biological monitoring studies because it they enable us to distinguish at lower the taxonomic ranks used during most biomonitoring, and permits an adequate level of genetic similarity among individuals of the same species or genus.

According to Kermarrec et al., 2013 the best region to use for diatom biomonitoring studies is the chloroplast gene *rbcL* (Figure 11). This marker was compared to SSU rDNA *18S* and *cox1* and obtained the highest detection rate of taxa present in a mock community (all taxa were detected) along with only one false positive (reads that are misidentified to species that are not present in the mock community). Hence this marker seems to be the best to discriminate diatoms at species level without producing false positive results (species incorrectly identified as being present in the sample when they are absent) or false negative (species present in the mock community but not detected). Following this study Rimet et al., 2018 created a method for diatom monitoring using Metabarcoding with an open reference library ready to use along with a set of five primers for *rbcL* (referred to as the diat.barcode mix of primers). This marker has the advantage of being specific to photosynthetic organisms, moreover the mix of primers amplifies a region of 312 bp that is very specific to diatoms. The variability of this short region is high enough to provide a species-level assignment for the majority of the sequences obtained during sequencing. This gene region is the one used commonly in the UK (Kelly et al., 2018). However, the primers used routinely in the UK amplify a slightly longer region and bind closer to the 5'end of the *rbcL* gene (Kelly et al., 2018) (Figure 10). The principal advantage is to avoid the use of a mix of primers, unlike the three different forward primers and two different reverse primers of diat.barcode.

Traditional ecological assessment based on diatoms relies predominantly on species determination because difference diatoms species within a genus could have very different

ecological preferences: e.g. the *Nitzschia* genus includes species that are very similar morphologically but with opposite environmental preferences, from oligotrophic to ultra-eutrophic environments. As the intention is to use diatom Metabarcoding to replicate traditional methods, it is therefore important to resolve all taxa down to species. Nevertheless, in Rimet and Bouchez, 2012, taxonomic assignment resolution to the genus level was found to be sufficient for most ecological assessments based on diatom biomonitoring.

Because using multiplexed markers (more than one barcoding gene on the same sample) increases the number of species detected (Zhang et al., 2018), the use of multiple markers (e.g. *18S* ,*SSU rDNA*, *rbcS*) during diatoms biomonitoring could be used as they have been used during diatoms taxonomy study (Trobajo et al., 2010). Nevertheless, we must consider three important points:

- This would complicate drastically the actual method (increased number of sequencing and PCR reactions, and primers) and also increase the cost.
- Several studies highlighted that ecological assessments based on genus rank identification are sufficiently accurate for biomonitoring (Rimet and Bouchez, 2012Lane, 2007).
- the *rbcL* marker is able to identify even down to subspecies level with a well curate taxonomic reference library (Rimet et al., 2019).
- The number of species detected would be higher, but the relative abundance of each taxon may differ drastically between primers and it would be difficult to correct the abundancy for ecological assessment index calculation.

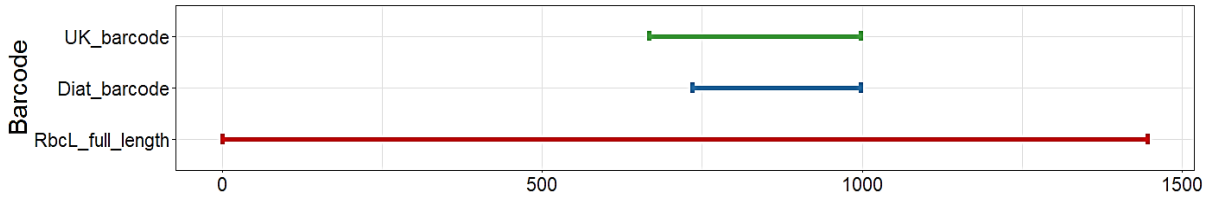


Figure 10 Localisation of the different amplicons used on the rbcL gene. Amplicon UK is from Kelly et al., 2018, Amplicon diat.barcode is from Rimet et al., 2019)

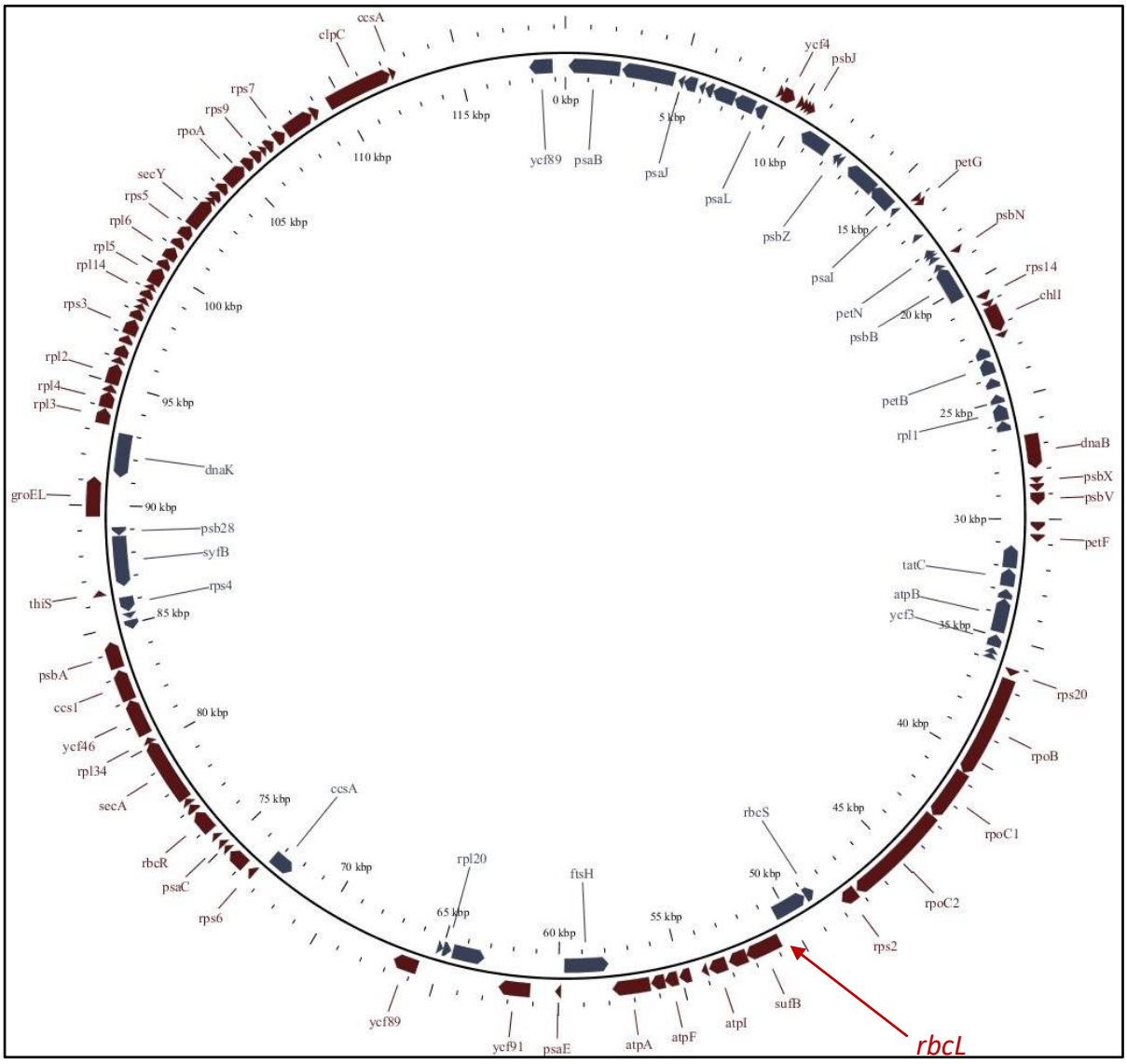


Figure 11 Position of the rbcL gene in the chloroplast genome of *Nitzschia palea* (Generated with Gview v1.7)

RELIABILITY OF DNA AND RNA METABARCODING COMPARED TO LIGHT MICROSCOPY

Despite the potential cost and time saved by a Metabarcoding-based approach compared to Light Microscopy, the reliability still needs to be fully evaluated and the results obtained by HTS need to be corrected (notably correlated to cell counts/abundance) before being used with the traditional environment indexes tools such as the trophic diatoms index (TDI) (Kelly and Whitton, 1995). The current way to assess reliability is to compare results to Light Microscopy results (although note the biological reasons these results may differ, as mentioned in sections 1.5 above), as is done in Chapter 3 and 4, or to use mock communities (as used in Chapter 5).

The association of read count with abundance is problematic; firstly, the traditional Light Microscopy method is based on count of single diatoms whereas the Metabarcoding method is based on reads of *rbcL* sequences, which are directly linked to the number of chloroplasts present in each cell. This is particularly problematic because this read number is correlated to different characteristics of the diatoms. For example, the biovolume of a diatom is positively correlated with the number of chloroplasts, subsequently a correction factor based on the biovolume of the diatoms was proposed (Vasselon et al., 2018). The corrected results are promising. Despite being a major driver of the number of chloroplasts, the biovolume alone does not seem to be the only driver, for instance centric diatoms are known to have tens of chloroplast whereas pennate diatoms commonly have less than ten chloroplasts (Bedoshvili et al., 2009). In the UK, the last version of DARLEQ, the official software used to calculate TDI (Juggins and Kelly, 2018) integrates a correction factor based on the average weighted proportion of each diatom. This is an empirical correction method but that seem to give good reliability with the Light Microscopy approach (Kelly et al., 2018).

In order to test the reliability of the biomonitoring using diatoms Metabarcoding, a few studies compare ecological Index values given for the community obtained with Light Microscopy (LM) and with Metabarcoding. This robustly evaluates the correlation between the results, and to create correction factors based on this knowledge (Visco et al., 2015; Zimmermann et al., 2015). Although providing promising results these studies are not generalizable because of the very small and particular area of study (Switzerland) and the

relatively low correlation factor between index values obtain in the samples. Nevertheless, the results are statistically significant and demonstrate that HTS read counts and LM counts are strongly correlated, but with large standard deviations.

In conclusion, the combination of diatom biomonitoring and HTS has the potential to provide a convenient and reliable method for water quality assessment. DNA Metabarcoding using the *rbcL* gene, rather than any other gene region, fits the ecological assessment requirements the best, but a direct comparison between the diat.barcode (Rimet et al., 2019) and the current (Kelly et al., 2018) *rbcL* barcode is needed.

Although being sufficiently accurate to answer particular ecological questions the Metabarcoding approach needs several improvements to be used routinely for diatom biomonitoring. In this context, this PhD Thesis aim is to improve the diatom biomonitoring method using Metabarcoding via:

1. Creating a DADA2 based bioinformatic pipeline and testing it against the traditional Light Microscopy method and the QIIME official bioinformatic pipeline.
2. Comparing the effectiveness of Illumina MiSeq to the Oxford Nanopore technology MinION for diatom Metabarcoding (mock communities).
3. Exploring the proportion of non-diatom to diatom reads generated and assign these with different *rbcL* barcodes and different reference libraries.
4. Exploring and quantifying the positive selection that could have been present during the evolution of the diatom *rbcL* gene history.

CHAPTER 2 GENERAL MATERIALS AND METHODS

This chapter aims to describe the methods that are shared between chapters in order to prevent redundancy. Detailed methods are presented in this section while chapters have briefer description of the methods and, if needed, details of the variations used in each chapter.

Some of the river sample data from this thesis originate from the UK and France (Chapters 3) this was provided as raw sequence data, but the methods used to produce this data are described below. Diatom and phytoplankton samples were also collected and processed for this thesis (Chapter 4) and followed the method described for the UK below. The methods differ slightly at most steps and are detailed separately below.

SAMPLE COLLECTION

UK

The biofilm samples from rivers were collected following the standard Environment Agency method (CEN, 2014; Kelly et al., 2018; M G Kelly et al., 1998) which involves the collection of 5 cobbles at each sampling sites. The cobbles were placed in a tray with about 50mL of water taken from the stream. The cobbles were brushed and the biofilm gathered/collected conserved in 70% ethanol solution after sampling. The samples from UK were preserved by mixing 5 mL of the suspension of biofilm and water with 5mL nucleic acid preservative made of 3.5 M ammonium sulphate, 17 mM sodium citrate and 13 mM ethylenediaminetetraacetic acid (EDTA). This was done for the samples collected in Chapters 3 and 4.

Alternatively, a pair of tiles was placed out at each sampling site of the river Foss. Biofilm was collected a month after the tiles were placed out following the standard Environment Agency method, with the tiles instead of cobles. This was done to standardise the method instead of collecting biofilm from different surfaces (in terms of size and roughness).

France

The biofilm sampling followed NFT 90 354 (Prygiel et al., 2002) which is very similar to the UK method: 5 to 10 cobbles were collected from the stream, biofilm was then brushed with water

from the stream. The samples collected were only kept in ethanol (70-80 % in the final volume).

Both samples from UK and France have been frozen at -30°C prior to DNA extraction and Light Microscopy slide preparation.

LIGHT MICROSCOPY IDENTIFICATIONS

The Light Microscopy inventories were executed in different laboratories but all followed the European Standards described in BSI (2003): Water quality — Guidance standard for the routine sampling and pre-treatment of benthic diatoms from rivers. (BSI, 2003).

The different steps included organic matter destruction with 30% hydroxide peroxide solution, Carbonate particles removal with 20 % HCl solution and mounting of diluted samples on slides with Naphrax (reflective index = 1.65).

The diatoms frustule identifications were done following the standard method using Light Microscopy with oil immersion and x400 to x1000 magnification. In order to determine the community compositions, at least 400 frustules were counted and identified in each French sample and 300 in each UK sample. Only diatom frustules with more than 75 % of integrity are considered.

The identifications were done with reference to "*Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment*" (Cantonati et al., 2017).

DNA EXTRACTION FROM BIOFILM

UK

The DNA extraction method follows Kelly et al., 2018 which includes spinning down samples (3,000g for 15min) to create a pellet, removal of buffer and then use of the DNeasy Blood and Tissue kit (Qiagen, Germany) following a standard protocol, with proteinase K incubation overnight .

France

The samples were extracted following the GenElute method (without the use of purification column, as described in Vasselon et al., 2017). This method involves a different sample lysis step, which was a thermal shock (-80°C/15min followed by 55°C/2min), sonification

(ultrasonic bath/20s) and enzymatic treatment (proteinase K); then the contaminant was removed by centrifugation and the supernatant collected. The DNA was then precipitated with GenElute TO-LPA (Sigma-Aldrich) followed by a step of centrifugation. The pellet was then resuspended in molecular water.

All DNA samples were preserved in cold storage (-30°C).

PCR AMPLIFICATION: SHORT BARCODES AND FULL LENGTH RBCL

UK

The PCR amplification followed the method from Kelly et al., 2018.

The volumes of amplification were 20µL and comprised of 0.3 µM of each primer (Forward primer *rbcL*-646F: ATGCGTTGGAGAGARCGTTTC, reverse primer *rbcL*-998R: GATCACCTTCTAATTTACCWACAACTG) including Illumina adapters, 4µl of HF buffer, 0.3 mM of dNTPs, 0.4 units Phusion high-fidelity DNA polymerase (New England Biolabs, UK), made up to a total volume of 19.5 µL using Nuclease-free water. 0.5µl of a 1:10 dilution of extracted sample DNA Nuclease-free water was used to make the final reaction volume.

The PCR reactions started with an initial denaturation at 98°C for 2 minutes followed by 35 cycles of: denaturation at 98°C for 20 seconds, annealing by lowering temperature to 55°C for 45 seconds, extension at 72°C for 60 seconds, a final extension at 72°C for 5 minutes. Electrophoresis on 1% agarose gels were used to assess the quantity and the length of the PCR product, dyed with ethidium bromide and visualised/printed on an ultraviolet (UV) transilluminator (Kelly et al., 2018).

France

The samples were amplified following Vasselon et al., 2019. The forward primer comprised an equimolar mix of Diat_*rbcL*_708F_1 (AGGTGAAGTAAAAGGTTCWTACTTAAA), Diat_*rbcL*_708F_2 (AGGTGAAGTTAAAGGTTCWTAYTTAAA) and Diat_*rbcL*_708F_3 (AGGTGAAACTAAAGGTTCWTACTTAAA); the reverse primer combined an equimolar mix of R3_1 (CCTTCTAATTTACCWACWACTG) and R3_2 (CCTTCTAATTTACCWACAACAG). For each DNA sample, PCR amplification was performed in triplicate in a final volume of 25 µL. Each PCR mix was composed of 1 µL of extracted DNA, 0.75 U of Takara LA Taq® polymerase, 2.5 µL of 10X Buffer, 1.25 µL of 10 µM of primers Diat_*rbcL*_708F_1_2_3 and R3_1_2, 1.25 µL of

10 g/L BSA, 2 µL of 2.5 mM dNTP, and made up to the final volume with molecular biology grade water. The PCR reaction conditions were initiated by a denaturation step at 95°C for 15 min followed by a total of 30 cycles of 95°c for 45s (denaturation), 55°C for 45s (annealing), and 72°C for 45s (final extension).

While overlapping, the two barcodes differ by their length with the diat.barcode amplicon shorter with is 263 bps (after primers trimming) compared to the 331 bps for the Kelly et al. (2018) barcode (Figure 9).The quantity and quality of each PCR product was evaluated by electrophoresis on 1.5% agarose gel.

SEQUENCING

The PCR products from French samples sites were sequenced at the "Plateforme Génome et Transcriptome de Toulouse" (GeT-Plage) and those from the UK were sequenced at Fera Science Ltd.

Both sets of PCR products followed the same process in the different laboratories. The PCR product was purified and the library preparation produced by adding specific tags and sequencing adaptors to each sample. The sequencing followed the paired-end multiplex method with Illumina MiSeq platform and V3 kit (2 x 250bp) for French samples and 2 x 300bp for UK samples.

BIOINFORMATIC PIPELINES

QIIME1 UK pipeline

The QIIME1 based pipeline is divided in a quality control step and a taxonomic assignment step, a complete description of the different steps is present in Kelly et al. (2018a).

The quality control part is divided into four steps:

- removal of the PCR primers located on both strands by the use of Cutadapt v1.9.1 (Martin, 2011);
- trimming of the poor quality 3' ends of sequences from both strands using Sickle v1.33 (Joshi and Fass, 2011);
- merging the paired-end reads using PEAR v0.9.6 (Zhang et al., 2013);
- removal of any sequences with a quality score lower than 30 and shorter than 250 bp using Sickle v1.33.

The taxonomic assignment was predominantly carried out by the QIIME platform (www.QIIME.org) and is divided into four steps following the in-house FERA bioinformatic pipeline:

- OTU *de novo* clustering at 97% similarity threshold using UCLUST (QIIME) (Edgar, 2010),
- selection of the most abundant sequence as representative sequence for each OTU (QIIME),
- taxonomic assignment of each representative sequence using BLASTn (QIIME) with 95% of sequence identity threshold, and
- calculation of the relative read abundances for each taxon presents in each single sample without filtering rarest OTU according to a minimum abundance threshold value.

DADA2 pipelines: QIIME2 and R

I created two different scripts of a new pipeline for more user flexibility, one written in R language and the other in QIIME2 language (Chapter 3).

Both QIIME2 and R script follow almost exactly the same steps, differing only in the graphical representations and in the quality filtering and trimming steps, which are run with other function, hence I consider them together hereafter.

The first steps of the pipeline are trimming steps and begin with PCR primer trimming of each read. A truncation step keeps the highest quality part of each read which are the 250 first nucleotides of the forward read and the 210 first nucleotides of the reverse read. The reverse reads commonly present a lower quality than the forward reads during Illumina dye sequencing due to a lower reagent concentration. Every read is afterward truncated at the first instance of an Illumina quality score less than or equal to 2.

A subsequent filtering step discards reads following these arguments:

- Presence of at least one ambiguous nucleotide (N),
- Reads with higher than 2 “expected errors” , which are calculated from the nominal definition of the quality score*: EE = sum(10^(-Q/10))* (Edgar and Flyvbjerg, 2015).

Thereafter the errors rate of every amplicon dataset is calculated by the use of a parametric error model. The method alternates estimation of the error rates and inference of sample composition until they reach a convergence with consistent solutions.

The dataset is then dereplicated prior to being denoised by using the core sample inference algorithm DADA2 (Callahan et al., 2016). This algorithm uses the error rates previously calculated to infer sample sequences (both forward and reverse) exactly and to distinguish sequences of as little as one nucleotide. This denoising process is an efficient alternative to an OTU clustering in order to deal with the common sequencing errors. The paired reads are then merged into contigs that are used to generate an amplicon sequence variant (ASV) table which has a similar structure to a traditional OTU table with the addition of a single nucleotide variant resolution.

Subsequently a chimera removal step is operated using the function provide in the DADA2 packages. The previous denoising step simplifies and shortens this process. The sequences are marked as chimeric (and so removed from the dataset) if there is an exact combination of a left-segment and a right segment from two different reads.

To assign taxonomy to each ASV I used the DADA2 implemented naive Bayes classifier method (Wang et al., 2007). The classifier is created by training it on a reference library made of diatoms and in some cases non-diatoms *rbcL* sequences. This method gives a rapid taxonomic placement with bootstrap value for each assignment. The output is a ready-to-analyse community inventory for each sample.

REFERENCE LIBRARIES

See Chapter 3 for a complete discussion and comparison of the different taxonomic reference libraries.

UK reference library (Kelly et al., 2018)

This reference library (Kelly et al. 2018) is based on 1483 sequences of 176 diatom species and available at https://github.com/rachelglover/diatom-analysis. Several sequences are from other phytoplankton taxa, but as the taxonomy of these algae is not specified, they appear as “NON-DIATOM” and “GREEN OR YELLOW ALGAE”. I used the latest corrected version which includes sequences file and taxonomy file to create a classifier useable by QIIME2 and another one useable with DADA2 (https://github.com/MathKarst/Diatom-izer).

Diat.barcode reference library

The well-curated and updated reference library provided in open access by the INRAE (France), diat.barcode, has been selected as an efficient alternative to the library used in the UK (Kelly et al., 2018). It contains sequences from the UK and also from all over the world which could improve the versatility of the assignment in term of geographical origins. The complete taxonomic lineage of each sequence is present.

RIVER QUALITY SCORE CALCULATION: THE TROPHIC DIATOM INDEX (TDI) VIA DARLEQ

In order to compare the results generated by each pipeline I compare the Trophic Diatom Index (TDI) values of the different communities obtained as well as the TDI EQR (Ecological Quality Ratio) which represent an ecological status class of either High, Good, Moderate, Poor or Bad (Juggins and Kelly, 2018; Kelly et al., 2014)

Both TDI and EQR is highly correlated to soluble P and nitrate-N which has been suggested to reflect the underlying inorganic nutrient pressure gradient (Kelly et al., 2008).

Environmental variables included are phosphate-P (P-PO4), nitrate-N (N-NO3), ammonium-N (N-NH4), alkalinity, conductivity and pH.

DARLEQ (version3) is used to calculate the TDI scores and EQR from each different river communities. I used the DARLEQ package with R software, including the interactive shiny app.

Calculation of the TDI derived from the weighted average equation of Zelinka and Marvan (1961):

$$Index = \sum_{j=1}^{n} \frac{aj.vj.ij}{aj.vj}$$

with *aj* = abundance (proportion) of species *j* in sample, *vj* = indicator value (nutrient preferences, 1-3) and *ij* = nutrient level sensitivity (1-5) of species *j*. The value of TDI can range from 0 (low nutrient concentrations) to 100 (high nutrient concentrations).

The expected reference values (eTDI) of each site was calculated using the annual mean alkalinity following this equation:

$$\mathrm{eTDi} = 9.933 * \mathrm{e}^{\mathrm{Log10(alkalinity)}*0.81}$$

The TDI EQR (Ecological Quality Ratio) is then calculated based on observed data (TDI) and, resulting in an overall EQR:

$$\text{EQR}_{DARES} = \frac{(100 - \text{observed value of river trophic diatom index})}{(100 - \text{reference value for river trophic diatom index})}$$

The last version of this software, Darleq3, enables the calculation of the TDI 5 NGS which implement a correction of the values from NGS by a heuristic method (correction factor for the taxa quantification which usually differs between molecular and Light Microscopy data). Other bioinformatic pipelines successfully used other corrections of the NGS data to reduce the overrepresentation or underrepresentation of some taxon comparing to the LM method. For example, a correction based of the biovolume of each diatoms has been used successfully (Vasselon et al., 2018)

I used the TDI5 version for the Light Microscopy data and the TDI4 and the NGS TDI5 version for Metabarcoding data.

The correlation between TDI4 and TDI5LM is 0.99 and Lin's concordance correlation coefficient is 0.99. (Kelly et al., 2020), as such they are very similar and I decided to only use the LM TDI5 version as it would not be informative to add the TDI4 in the studies. However, the TDI4 and the NGS TDI5 are not as similar.

Nevertheless, the NGS TDI5 version is the recalibrated version using a larger dataset in term of environment data and species sequences. Only a small overall improvement was found by the creator and optimisers of the index, which is the Environment Agency (Kelly et al., 2020). The NGS TDI5 versions have been designed to be more suitable for NGS data than the traditional TDI4 as it has been calibrated with NGS data instead of LM data.

CHAPTER 3 DIATOM-IZER: A DADA2-BASED BIOINFORMATIC PIPELINE DESIGNED FOR DIATOM BIOMONITORING USING METABARCODING

INTRODUCTION

Diatoms are ubiquitous freshwater microalgae known to be reliable water quality indicators due to their environmental preferences and fast response to ecological changes. This had led the analysis of diatom communities to be routinely used to assess the ecological quality of rivers in many countries, including the EU and the UK (Kelly, 1998; Prygiel and Coste, 2000). The traditional identification method relies on the diatoms observations from river biofilm using microscopy which is time-consuming and requires highly skilled taxonomic experts. Alternatively, a molecular based method, Metabarcoding, has been developed that relies on the identification of multiple species from a single environmental sample using variations in conservative short sequences of DNA/RNA (called a barcode) and High-Throughput Sequencing (HTS; also referred to as Next Generation Sequencing, NGS) (Taberlet et al., 2012). The diatom Metabarcoding method developed in the UK (Kelly et al., 2018) uses a short barcode located in the chloroplast gene *rbcL*, which encodes the large subunit of the RuBisCO enzyme. The diat.barcode 'European' method relies on a slightly smaller *rbcL* barcode which is located closer to the 3' end (Frederic Rimet et al., 2018).

There are known biases and errors associated with Metabarcoding, notably the formation of 'PCR chimera sequences' (a single DNA strand originated from more than one transcript) (Smyth et al., 2010), primer biases and sequencing errors associated with the different sequencing platforms (Nearing et al., 2018). These artefacts can impact the results if they are not considered and corrected for. However, it is not always simple to distinguish artefact sequences from real biological sequences. The substantial amount of data generated by sequencing needs to be bioinformatically processed. Several bioinformatic pipelines have been created to deal with these issues by using algorithms that can handle chimera removal such as Uchime (Edgar et al., 2011) or Vsearch (Rognes et al., 2016) and that can also deal with the sequencing errors by modelling and correcting amplicon errors to obtain amplicon sequence variants (ASVs) as an alternative to Operational Taxonomic Units (OTUs), which make similarity-related sequences clustering obsolete. This latest process, sometimes called denoising, is performed by the use of different sample inference algorithms, the most well-

known of which are Deblur (Amir et al., 2017), DADA2 (Callahan et al., 2016) and Unoise (Edgar, 2016).

The current bioinformatic pipeline created in the UK for diatom Metabarcoding is based on QIIME1 and relies on OTU clustering instead of ASV clustering-free approach. In order to update the method, I have created a bioinformatic pipeline that integrates a "denoising" step managed by DADA2. I have written two scripts, one in R and one for use with QIIME2 to provide greater flexibility of use and ease of integration. I compared this bioinformatic pipeline with the previously published UK QIIME-based bioinformatic pipeline in terms of speed of processing and accuracy of ecological assessment. I also compared these molecular approaches with the result given by the traditional Light Microscopy (LM) method which is the standard of comparison. The bioinformatic pipelines have been compared using two different datasets, one from the UK and one from France in order to evaluate whether the differences in the method have an influence on the outputs generated by the bioinformatic pipelines.

The aims of this study are to create a new pipeline and to compare the results given by my new bioinformatic pipeline with the bioinformatic methods created for UK biomonitoring in the UK (Bailet et al., 2020; Kelly et al., 2018). The results cannot be used to compare each setting separately but gives a general comparison of the pipeline outputs by which I propose and compare my own pipeline by this means.

MATERIALS & METHODS

DATASET ORIGINS

The study uses two diatom datasets, one from the UK and one from France. The UK dataset originated from 171 samples used during the 2016 analysis for the EU Water Framework Directive, collected following the standard Environment Agency method (CEN, 2014; Kelly et al., 2018; M G Kelly et al., 1998). The French dataset is composed of 371 samples from the public dataset from INRAE/AFB (Rivera et al., 2020), collected following the NFT 90 354 diatom collection protocol (Prygiel et al., 2002) which is very similar to the UK method. Both datasets are composed of the LM inventory, the output from an Illumina MiSeq sequencing

run and the alkalinity of the sampling sites (necessary for the calculation of environment indexes).

DNA EXTRACTION PCR AND SEQUENCING

The molecular methods used for the UK data set are given in Kelly et al., 2018 and detailed in the Materials and Methods chapter (Chapter 2) above.

For the French data, the samples were extracted following the GenElute method (without the use of purification column, as described in Vasselon et al., 2017).

Both set of PCR products followed the same process in the different laboratories. The PCR product was purified, and the library preparation produced by adding specific tags and sequencing adaptors to each sample. The sequencing followed the paired-end method with Illumina MiSeq platform and V3 kit (2 x 250bp) for French samples and 2 x 300bp for UK samples).

BIOINFORMATIC PIPELINES

Three different pipelines were used during this experiment and are all described in the General Materials and Methods Chapter: QIIME1 (Kelly et al., 2018), DADA2 R and DADA2 QIIME2. The DADA2 scripts are interchangeable and were created for this study.

REFERENCE LIBRARIES

Only the reference library from Kelly et al. 2018 was used on all samples. See the Materials and Methods Chapter 2 for more details.

RESULTS

TDI COMPARISON

Linear regression models have been built to calculate the correlation between the river quality scores given by each method. The hypothetical situation in which no single sensitive species was present in the site results in a 0 for both LM and molecular method, which means

the intercept passes through 0, therefore, the intercept was forced to 0 during the linear regression models building.

The R-squared and the equation of the linear models were the main comparable index and showed a very strong overall correlation between the result calculated from LM data and Metabarcoding. The French dataset gave stronger correlations due to the larger number of samples in the dataset.

UK	Linear regression		FR	Linear regression	
TDI NGS	y	R^2	TDI NGS	y	R^2
DADA2 / LM	0.941	0.93	DADA2 / LM	0.989	0.97
CLASS DADA2 / LM	1.01	0.95	CLASS DADA2 / LM	1.02	0.97
Q1 / LM	0.951	0.95	Q1 / LM	0.934	0.96
CLASS Q1 / LM	0.996	0.96	CLASS Q1 / LM	0.995	0.97
Q1 / DADA2	0.987	0.97	Q1 / DADA2	0.942	0.99
CLASS Q1 / DADA2	0.972	0.97	CLASS Q1 / DADA2	1.02	0.99

Table 3 Correlation factors (y) and R-squared values for the linear regression between TDI (Trophic Diatom Index) or TDI Class (CLASS) generated by each method. Method: Q1=QIIME1, LM = Light Microscopy, DADA2: Diatom-izer. Left-hand panel: UK dataset, right-hand panel: French dataset

UK Sites

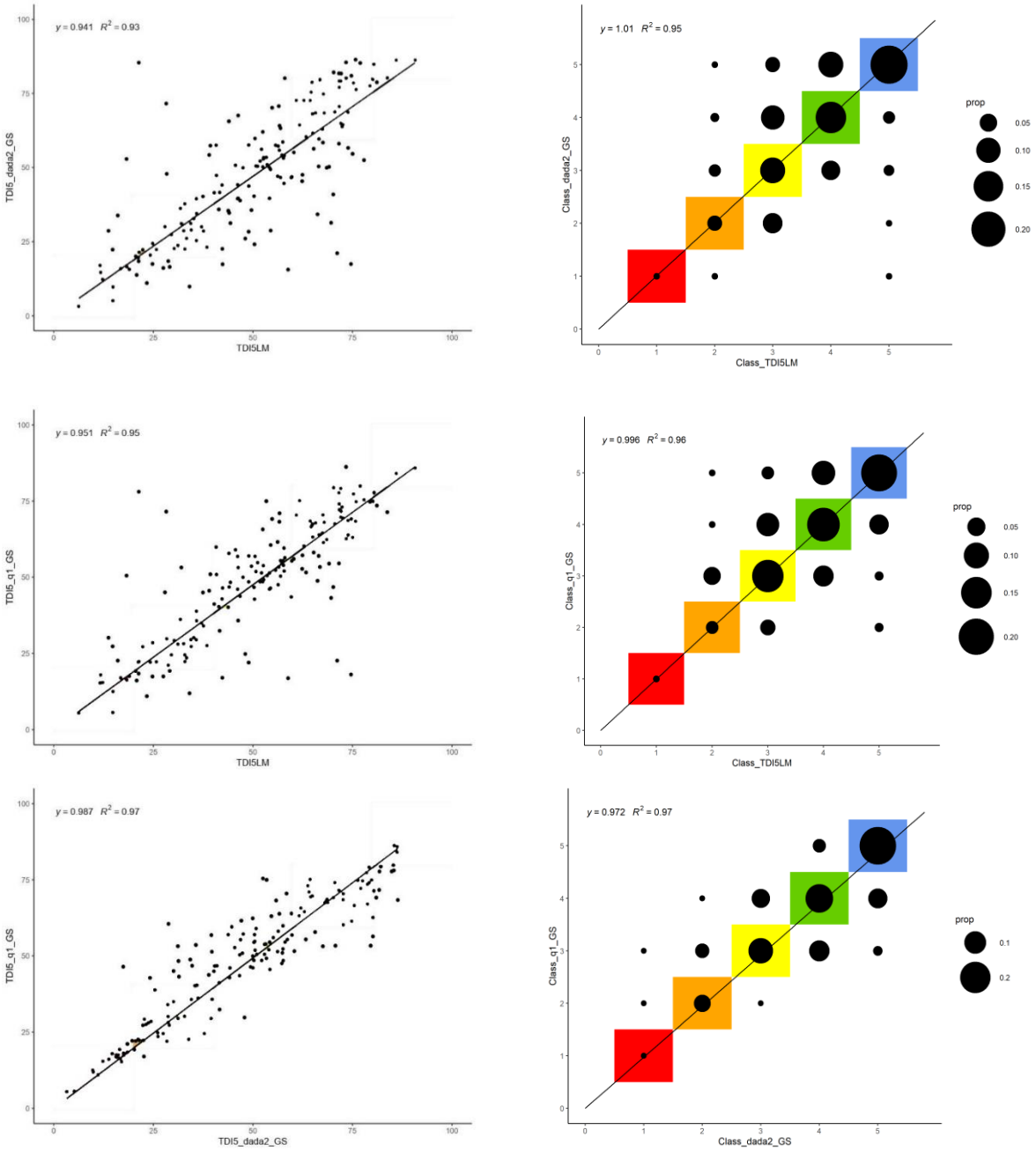


Figure 12. Linear regressions of the TDI values (left) and TDI classes (right) assigned by each method for the UK sites. The colours represent the different ecological classes: 1-BAD (red), 2-POOR (orange), 3-MODERATE (yellow), 4-GOOD (Green), 5-HIGH (blue). On the TDI class graphs (right) the presence of a point in a coloured region means that both methods assigned the site to the same TDI class. Note: the alkalinity is integrated in the calculation of the TDI ecological class. Important Note: TDI is negatively correlated with the water quality, as such the highest TDI, the lowest the quality and vice versa.

France sites

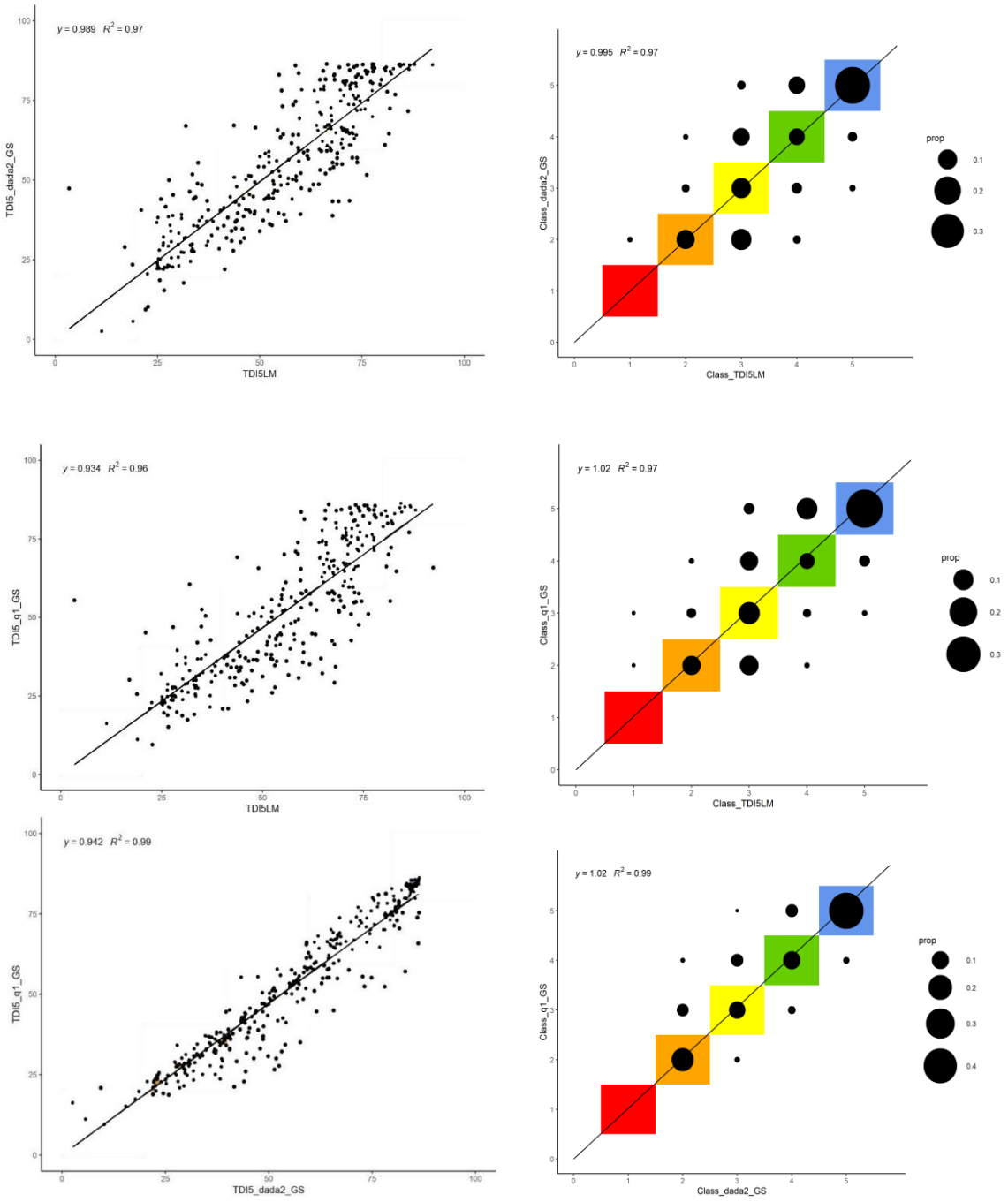
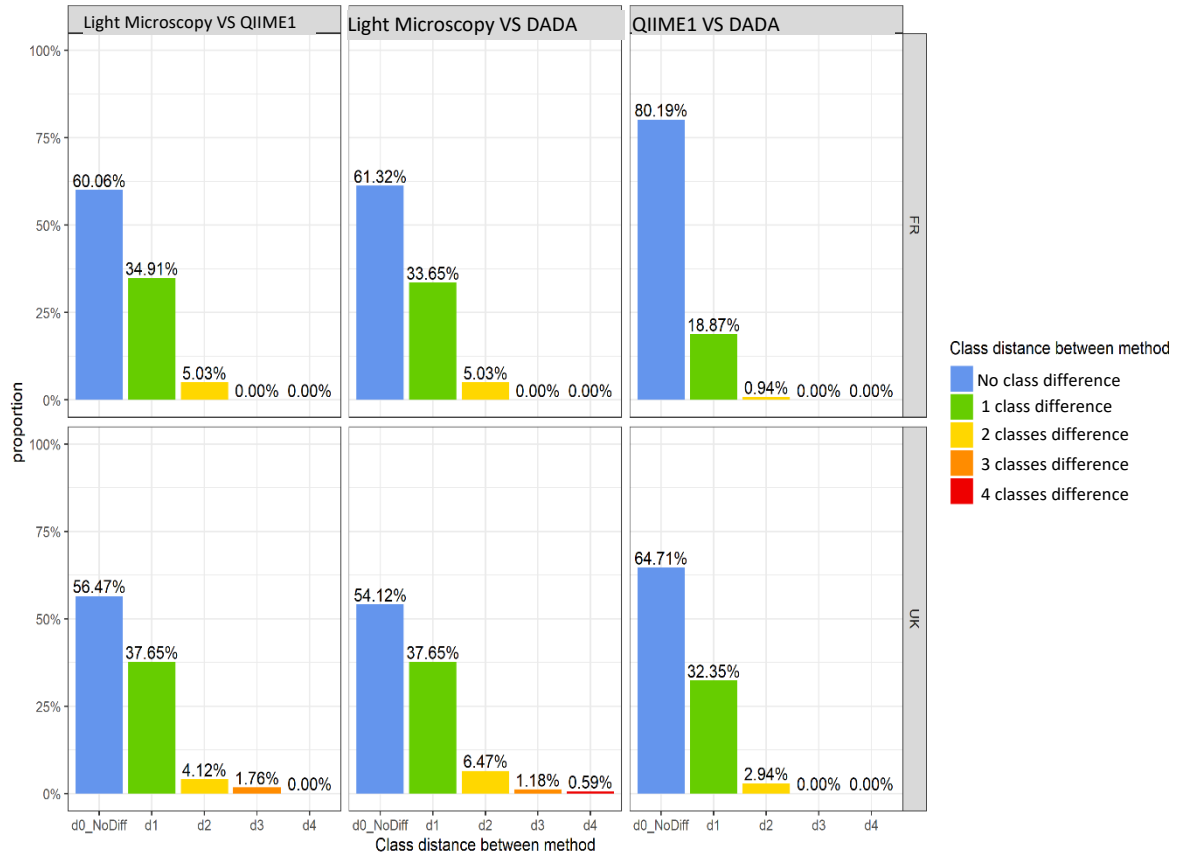


Figure 13 Linear regressions of the TDI values (left) and TDI classes (right) assigned by each method for the French sites. The colours represent the different ecological classes: BAD (red), POOR (orange), MODERATE (yellow), GOOD (Green), HIGH (blue). On the TDI class graphs (right) the presence of a point in a coloured region means that both methods assigned the site to the same TDI class. It is not exactly the case for the TDI plot because the alkalinity is integrated in the calculation of the TDI ecological class.



` Figure 14. Percentage of samples assigned to the same or different TDI classes by different methods.

The above graphs show the proportion of sites that were assigned to the same or different ecological (TDI) classes according to the method used.

DADA2 and QIIME1 assigned the same class in 80.19% of the French sites and 64.71% of the UK sites, with less than 3% of the sites assigned to a class distant from more than one class to the other.

When compared to the LM generated TDI class results, both pipelines gave similar results with QIIME1, assigning 60.06% of the French sites and 56.47% of the UK sites to the same class as the LM method. DADA2 assigned 61.32% of the French sites and 54.12% of the UK sites to the same class as the LM method. Less than 5.5% of the sites were assigned to a class distant from more than one ecological class to the one assigned by the LM method.

	LMvsQ1	LMvsDADA2	Q1vsDADA2
TDI UK	0.493	0.282	0.707
TDI FR	0.0241*	0.890	0.0258*

Table 4 Paired Wilcoxon test p value result. * Values are significantly different with alpha = 0.05. LM=Light Microscopy, Q1=QIIME1 pipeline, DADA2 = DADA2 pipeline

For the UK dataset, the paired Wilcoxon tests show that the raw TDI values were not significantly different between LM and each bioinformatic pipeline, furthermore, the TDI values from each bioinformatic pipeline were not significantly different.

The paired Wilcoxon tests on the French dataset gave different results: the TDI values from QIIME1 were significantly different from the TDI values of both LM and DADA2. Moreover, TDI values from DADA2 and LM were not significantly different.

COMMUNITY STRUCTURE: MANTEL TEST AND EVENNESS INDEX

Mantel test on dissimilarity matrices

In order to compare the correlation between the communities created by the Metabarcoding pipelines and the LM identifications, Bray Curtis dissimilarity matrixes were built for each method. The matrixes were used to run Mantel testes (Pearson product-moment correlation with 10,000 permutations).

The results showed a strong correlation between the community created from LM identifications and the ones created from the output of each bioinformatic pipelines. The results comparing the two bioinformatic pipelines (QIIME1 vs DADA2) presented the strongest correlation.

This confirmed that each method gave comparable and correlated community structures, which is promising in the perspective of Metabarcoding replacing or being used alongside the LM method.

	UK		FR	
	Mantel statistic	Significance	Mantel statistic	Significance
Q1 vs DADA2	0.8563	>0.0001	0.8436	>0.0001
Q1 vs LM	0.5144	>0.0001	0.6627	>0.0001
LM vs DADA2	0.3931	>0.0001	0.6239	>0.0001

Table 5 Mantel test result. The positive Mantel statistic values indicate a positive correlation beetween the matrixes of each method.

Evenness

The evenness index is one of the most used diversity indexes which has the advantage of including the abundance of each OTUs/ASVs in the community as well as the number of OTUs/ASVs/species (alpha diversity) in order to provide a metric that gives a clear and understandable evaluation of the dominances and uniformities among the studied community. This metric is frequently difficult to use with molecular data because the number of OTUs/species is easily overestimated with the errors during sequencing and PCR that creates singletons (or very rare sequences) during sequencing, and chimeras, during PCR, that are not representative of the true genetic diversity of the samples. Both pipelines include a rare sequences removal step that should remove the sequences created during sequencing, but only the DADA2 pipeline includes a PCR chimera removal step.

I used the evenness index (Pielou, 1966):

$$\mathrm{J} = \mathrm{H}/\log S$$

With J =Evenness index, H = Shannon diversity index, S= total number of OTUs/ASVs

I decided to work on the biggest dataset which is the one containing sampling sites from France. The comparisons between the evenness values generated by LM and every method shows evident correlations. The QIIME1 and LM values were correlated with a coefficient factor of 0.81, but they were significantly different using the Wilcoxon signed-rank test. The

DADA2 and LM values were strongly correlated with a coefficient factor of 0.96, the Wilcoxon signed-rank test indicated that the values were not significantly different.

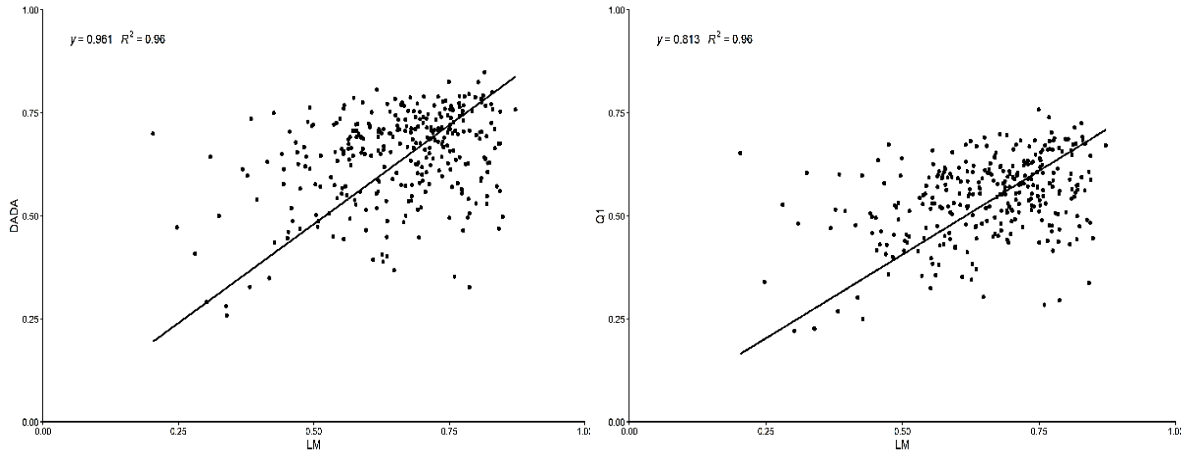


Figure 15 Correlation between the evenness values generated with metabarcoding (vertical axis) and LM (horizontal axis) on the dataset from France. Left: DADA2 pipeline; right: QIIME1 pipeline

The Wilcoxon paired test showed that QIIME1 TDI values were significantly different from both Light Microscopy and DADA2 TDI values while DADA2 TDI values were not significantly different from the LM TDI value. This is likely to be explained by the highest number of OTU given by QIIME because there was no chimera-removal step present. The DADA2 pipeline seemed to give a better image of the community structures.

Evenness values Wilcoxon paired test	UK	FR
	p-value	
Q1 vs DADA2	< 0.0001*	< 0.0001*
Q1 vs LM	< 0.0001*	< 0.0001*
LM vs DADA2	0.4828	0.279

Table 6 Wilcoxon-paired test with evenness values from different methods. As a Wilcoxon-paired tests, any value p-value greater than shows no significant difference between the two-paired dataset, lesser value shows significant differences between the two-paired dataset.

PROCESSING TIME

The processing time for each bioinformatic pipeline was measured for the 171 samples from UK. Only the identification steps were considered here. For the microscopy I only considered the time taken to identify the slides, and for the molecular approach I focussed only on the bioinformatic pipeline time requirement with the raw sequences from the sequencing platform. As such, laboratory work is not considered, namely the chemical treatments and slides preparation for microscopy, and DNA extraction and sequencing for the Metabarcoding approach. Both of these usually require a few days of work for both approaches.

The data was processed on the same server with 20 threads. It required 69 hours to process the 171 samples using QIIME1 (20 threads) from raw data output from the sequencing platform. It required 4 hours using DADA2_R pipeline (20 threads) from raw data output from the sequencing platform. This makes DADA2 more than 17.25 times quicker for this MiSeq output than for QIIME1, using this very typical diatom Metabarcoding dataset.

Method	**Time requirement**
Light Microscopy	4-6 weeks
QIIME1	69 hours
DADA2	4 hours

Table 7 Comparison of the time required to execute each of the three different methods: Light Microscopy, Metabarcoding with QIIME1 pipeline, Metabarcoding with DADA2 R pipeline.

DISCUSSION

The two different Metabarcoding bioinformatic pipelines gave very similar results for TDI ecological assessment. Compared to the LM method the ecological status evaluated were the same in ~58% of the sample sites, which can seem low, but the assessments differed by more than one ecological class in only ~5% of the samples. The French dataset result analyses showed that the TDI generated with the DADA2 method were not significantly different from the one generated with the LM inventories whereas the TDI generated from the QIIME1 method were significantly different from the LM outputs. This is a significant proof that

DADA2 ecological assessments were more similar with LM ecological assessment than QIIME 1 with LM.

The community structure, in terms of evenness index, given by the DNA Metabarcoding with DADA2 pipeline were not significantly different from LM-based community structure data. On the contrary, the QIIME1 pipeline community structures were significantly different from the ones generated with DADA2 or LM. The lack of chimera removal step induced an overestimation of species diversity attributable to the presence of singletons and rare sequences originating from sequencing errors rather than true genetic diversity. Consequently, the DADA2 pipeline performed better than the QIIME pipeline for diatoms community structure analyses if I considered the LM results as the standard.

In this study the two different Metabarcoding methods (for the dataset from France and from the UK) were close but not identical as the *rbcL* primers differed as well as the slightly difference during the PCR processes or the DNA extractions. These might have contributed to create different results between the two Metabarcoding methods, but this cannot be confirmed. Nevertheless, both methods provided quality output data that can be used as effectively as the LM identification for diatom biomonitoring studies and the gain in term of computational and time requirement was undoubtedly in favour of the Diatom-izer/DADA2 bioinformatic pipeline. The difference in term of evenness values was logically explained by the difference in term on bioinformatic pipeline and there is no reason to attribute this difference to the slight differences during the PCR or DNA extraction steps or the minor changes of the *rbcL* location and length.

The use of Metabarcoding for biomonitoring could be conjointly of the LM method. Potentially, DNA Metabarcoding is well suited to analyse numerous samples and LM can handle the most problematic samples (Kelly et al. 2018, Vasselon 2018). The complementarity of the two methods could potentially be efficient for the majority of river ecological assessments, thanks to the rapidity and cost effectiveness of Metabarcoding and the accuracy of LM identification.

Notwithstanding the improvement that could be done in the future, this is very promising for the future of diatom biomonitoring. Moreover, the DADA2 pipeline performed at least as well as the QIIME1 pipeline while improving some of the most problematic issues that used to be

constraining and demotivating the routine use of DNA Metabarcoding. These improvements are the convenience of use, the low computational and time requirement as well as the naïve Bayesian classification outputs that are useable straightaway for several analysis such as community structures or phylogenetic analysis.

DATA AVAILABILITY

Dataset from INRAE (France): https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/9EG5Z4

Diatom-izer repository (GitHub), including scripts and reference libraries: https://github.com/MathKarst

Kelly et al. 2018 reference library https://github.com/rachelglover/diatom-analysis

CHAPTER 4 COMPARING LIGHT MICROSCOPY AND MISEQ SEQUENCING FOR DIATOM METABARCODING IN BOTH CONTROLLED AND NATURAL FRESHWATER STREAMS.

INTRODUCTION

The increasing numbers of river ecological studies based on diatoms using Metabarcoding is fuelled by the latest studies that demonstrate that the Illumina HTS technology generates reliable data, suitable for water quality assessment (Bailet et al., 2020, 2019; Kelly et al., 2020, 2018). In the UK, the measurement standard is the Trophic Diatom Index (TDI; Kelly, 1998) which is calculated by taking the mean of the ecological preferences, and especially the nutrient optimum, of each diatom taxon, weighted by the relative abundance of each taxon. The TDI was designed for natural waterbodies such as rivers, which raises the question of the versatility of this index: will the TDI be more suitable for natural environments rather than artificial waterways such as canals or mesocosms? Fera Science Ltd has created the largest flow-through mesocosm in Europe, offering the perfect experimental system for testing the effectiveness of the TDI for trophic level assessment in diverse controlled environments. Moreover, mesocosms are powerful tools to experiment association between diatoms and water quality. They can help to determine whether some factors affect the Light Microscopy identification more than biomolecular method, as some characteristics should influence differentially the preservation of frustules compared to DNA, and vice versa.

In diatom community studies, the influence of upstream communities on the composition at a downstream site needs to be evaluated. The water quality assessment provides information about one precise location uncontaminated by cells from the upstream community that may be adapted to different conditions. As we are using DNA and frustules rather than RNA, there is a risk of “dead diatom” DNA or shells from the upstream community that could dilute our signal. In view of this, a study along rivers and using mesocosms is needed to reveal the influence of this upstream community on the downstream community and the effect of this on TDI calculation. The mesocosms offers us a very short “canal” that we can compare to two Yorkshire rivers: the Aire and the Foss.

Here we use both MiSeq Metabarcoding and Light Microscopy to characterise the benthic diatom community in sites from two Yorkshire rivers, and from artificial streams in the Fera

Science Ltd Mesocosm site. The main goals were to determine if the TDI is suitable for both natural and controlled environments, if the Inter-site and Intra-site differences will group sites preferentially by origins (from the same river) and/or by conditions (same nutrient level). Moreover, a tile-based sampling method was used and tested during this study to determine if this methodology is suitable for diatom biomonitoring. This method is inspired by Kelly et al. (1998). Finally, the comparison of Metabarcoding and LM identification method can give us information regarding the most suitable method for surveying or characterising the waterbodies studied.

MATERIALS AND METHODS

STUDY AREA

The chosen waterbodies for this study are both Yorkshire rivers that are tributaries of the River Ouse.

The mesocosm location is also in Yorkshire.

The relative proximity of the studies area (Figure 15 & 18) implies a similar climate.

Alkalinity and nutrient concentration used in this study was taken from the Environment Agency water quality data from the Water Quality Archive.

Figure 16 Location of the selected river starts (Aire= Yellow, Foss = Red) and the Mesocosm (Blue) in the UK map

River Aire

The River Aire is located in North and West Yorkshire and starts its course at Malham Tarn (altitude 377 meters) in the Yorkshire Dales, but it alternates between surface and underground flow until downstream of Malham Cove. It is 148 km in length and reaches Leeds 70 km from its source; this upper section of the river is the one of interest for this study. The total catchment area before Leeds is 690km^2 and is composed of a succession of Carboniferous geologies, limestones series between Malham and Skipton, Millstone grit from Skipton to Bradford and then coal measures (siltstone, mudstone, and sandstone) downstream to Leeds (Vercruysse et al., 2020). Alluvial deposits are noticeably present in the

near river area geology from the area upstream of Skipton to downstream of Keighley and from Bradford until it joins the River Ouse (Figure 16)(Carter et al., 2006).

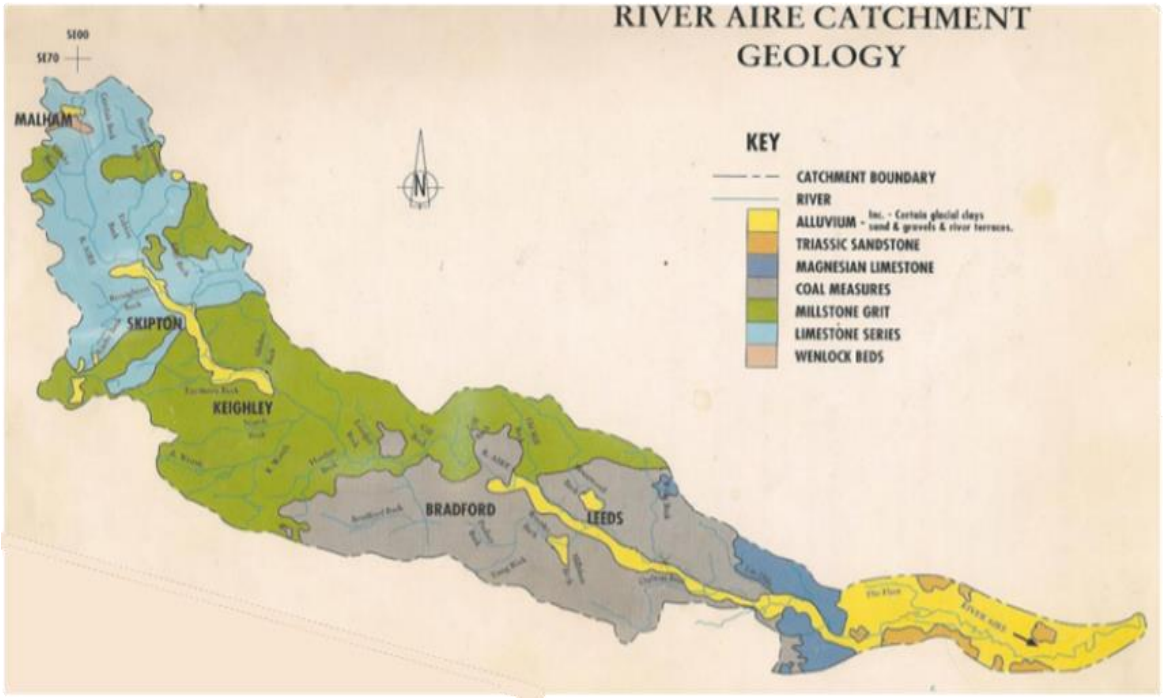


Figure 17 Geological map of the River Aire Catchment (from NRA, 1993)

Downstream from Keighley, the River has a history of heavy pollution as it flows through the former industrial landscape of West Yorkshire. Sewage still affects the ecology of the River Aire with 4,085 raw sewage events in 2021, nevertheless important investments have been made to modernise Wastewater Treatment Works, for example in Castleford and Esholt. These have improved the water quality and enabled the return of semi aquatic mammals such as Eurasian otters and European water voles.

River Foss

The River Foss runs through the Vale of York, from a small spring (altitude 160 m) flowing into Oulston Reservoir, to the River Ouse in the city centre of York, for a distance of 31 km (Fife and Walls, 1981). The geology is sandstone and alluvium (Figure 17). The history of regular flooding events drove the enaction of river modifications to prevent damage to cultivated land and urban areas around the river course. Additionally, wastewater treatment plants discharge into the stream. The upper section is mainly farmland and small villages, and the lower part is characterized by more urban area as it runs through York.

The River Foss geology is mainly sandstone which is noticeably more consistent than the series of different geological sequences which characterize the River Aire watershed.

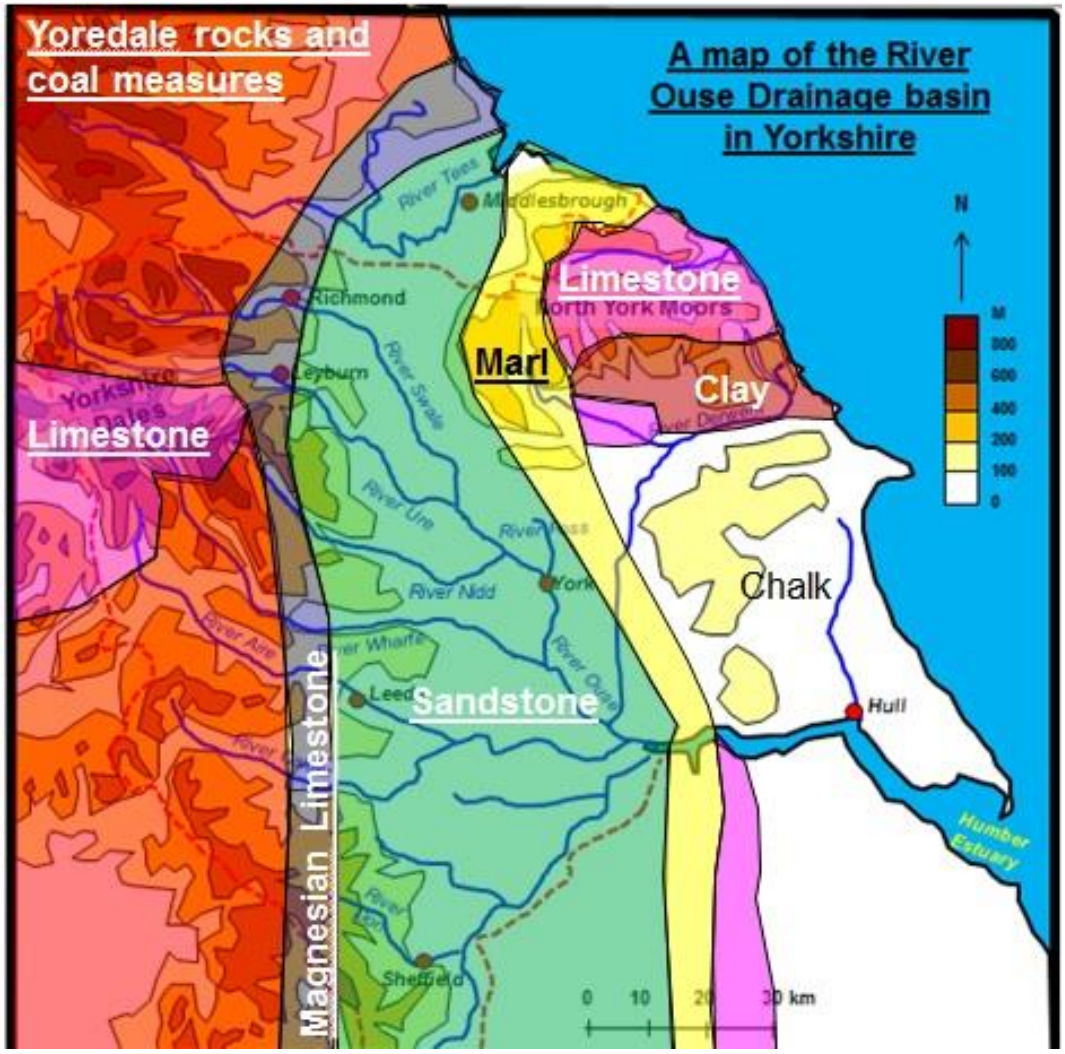


Figure 18 Ouse Basin Geology Map. From https://www.coolgeography.co.uk/

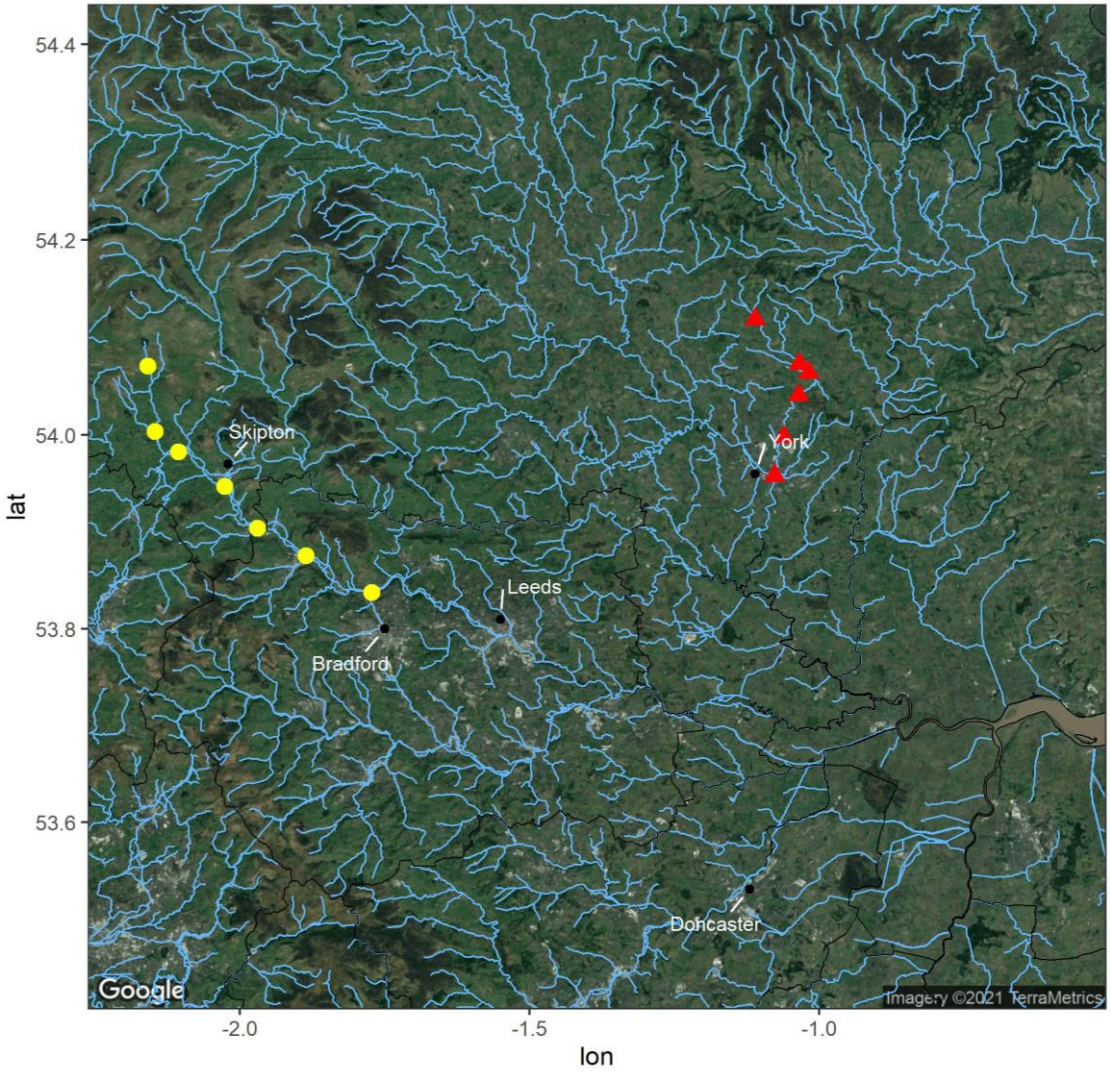


Figure 19 Location of sampling sites on the River Aire (Yellow dots) and River Foss (Red dots).

E-Flow Mesocosm

Based in Fera Science Ltd (York, UK), this mesocosm is the largest research platform for totally controlled water bodies in Europe, with 60 artificial streams and ditches. Each independent runnel is 10m long and they are all supplied with water from the same highly monitored source, which is a class A sandstone aquifer. Water is aged for a minimum of five days in controlled lagoons before entering the mesocosm (Figure 19 & 21).

In order to create diversity in the data I used

In 2019:

- two runnels with a single row of *Juncus effuses* plus soil ("M1" and "M2");

- two runnels with a single row of *Juncus effuses* minus soil (“M3” and “M4”).

In 2020:

- two runnels with “slow” flowing water (“24S”,”40S”,”48S”) 0.2 litres per minute;
- two runnels with “fast” flowing water(“29F”,”41F”,”43F”) five litres per minute.

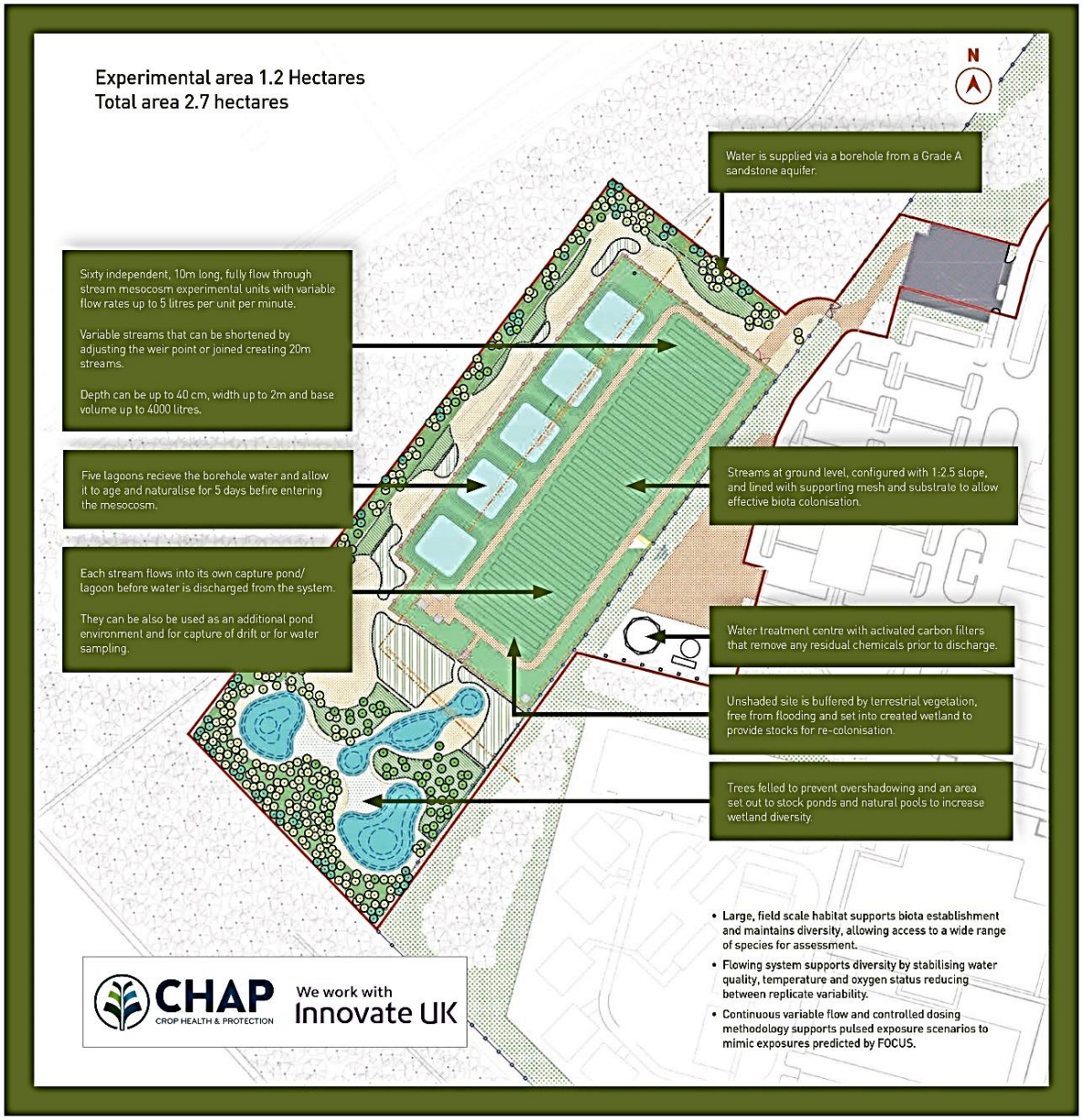


Figure 20 E-flow Fera Science Ltd Mesocosm experimental area detailed map.

Figure 21 E-flow Fera Science Ltd Mesocosm experimental area aerial photography

SAMPLING METHOD

As there are no cobbles or any other collectable mineral medium that I could brush to sample biofilms in the mesocosm, a standardised sampling method was designed and used in the mesocosm as well as in the River Foss in order to evaluate the versatility of the method. A pair of tiles (Terracotta, 15cmx15cm) were deposited at each sampling point: in the middle of the river channel (River Foss June 2020) or in the middle of the stream for the mesocosm. The tiles were set in place for a duration of at least one month (Summer 2019 and Summer 2020) in order to allow time for a biofilm to form on the surface of each tile. The biofilm was collected on site by brushing each tile with a toothbrush and ethanol. The biofilm preservation, DNA extraction and Illumina sequencing steps were carried out following the protocols described in the Materials & Methods Chapter.

The sampling method for the River Aire was a more traditional method (European Committee for Standardization, 2003) which involved just collecting 5 cobbles in each site before brushing biofilm from them (Cf Chapter 2 : Materials and Methods).

The list and details of each sample are present in the Table 1.

SAMPLE ID	ENVIRONMENT	SITE	FLOWING WATER	BIOFILM MEDIUM	POSITION IN CHANNEL	STREAM
24SB	Mesocosm	2020	Slow	Tiles	Bottom	24
24SM	Mesocosm	2020	Slow	Tiles	Middle	
24ST	Mesocosm	2020	Slow	Tiles	Top	
29FB	Mesocosm	2020	Fast	Tiles	Bottom	29
29FM	Mesocosm	2020	Fast	Tiles	Middle	
29FT	Mesocosm	2020	Fast	Tiles	Top	
40SB	Mesocosm	2020	Slow	Tiles	Bottom	40
40SM	Mesocosm	2020	Slow	Tiles	Middle	
40ST	Mesocosm	2020	Slow	Tiles	Top	
41FB	Mesocosm	2020	Fast	Tiles	Bottom	41
41FM	Mesocosm	2020	Fast	Tiles	Middle	
41FT	Mesocosm	2020	Fast	Tiles	Top	
43FB	Mesocosm	2020	Fast	Tiles	Bottom	43
43FM	Mesocosm	2020	Fast	Tiles	Middle	
43FT	Mesocosm	2020	Fast	Tiles	Top	
48SB	Mesocosm	2020	Medium	Tiles	Bottom	48
48ST	Mesocosm	2020	Medium	Tiles	Top	
LG	Mesocosm	2019	Medium	Tiles	Average	Lagoon
M1	Mesocosm	2019	Medium	Tiles	Average	Juncus effusus with soil
M2	Mesocosm	2019	Medium	Tiles	Average	
M3	Mesocosm	2019	Medium	Tiles	Average	Juncus effusus without soil
M4	Mesocosm	2019	Medium	Tiles	Average	
A1_AR6	River	2019	Natural	Cobbles	Average	Aire
A2_AR7	River	2019	Natural	Cobbles	Average	
A3_AR8	River	2019	Natural	Cobbles	Average	
A4_AR1	River	2019	Natural	Cobbles	Average	
A5_AR2	River	2019	Natural	Cobbles	Average	
A6_AR3	River	2019	Natural	Cobbles	Average	
A7_AR4	River	2019	Natural	Cobbles	Average	
A8_AR5	River	2019	Natural	Cobbles	Average	
R1_RF2	River	2020	Natural	Tiles	Average	Foss
R2_RF3	River	2020	Natural	Tiles	Average	
R3_RF6	River	2020	Natural	Tiles	Average	
R4_RF1	River	2020	Natural	Tiles	Average	
R5_RF4	River	2020	Natural	Tiles	Average	
R6_RF5	River	2020	Natural	Tiles	Average	

Table 8 Detailed description of the river and mesocosm samples

BIOINFORMATIC METHOD

The DADA2 pipeline was used (Callahan et al., 2016) with R (R Core Team, 2022) for the molecular data, for more details, see the Materials and Method Chapter and the Chapter 3 "DIATOM-IZER: a DADA2 based bioinformatic pipeline designed for diatom biomonitoring using Metabarcoding." In brief, it is composed of a filtering step (for both sequence length and quality), a trimming step, error rate estimation, sample inference using the estimated error rate, then merging of paired reads (forward and reverse), chimera removal, taxonomic assignment (diat.barcode custom, see results) and finally combines the OTU table and taxonomic assignment to create a diatom inventory table ready for ecological assessment indexes calculation.

As detailed in the Materials and Methods Chapter I used the Trophic diatom index (TDI) and more specifically TDI5 for the LM data and both the TDI4 and NGS TDI5 for the Metabarcoding data. LM TDI5 and NGS TDI5 are recalibrated versions of TDI4, nevertheless the LM TDI5 version generates extremely close results to the LM TDI4 (Kelly et al., 2018), so I decided to use the LM TDI4 as a "gold standard" for evaluation of recalibration of NGS TDI4 to TDI5.

Direct comparison using linear regression was used. The origin was forced to pass through 0 because an entire community composed of species without pollution sensitivity values would generate the same result of 0 for both NGS and LM method.

R software was used to run the two-sided Wilcoxon rank sum test to evaluated whether paired samples (from the same site but with different method) TDI results were significantly different and Kruskal-Wallis rank-sum test (alternative to ANOVA -Analysis of Variance- as the data do not follow a normal distribution) to compare the means between groups (LM vs Metabarcoding).

In order to find correlations between the species inventories generated by each method (microscopy and OTU table from MiSeq), Bray-Curtis dissimilarity matrices were calculated and a Mantel statistic test performed using R with the vegan package (Oksanen et al., 2022).

Correlation tests were performed in R using Pearson's product-moment Correlation test.

Non-metric multidimensional scaling (NMDS) plots were built in R to ordinate the distance matrix mentioned previously and condense multidimensional data. NMDS have a long history

of successful uses for ecological studies (Kenkel and Orloci, 1986) and more specifically with phytoplankton (Salmaso, 1996) and diatom data (Lane, 2007). The ellipses of each group are 95% confidence level for a multivariate t-distribution. This ease the visualisation of each group and see how each site is similar or different from the rest of its group.

Reference library

This study is an opportunity to compare the results from different taxonomic reference libraries. The descriptions of the different reference libraries are present in the Materials and Methods Chapter 2.

I created a new custom reference library by adding 19 non-diatom taxa from GenBank to the diat.barcode reference library (Rimet et al., 2019) : *Cryptomonas curvata, Chiloscyphus polyanthos, Vaucheria repens, Diplosphaera mucosa, Trebouxia sp, Chrysosporum ovalisporum, Pseudendoclonium akinetum, Heterococcus mainxii, Batrachospermum helminthosum, Heribaudiella fluviatilis, Oedocladium carolinianum, Interfilum paradoxum, Gonyostomum semen, Spirogyra fluviatilis, Ulvella repens, Planktothrix agardhii, Chlorella vulgaris, Chlamydomonas reinhardtii, Botryococcus braunii.* The new reference library is accessible GitHub: https://github.com/MathKarst/diat.barcode_custom/

Both the library used in Kelly et al. (2018) and diat.barcode custom generate very close taxonomical assignments at genus and species levels. The UK reference library suffers from having all of its non-diatom sequences details as “GREEN OR YELLOW ALGAE” or “NOT DIATOM” without lineage details, which is significantly less informative than the complete lineage of each assigned reads given by the diat.barcode reference library. Moreover, the naïve Bayesian taxonomic assignment method (Wang et al., 2007) is particularly affected by this lack of lineage as it uses the full lineage to localise the read in the taxonomic tree rather than just seeking characteristic read patterns (the Basic Local Alignment Search Tool : BLAST). This will theoretically result in less robust assignment as the taxonomy built from the reference library will agglomerate each “GREEN OR YELLOW ALGAE” together rather than trying to find similar pattern for particular phylogenetic groups. This motivated the inclusion of non-diatom reads (with a wide range of genetic distances from diatoms) from trusted sources in the reference library update.

It now has the potential to identify the origin of a read: from microalgae or from other photosynthetic organisms (from cyanobacteria to higher plants). Future surveys will probably consider other photosynthetic non-diatom taxa for biomonitoring and it is necessary to be able to discard DNA reads from photosynthetic organism that are not used for ecological assessment.

I trialled two databases; the diat.barcode and the current reference library used in the UK.

RESULTS

REFERENCE LIBRARY CHOICE

For this study the use of the diat.barcode reference library managed to identify 66 different genera whereas the current reference library identified 51 different genera.

Hence the decision to use the diat.barcode custom reference library rather than the current UK reference library as the diat.barcode maximises the information available which make it more adequate to my naive Bayesian based bioinformatic pipeline (DADA2/ Diatom-izer, see Chapter 3).

Therefore, for the later experiment I only used the diat.barcode custom reference library.

The relative abundancy plots, for each reference library, are present below in Figures 21 & 22.

DIATOM ASSEMBLAGE

An important part of the assemblage in the mesocosms was composed of the *Epithemia* genus for both Light Microscopy identification and MiSeq outputs. Moreover, the Light Microscopy identification showed an unusually high proportion of the taxon *Mastogloia* and *Rhopalodia*. They are all usually not dominant species and *Epithemia* and *Rhopalodia* are known to have endosymbiosis with cyanobacteria that are able to fix the atmospheric Hydrogen, and therefore these taxa are generally found in environment limited in Nitrogen. Moreover, *Epithelia*, *Rhopalodia* and *Mastogloia* are typical of hard spring-fed standing waters in northern England.

UK Kelly et al. 2018 reference library

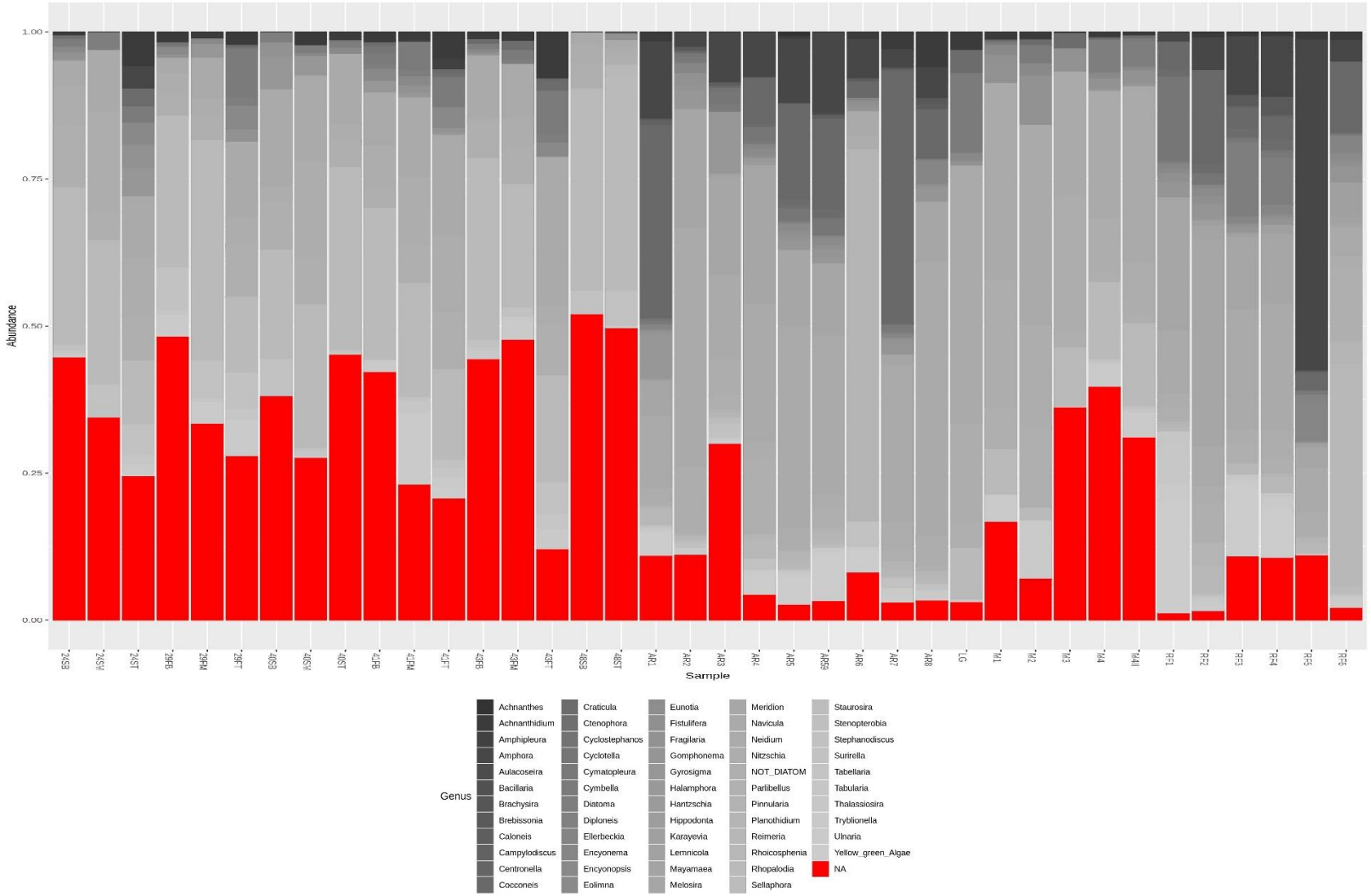


Figure 22 Taxonomy bar plot with the original reference library (without taxonomical lineage of non-diatom taxa) showing the abundancy of each genus identified in all the sites. Identified genera are displayed in a shade of grey and unassigned reads are in red.

Diat.barcode completed with non-diatom taxa reference library

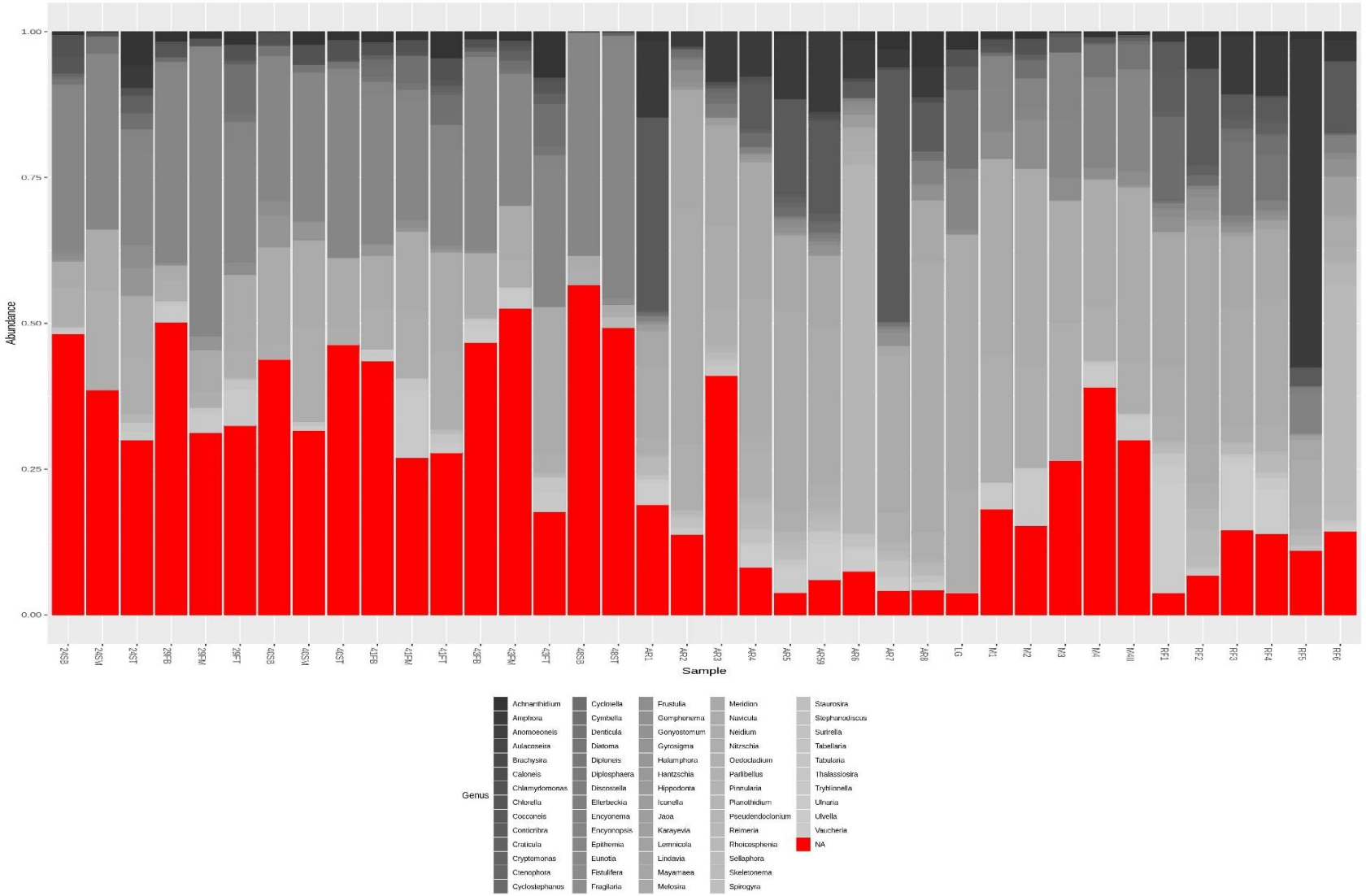


Figure 23 Taxonomy bar plot with the custom diat.barcode reference library showing the abundancy of each genus identified in all the sites. Identified genera are displayed in a shade of grey and unassigned reads are in red.

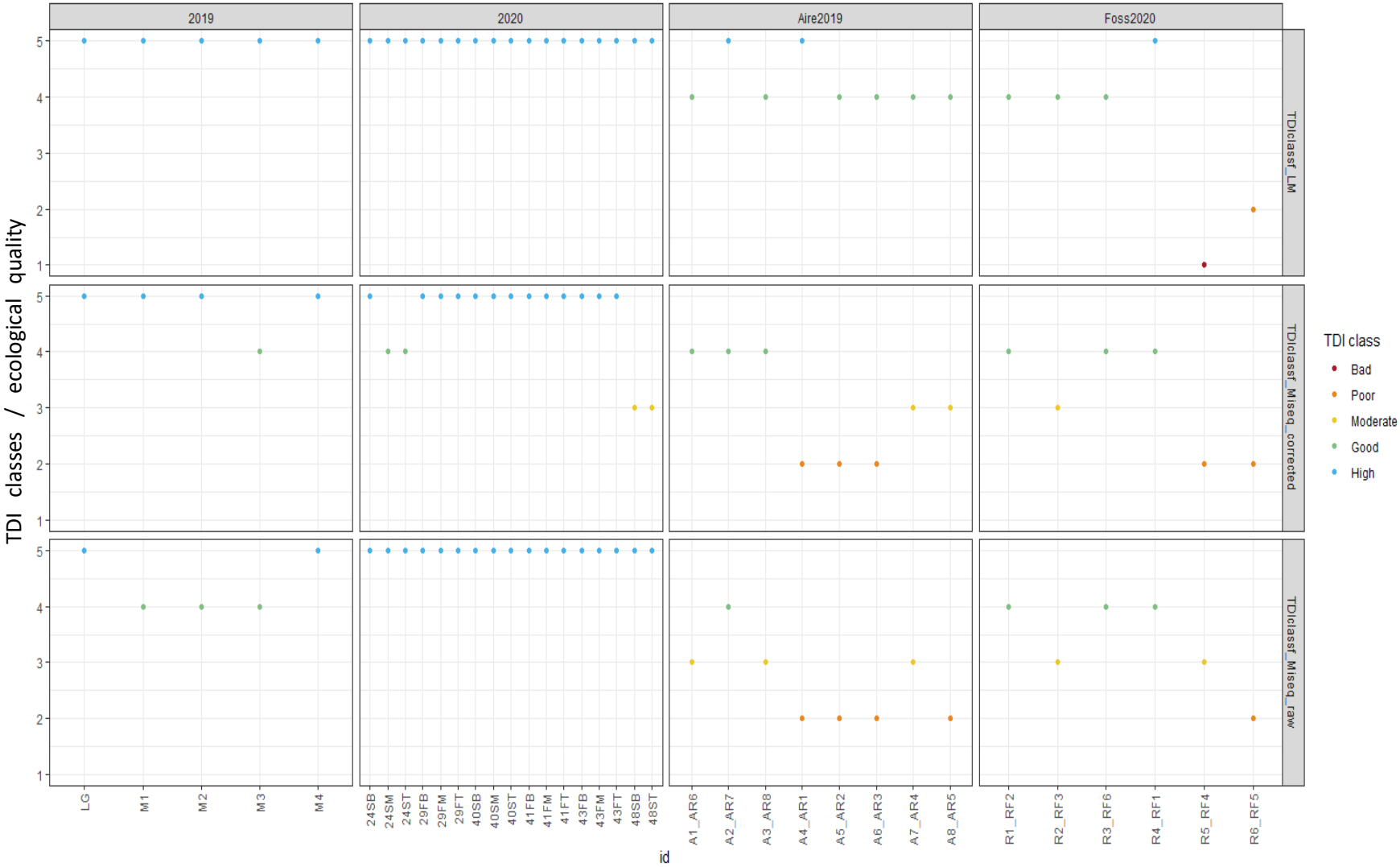


Figure 24 Comparison of the TDI ecological classes from the different sites (2019 and 2020 mesocosm, the Aire and the Foss) calculated with Light Microscopy TDI, MiSeq TDI with TDI5 correction and MiSeq raw TDI4 values. TDI ecological classes = ecological quality ratios: Bad (red), Poor (orange), Moderate (yellow), Good (green) and High (blue).

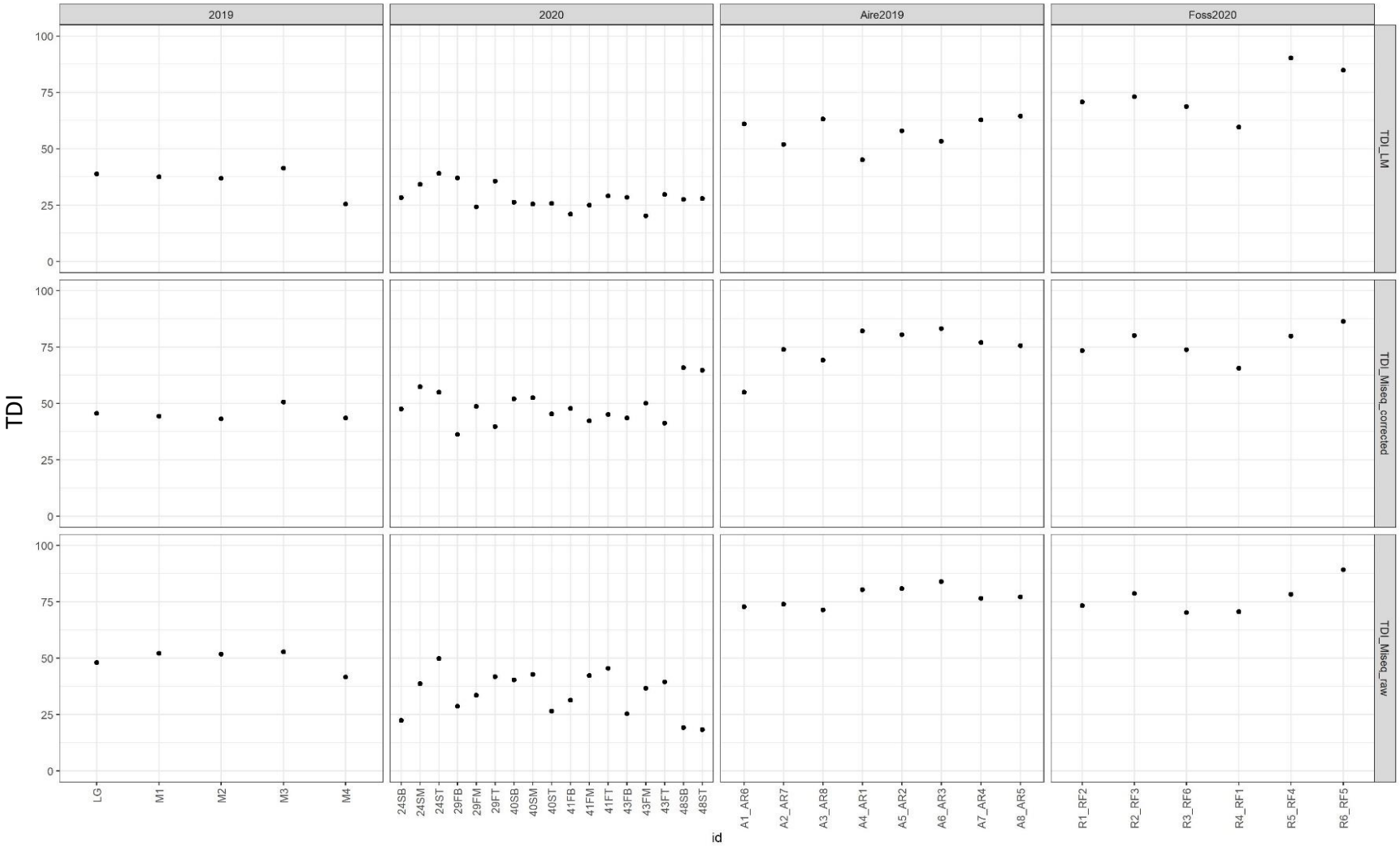


Figure 25 Comparison of the TDI values from the different sites (2019 and 2020 mesocosm, the Aire and the Foss), calculated with Light Microscopy data , MiSeq data with TDI5 correction and MiSeq data with raw TDI4 values.

MESOCOSM : VARIABILITY ALONG THE SAME RUNNEL : INTRA-VARIABILITY

Wilcoxon paired tests were performed in order to compare the TDI values of the sample from the top and the bottom section of each same runnel. There were no significant differences found during this test. Nevertheless, the Kruskal-Wallis rank sum test was not significant, and I cannot say that all group means were significantly different, which is not surprising considering the high number of groups (6 runnels) and the low number of duplicates per group (3).

EFFECT OF WATER FLOWING SPECIFICITY ON THE TDI RESULTS

The Wilcoxon ranked test was not able to find any significant differences between the TDI calculated for each group of flowing water, which tends to indicate that the water flowing rates of the runnels were not sufficiently impacting to create a significant effect on the TDI calculation. For better visualization, the average TDI value of each runnel can be seen in Figure 25.

M48S showed a noticeably high TDI value for the NGS TDI5 compared to the LM TDI4. Overall, the NGS TDI5 was higher than the LM TDI4 in five of the six runnels. Moreover, the LM TDI4 was always closer to the NGS TDI4 than the NGS TDI5.

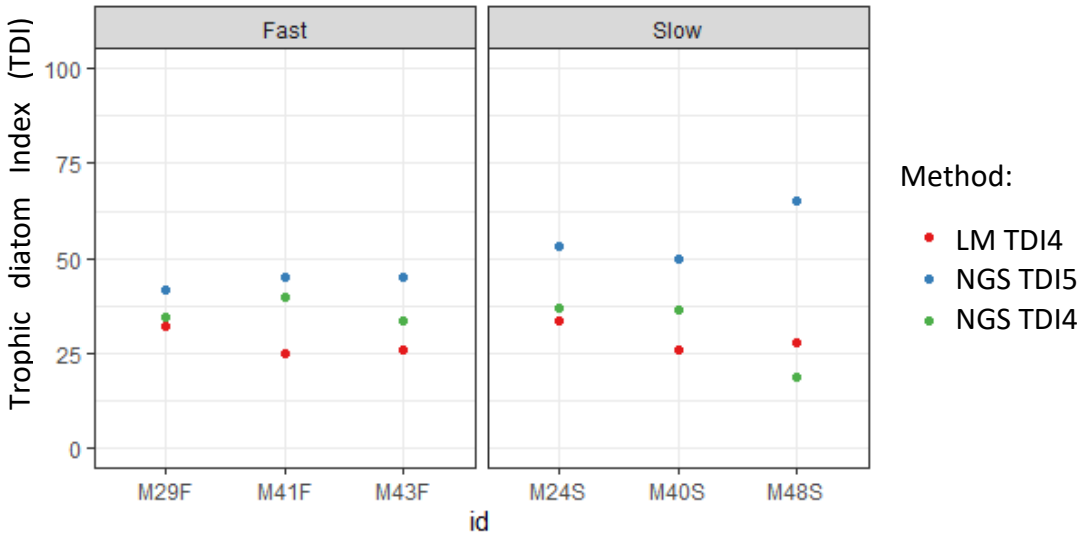


Figure 26 TDI value comparison between samples from the fast-flowing runnels (left grid) and the slow-flowing runnels (right grid). Light Microscopy method in Red, TDI4 Miseq method in Green and TDI5 Miseq method in blue.

NITROGEN AND ORTHOPHOSPHATE SNAPSHOT COMPARED TO TDI IN RIVER

The average 2021 orthophosphate (most readily bioavailable form) and Nitrogen concentrations were compared to the TDI value in the river sites (see Figure 26). The orthophosphate concentrations were more consistent than the Nitrogen values in both the River Aire (Orthophosphate mean = 1.23 $mg.L^{-1}$, variance = 0.223; Nitrogen mean = 2.80 $mg.L^{-1}$, variance = 9.39) and the River Foss (Orthophosphate mean = 3.51, variance = 0.00839; Nitrogen mean = 9.13, variance = 19.2).

There was a significant higher average nutrient level in the River Aire samples downstream Skipton (sample name: AR04) than in sites that are upstream. For both orthophosphate (2.4 $mg.L^{-1}$ > 0.022 $mg.L^{-1}$) and Nitrogen (4.4 $mg.L^{-1}$ > 0.87 $mg.L^{-1}$)

A significant correlation was not observed between TDI and nutrient concentration for either of the two small datasets.

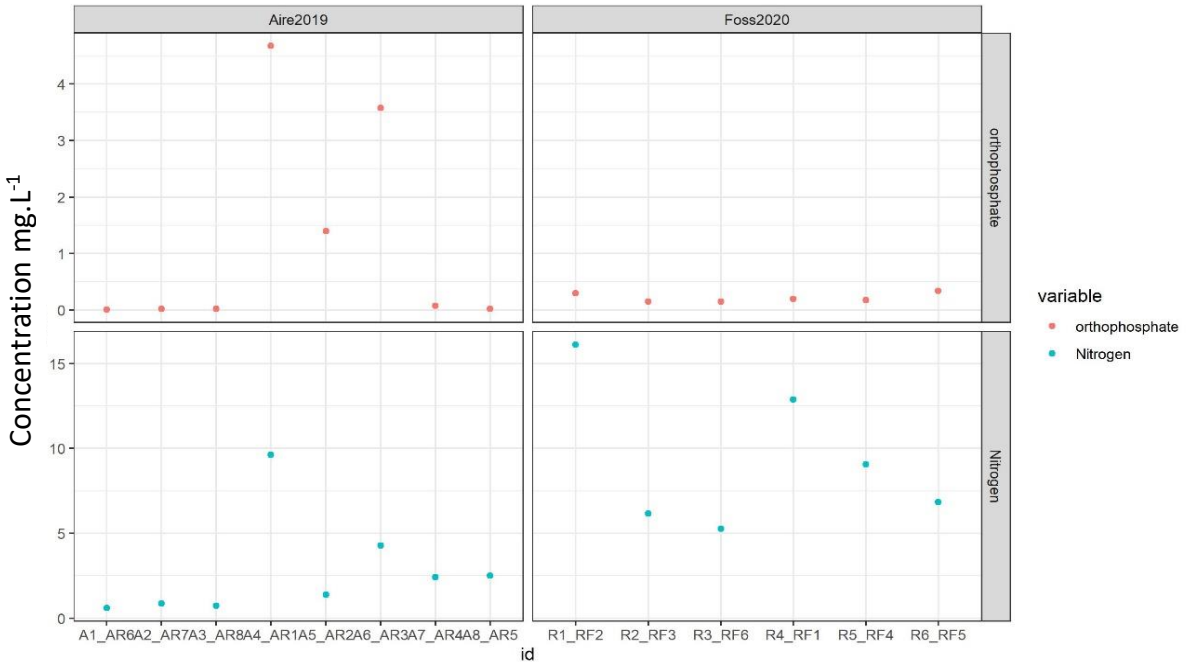


Figure 27 Nutrient level found in the River Aire (left) and the River Foss (right). Orthophosphate concentration: Top, red; Nitrogen: Bottom, Blue

EFFECT OF SOIL ADDITION IN THE RUNNELS

Significant differences of TDI were not found between samples from the lagoon, the runnels with soil or the runnels without soil (Figure 27). The two MiSeq methods were closer together than with the LM method. Moreover, the LM method tended to generate lower TDI values than both MiSeq methods.

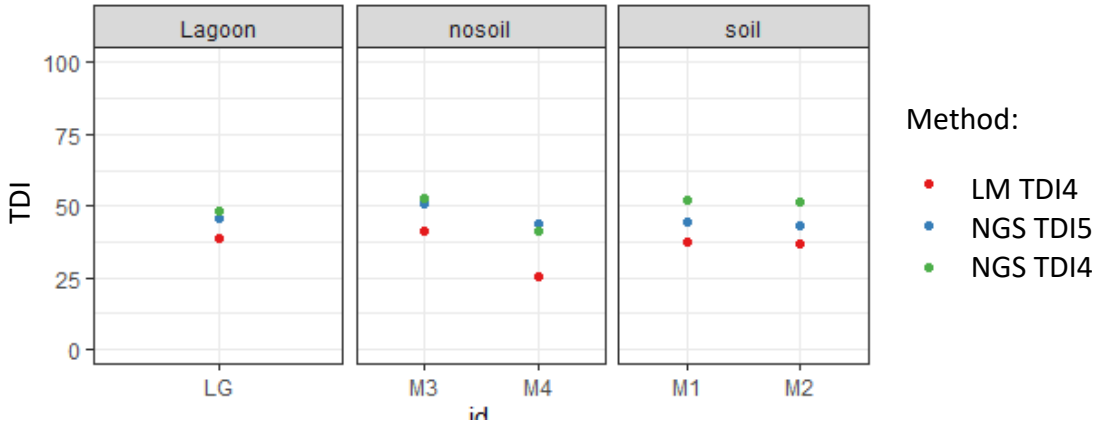


Figure 28 TDI value comparison between samples from the Lagoon (Lagoon) and the Runnels with soil (centre) and without soil (right). Light Microscopy method in Red, TDI4 Miseq method in Green and TDI5 Miseq method in blue.

MANTEL TEST

A high significance and a positive Mantel statistic r were observed between the species inventories generated by Light Microscopy and Metabarcoding (Table 9), which means that both matrices were positively correlated and therefore the community structures given by both methods were similar.

Mantel statistic based on Spearman's rank correlation rho			
Mantel statistic r		0.7851	
Significance		1e-04	
Upper quantiles of permutations (null model):			
90%	95%	97.5%	99%
0.0676	0.0951	0.1277	0.1674

Table 9 Mantel test results between Light Microscopy data and raw OTU table from HTS Illumina sequencing metabarcoding. Number of permutations: 9999

CORRELATION TEST TDI : NGS TDI4 OR NGS TDI5 FOR MISEQ DATA

Both NGS TDI5 and NSG TDI4 calculated from MiSeq dataset were analysed to determine their correlation with the LM TDI5 data (Table 10 and Figures 28 & 29). Surprisingly the TDI4 MiSeq was more strongly correlated to the LM TDI4 values than the NGS TDI5 with a lower correlation test p-value (6.55E-12 < 1.73E-08) and a higher correlation coefficient (0.869>0.782). A strong positive correlation is necessary if the MiSeq approach is to be used interchangeably with the original TDI from Light Microscopy.

	Sample estimated correlation factor	p-value	linear equation	Coefficient of determination R^2
TDI4 LM / TDI5 (corrected) MiSeq	0.782	1.73E-08	y=1.23x	0.93
TDI4 LM / TDI4 (raw) MiSeq	0.869	6.55E-12	y=1.17x	0.96

Table 10 Correlation test between Light Microscopy TDI result and each molecular TDI: TDI4 uncorrected, TDI5 corrected for molecular data.

Notwithstanding the strong correlation between NGS TDI5 MiSeq and LM TDI4, the result shows that NGS TDI4 was more strongly correlated to LM TDI4. Therefore, it seems that the recalibration of the TDI raw values from the MiSeq data is not beneficial with this particular dataset. Due to the limited size of the dataset, it is not possible to generalize this statement.

The linear equation with NGS TDI4 and LM TDI4 values was y=1.17x and a R^2 = 0.96 whereas that of NGS TDI5 and LM TDI4 is y=1.23x with a R^2 =0.93. It appears that NGS TDI4 values with LM TDI4 had a coefficient of proportionality closer to 1 than the corrected TDI with LM TDI4. The coefficient of determination (R^2) was higher when comparing raw TDI with LM TDI than corrected TDI with LM TDI. Even though the correction of TDI was created to make result from NGS and LM data more equivalent, it appears that in the context of my experiment the NGS TDI4 shows more similarities to the LM TDI5 outputs that the NS TDI5.

NMDS

Non-metric multidimensional scaling visualizes the differences and similarities of the overall community structure based on the communities identified with LM and from the raw (before taxonomic assignment) OTU tables.

The NDMS spatial analysis showed the clustering of each individual data group (Mesocosm 2019, Mesocosm 2020, River Aire 2019, or River Foss 2020; Figures 30 and 31).

Two main clusters were present, and the key driver was whether it was a controlled environment or a natural environment (a controlled mesocosm environment and two Yorkshire rivers). In fact, the Foss and the Aire clusters were overlapping, as were the two mesocosm clusters (2019 and 2020). Therefore, factors such as climate or localisation (The Aire and the Foss are much more distant from each other than the Foss and the Mesocosm) does not seem to drive the community structures compared to the difference between the artificial and natural habitats.

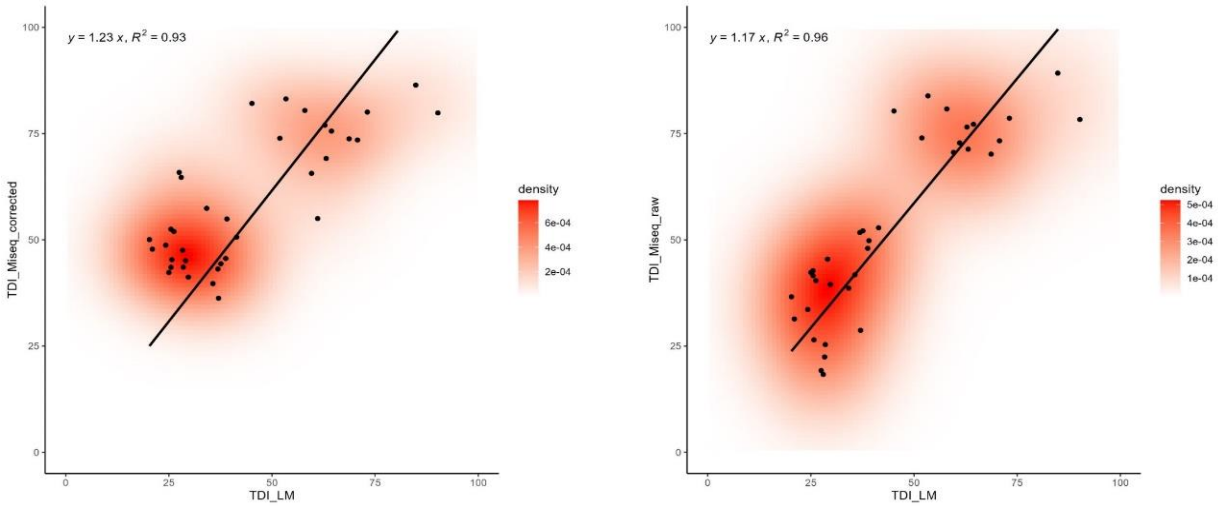


Figure 30 Correlation between TDI values calculated using the data provided by the MiSeq metabarcoding method or Light Microscopy identification. The metabarcoding date is corrected following TDI5 on the left and is raw TDI4 without correction on the right. TDI_LM is TDI4 version. Density is present as an indicator of the composition of the data.

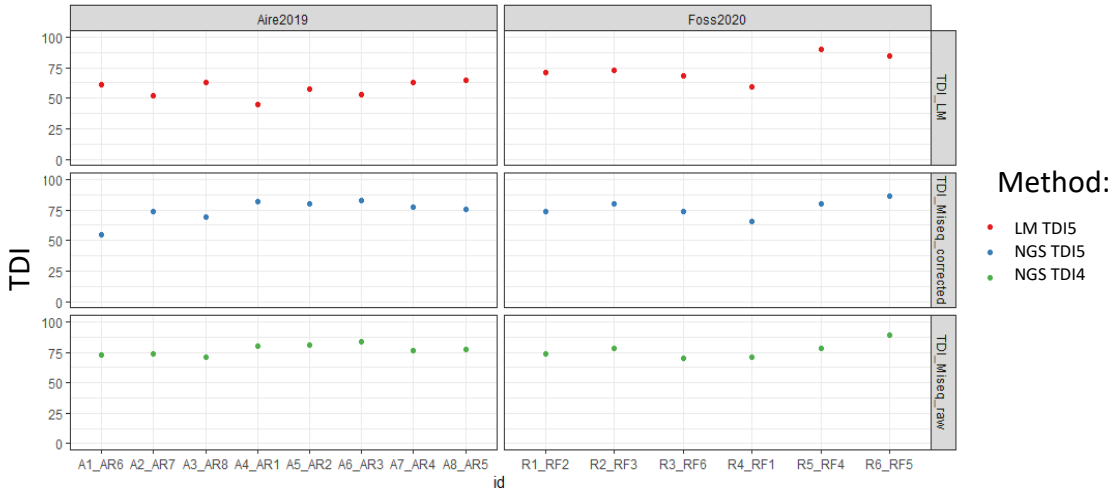


Figure 29 Comparison of the Trophic diatom index values in River Aire (left) and River Foss (right) regarding the TDI version : TDI4 with Light Microscopy data (top), TDI5 with Metabarcoding data (centre) or TDI4 with metabarcoding (bottom)

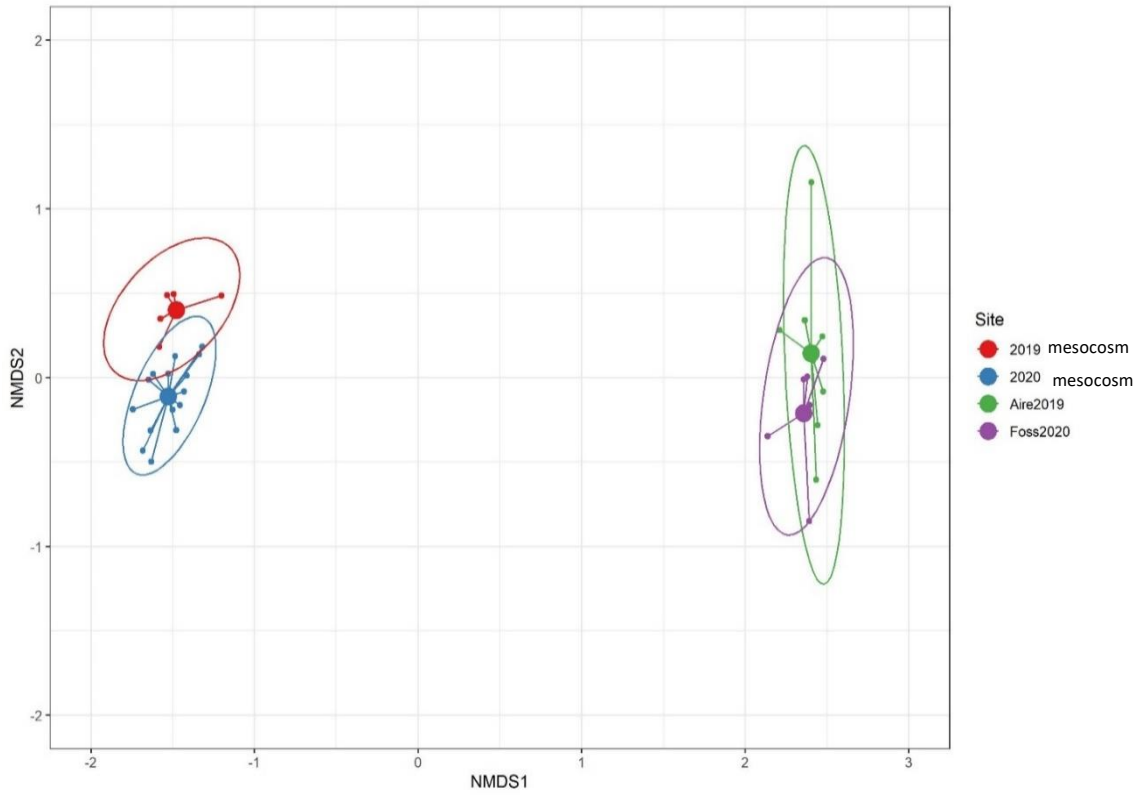


Figure 31 Non-metric multidimensional scaling (NMDS) analysis of the data from the metabarcoding method (OTU tables). Mesocosm sites in red (2019) and blue (2020), natural rivers in green (River Aire 2019) and purple (River Foss 2020). The ellipses are 95% confidence level for a multivariate t-distribution.

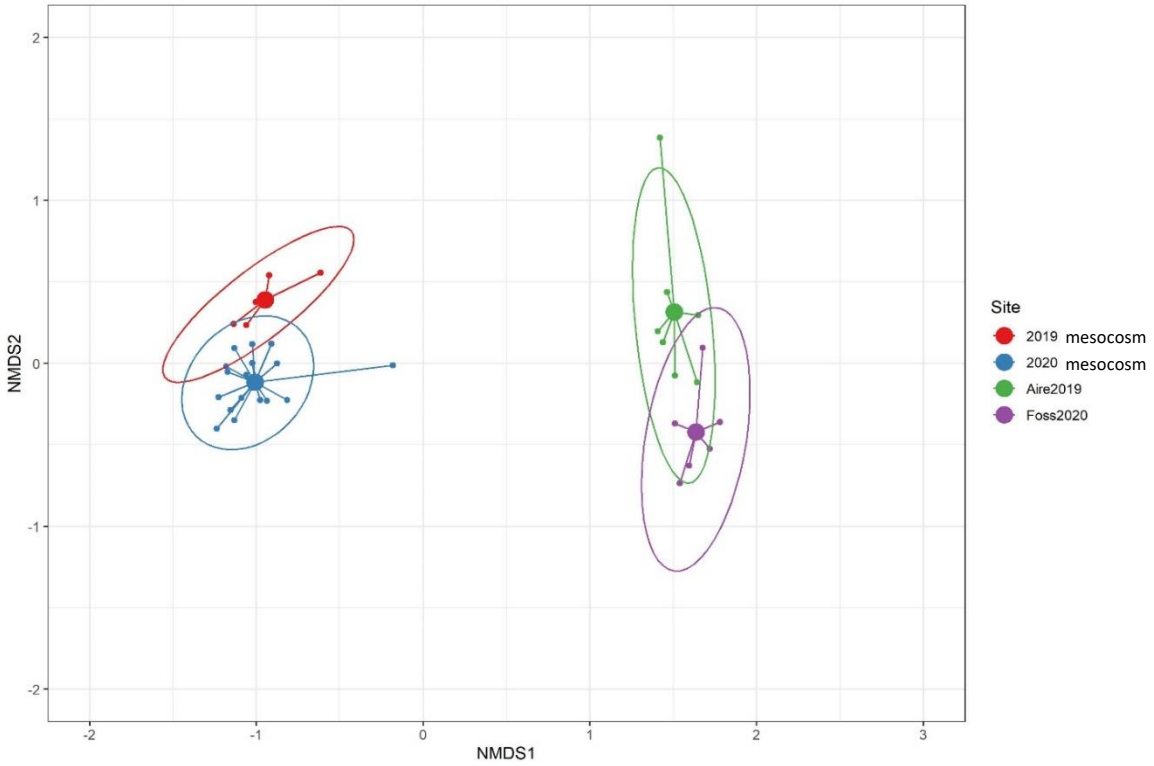


Figure 32 Non-metric multidimensional scaling (NMDS) analysis of the data from the Light Microscopy method (identification table). Mesocosm sites in red (2019) and blue (2020), natural rivers in green (River Aire 2019) and purple (River Foss 2020). The ellipses are 95% confidence level for a multivariate t-distribution.

DISCUSSION

SAMPLING METHODOLOGY

The overall adequate quantity and quality of data obtained from the tiles method is encouraging. This method showed controlled exposure time coupled with a standardised surface for biofilm sampling, two characteristics that reduce errors and, thus, are highly sought after by environment managers. Moreover, the presence in my study of both controlled and natural environments with distinctive characteristics, highlights the versatility of the method in a wide variety of situations and environments. All these points confirm the suitability of this method for diatom biomonitoring studies: we could imagine setting the tiles in diverse waterbodies and environmental managers could simply brush them to collect the biofilm when they need to assess the diatom assemblage of a site. This can solve some problems present when the site does not have biofilm covered surface that can be collected (including, but not limited to, mesocosm runnels), and the standardisation of the surface is beneficial as the nature of the surface can influence the diatom composition present in the biofilm. It has been already tested for experiment with, for example, ropes, tiles, sandstone and stainless steel artificial substratum (M. Kelly et al., 1998; Ramachandra, 2010; Richard et al., 2017; Rimet et al., 2009). But they are not used widely for biomonitoring in the UK.

INTRA AND INTER-VARIABILITY

The comparison of samples along the same runnel (top-bottom samples) seems to show that the intra-variability is low but does not prove that inter-variability between runnels is strong in my experiment which is not surprising as there is a high number of groups, low number of replicates and environment with subtle differences rather than completely different (same location, same water input, etc.). This is reassuring as the current sampling methodology (CEN, 2014) is focused on collecting samples representative of the site by standardized bulking several samples in the site to avoid the effect of minor river structure on the whole site assessment. Thus, the samples are mainly representative of the site and the bulking of different samples adds an extra layer of confidence in representativity of the site.

APPLICABILITY TO NATURAL AND ARTIFICIAL WATERBODIES

There is a wider range of communities within the natural environments because samples were taken along approximately 50km of river, featuring a great diversity of pressures (mainly nutrient inputs from agriculture and wastewater treatment plants), riverbanks, riverbeds and microclimates, whereas the mesocosms are all in the same field and share the same structure, climate and substratum, which leads to more similar communities because they are structurally and spatially very close. The assemblages found in the mesocosm include species which are typical of calcareous springs and ponds, which is a very different assemblage compared to the assemblages in the two Yorkshire rivers.

The slow and fast flowing experiment seems to show that the TDI is not significantly affected by this range of flowing water. This tends to confirm that the focus on periphytic diatoms is an adequate method to evaluate the quality of a waterbody without being overly influenced by the discharge, of the stream (Stevenson, 2014). The soil experiment tends to show that the presence of soil on a 10-meters-long riverbed does not strongly impact the TDI results, which could imply that the TDI is more correlated to the overall stream environment and water quality, than the particular characteristics of the precise location of the samples. This suggests that while keeping good practice during sampling (bulking five samples, finding the spots with the strongest waterflow) is still important to obtain good quality samples, the TDI calculation is robust to the precise location of the sample within a reach. Additionally, it corroborates the fact that diatom community analysis is a tool-of-choice for freshwater environmental assessment (Bailet et al., 2020; Kelly et al., 2016; Mora et al., 2019; Pawlowski et al., 2018).

The NMDS analysis of raw community data (Diatom inventories and OTU tables) shows a clear clustering related to the natural or artificial setting of the ecosystem. Neither the variation of flow velocity nor the addition of soil significantly affected the TDI values, it seems that these two factors are not the main drivers of the specificity of artificial ecosystems.

LIGHT MICROSCOPY AND METABARCODING

The main conclusions are that both molecular and Light Microscopy generates significantly similar results to be used interchangeably in most case.

Independently from the method used, the analysis of diatom communities along the two rivers highlights that the TDI results and the similarities amongst the communities from the same river are more related to the overall characteristics of the river rather than diatom communities from upstream. If it was not the case, we would have expected a significant difference between the results given by Light Microscopy (frustule from both living and dead diatoms) and by eDNA Metabarcoding as DNA should be less resistant in the environment. There is still a possibility than DNA trapped in dead diatoms and protected by the frustule would interfere with the eDNA Metabarcoding studies of downstream sites. An interesting test would be to try RNA Metabarcoding as RNA is more degradable in the environment and a majority should be from living organisms. Nevertheless, the RNA degradability is an important challenge and requires an advanced methodology especially for Metabarcoding, long-fragments and environmental samples (due to the presence of RNase, inhibitors, etc.) (Wood et al. 2019, Kagzy et al. 2022).

TDI VERSIONS COMPARISON

During this study the comparison of correlation suggests that it was not significantly better to use the recalibrated NGS TDI5 rather than the NGS TDI4.

This conclusions is limited to my experiment which is only focussed on two rivers in Yorkshire and runnels in a mesocosm, the recalibration of TDI was made among 1,367 paired LM and NGS samples (Kelly et al., 2020) which is a substantially greater dataset. Nevertheless, the LM TDI5 and LM TDI4 were proven to be extremely similar in the recalibration paper (99% correlation between TDI4 and TDI5 LM, 99% Lin's concordance correlation coefficient), so the recalibration is not significantly deleterious or advantageous in most cases. For the NGS TDI the recalibration still generates results with a tendency to slightly overestimate at low TDI values and underestimates high TDI values, and the conclusion of the recalibration paper (Kelly et al., 2020) indicates that while having a neater fit between NGS TDI and LM TDI after calibration, the overall improvement in correlation and concordance is small but give more linear fit than the original version. In this case the recalibration is slightly deleterious on the correlation between NGS TDI and LM TDI.

In the mesocosm samples few species have been found in rather unusually high proportions (dominant in several mesocosm assemblages): *Epithemia sorex*, *Mastogloia smithii* and

Rhopalodia gibba. While *Epithemia* was found without problem with the MiSeq method, *Mastogloia* and *Rhopalodia* were difficult to detect and in rather low relative abundance compared to the assemblage made with LM method. The LM method seems to be more efficient to find unusual taxa such as these three taxa. The primer affinity to *Mastogloia* and *Rhopalodia* has not previously been recorded as problematic in the literature, instead the accessibility of their DNA (number of copies per cell and frustule fragility) could be a probable reason of the low number of reads counts, but the low number of these sequences in the reference library is an important source of low identification reads. Nevertheless, this does not influence the TDI calculation to a problematic extent. *Epithemia* and *Rhopalodia* are found in poor Nitrogen environments such as the mesocosm and specifically in low N:P environments. They are known to be hosts of endosymbiosis with Nitrogen-fixing cyanobacteria. This uncommonly high abundance tends to indicate a low Nitrogen in the mesocosm, which corroborates the Nitrogen concentration measured in the mesocosm runnels. *Epithelia*, *Rhopalodia* and *Mastogloia* are also typical of spring-fed waterbodies, suggests that the water used in the mesocosm has kept groundwater characteristics even after the aging process in the lagoon.

The Diatom assemblage in the River Foss was frequently dominated by *Amphora pediculus* which is a pollution tolerant species, associated with human activity along the stream (confirmed by the sewage and agriculture activity described previously).

The upper River Aire sites were noticeably colonised by *Achnanthidium minutissimum*, a rather ubiquitous and mildly pollution sensitive species. *Planothidium lanceolatum* was dominant in the site downstream of Skipton. This taxon is tolerant to pollution and indicates pollution (Lange et al., 2011; Sbihi et al., 2014). This site is located downstream outputs of several sewages where very high orthophosphate concentration (4.7 $mg.L^{-1}$) and Nitrogen concentration (9.6 $mg.L^{-1}$) were found.

The higher average nutrient level downstream of Skipton reflected the high impact of wastewater on the nutrient concentrations in the river. This is known to be linked to eutrophication and should be monitored. There is a clear shift in the Metabarcoding TDI class (EQR) between samples from upstream and downstream of Skipton. This was not the case with TDI classes generated from LM. As such, Metabarcoding seems to be more efficient to

detect shift change in the nutrient level of series of site in a stream. It is difficult to explain the reason with my study but one of the main concerns of LM is the potential contamination by dead diatom shells from upstream that can influence the TDI calculation, as mentioned before in this section (cf. Discussion- Light Microscopy and Metabarcoding).

The value from the Water Quality Archive (Environment Agency) used to calculate the nutrient level average of the year of sampling was rather unstable and seems to have a high sensitivity to short-term events such as flooding and heavy rain because of the lixiviation of the soil, and irregular wastewater outlet (treatment plant or urban sewage; Eyre and Twigg, 1997; Fasching et al., 2019). For example, the most upstream site of the Foss River had Nitrogen concentrations which ranged from less that the detection limit to 29.1 $mg.L^{-1}$ during the same year. TDI was designed with the assumption that Phosphorus was generally the limiting nutrient (Kelly and Whitton, 1995). As TDI is derived from the diatom assemblage, which is affected mainly, but not only, by nutrient change, TDI is an indicator of the average recent nutrient level rather than the instant nutrient level. As such it is less sensitive to short-term events than the isolated nutrient values.

CONCLUSIONS

In conclusion my experiments were adequate to answer my questions: for the biomonitoring experiment, the Metabarcoding method gave similar results to traditional Light Microscopy method regardless of the artificiality of the water stream. The limitations of Metabarcoding induced by the sequencing accuracy and PCR were not deleterious to the water quality assessments. In fact, diatom bioindication was mainly driven by the proportion of the main diatom species/genera and the small errors during sequencing did not have a major effect on the final assessment. Moreover, the LM and molecular methods generated quite similar taxonomic inventories, although the nutrient increase downstream of Skipton was not associated with a TDI shift with the LM method, whereas the TDI shift was present with the Metabarcoding method. Considering this I could hypothesise that the LM method based on preserved frustule was more influenced by the surrounding sites than the DNA-based method. This corroborates my hypothesis that frustules have a greater tendency to influence downstream TDI calculation than nucleic acids do, as frustules are more preserved than DNA and can originate from upstream (Whitton et al., 2009).

A larger scale experiment would be able to answer other questions and to have more robust conclusions. For example, increasing the number of runnels could increase the number of factors tested as well as the number of replicates. Moreover, experimenting on a greater number of rivers has the potential to generate a better understanding of the effect of the variability of condition in these natural waterbodies on TDI. This kind of experiment was done in the UK and Ireland in 2018 (Kelly et al., 2018) in order to calibrate the TDI5, and as such it is calibrated for samples from the UK and Ireland rather than more diverse rivers. The diat.barcode reference library (Rimet et al., 2019), created and curated by an European wide study group, sometimes use the TDI but different standards exist: the IPS (Specific Polluosensitivity Index; (CEMAGREF, 1982; Descy and Coste, 1991; Prygiel et al., 2002) and, in a less systematic way, the IBD (Biological Diatom Index;(Lenoir and Coste, 1996; Prygiel and Coste, 2000). In order to compare samples from either inside and outside of the UK, it would be interesting to calibrate the TDI with samples from outside the British Isles or to use a more widely used index. The last TDI recalibration is to be used carefully as it does not improve the correlation between LM TDI and NGS TDI in this experiment.

My tiles-based methodology revealed a great potential for diatom sampling standardization which could ease the diatom biofilm sampling step in already monitored river sites, but also in deep rivers where cobbles are not accessible. In fact, the tile could be easily attached to a wire and enable benthic diatom biofilm to be sampled. The main downside of this method is that it requires tiles to be positioned beforehand, which is not problematic in the case of monitored sites but prevents the analysis of newly discovered sites of interest. Due to the similarities between the tile-based method and the traditional cobble-based method, it seems that using tiles on the monitored sites and cobbles as option for other sites is an interesting compromise.

It appears that the characteristics of the diatom community are heavily impacted by the degree of artificiality of a water stream. Whilst mesocosms seems to be excellent for experiments, there is still room for improvement for them to mimic natural environment. For now, they are more similar to canals and other artificial waterways, which is exactly what they are.

Finally, the mesocosm is an effective way of testing the interaction of conditions on diatom communities. Overall, Metabarcoding generates very similar results to Light Microscopy for environment quality assessment but they should be used conjointly as Metabarcoding enables the simultaneous analysis of a large number of samples but Light Microscopy, as a standard, should be the tool of choice for the potential ambiguous samples with particular assemblage and sometimes morphological anomalies (such as pollution induced Teratological forms; (Falasco et al., 2021) that cannot be detected by a molecular approach.

CHAPTER 5 COMPARISON OF MICROALGAL MOCK COMMUNITY STRUCTURES GENERATED BY DIFFERENT METABARCODING PLATFORMS (MISEQ VS MINION)

INTRODUCTION

Diatom biomonitoring is a powerful tool for river quality assessment (Bailet et al., 2020; Kelly et al., 2008; Prygiel et al., 2002).The measurement standard for this kind of survey is morphological identification of diatoms via traditional Light Microscopy (LM). The identification and individual cell counts are coupled to calculate ecological indexes such as the Trophic Diatom Index, the “Indice Biologique Diatomées” (IBD, Biological Diatom Index) (Lenoir and Coste, 1996; Prygiel and Coste, 2000), or the “Indice de Polluosensibilité Spécifique” (IPS, Specific Pollution Sensitivity Index) (CEMAGREF, 1982; Descy and Coste, 1991; Prygiel et al., 2002), designed to be calculated with LM data.

PRESENCE OF NON-DIATOM TAXA: USE AND POTENTIAL BIASES

Although being an important part of the phytoplankton community, diatoms often share their biotopes with other phytoplankton groups, such as yellow and green algae, Chrysophyceae, Dinoflagellate, Trebouxiophyceae, Cryptophyceae, and Cyanobacteria (Descy et al., 2012). Those taxa proved to be efficient bioindicators (Shams El-Din et al., 2022; Smol, 1985; Tsarenko et al., 2021) and yet are not routinely used in the standard biomonitoring. This is explained by the diatom exoskeleton that is historically easier to identify with Light Microscopy (LM). As molecular based biomonitoring methods are being used more frequently, it is valuable to integrate other phytoplankton/unicellular photosynthetic organisms in the ecological assessment because the limitations of the Light Microscopy are not shared with the Metabarcoding approach.

Due to the diatom exoskeleton (called the frustule), diatom DNA is less accessible than that of the yellow and green algae DNA (Vasselon et al., 2017a). Intuitively, this is likely to underrepresent diatom DNA during the extraction process when the samples are not exclusively composed of diatoms, which is the most common case during natural environmental samples (Del Carmen Pérez et al., 2009). Due to the high number of factors that should affect the DNA ratio between diatoms and other species, within this experiment

I investigate the phytoplankton DNA ratio (via HTS read count) between species and the consequences on the use of molecular data for biomonitoring compared to the tradition light microscope counting.

METABARCODING: LIMITATIONS TO ABUNDANCE ESTIMATES

In recent years, the use of Metabarcoding for diatom biomonitoring has been of increasing interest as its capacity to generate accurate qualitative species composition data has been shown. Notwithstanding this ability, the accurate quantitative composition data for Metabarcoding has shown to be more difficult to obtain (Mora et al., 2019; Vasselon et al., 2018). This is problematic as diatom biomonitoring relies on the abundance (and the ecological preferences) of each species identified in the samples (Kelly and Whitton, 1995; Prygiel et al., 2002), and therefore the proportion of each taxon must be accurate. While Metabarcoding is a powerful tool for biomonitoring, the difficulties in extracting DNA from some species relative to others, the nature of PCR amplification and the number of genes copy per specimen can result in inaccurate estimation of the relative abundancy of each species (Kelly et al., 2020). An additional consideration is the sequencing platform, where the characteristics of the platform (error rate, read length, read depth) affect the final sequencing data. The bioinformatics pipeline and reference database also create differences in the final data.

EFFECT OF DIFFERENCES IN RBCL COPY NUMBER AND BIOVOLUME

The amplicon region typically used in diatom Metabarcoding is the *rbcL* gene (Kelly et al., 2018), contained within the chloroplast. As each chloroplast contains at least a copy of the *rbcL* gene and the number of chloroplasts per cell is depending of the phytoplankton species (Bendich, 1987; Rauwolf et al., 2010; Round et al., 1990), the quantification of *rbcL* copies during Metabarcoding could affect the community inventories. The biovolume has been proven to be correlated to the number of *rbcL* copies in diatom mock communities (Vasselon et al., 2018). Moreover, centric diatoms are known to contain multiple chloroplasts compared to the single to few chloroplasts present in pennate diatoms cytoplasm (Bedoshvili et al., 2009).

Here I test how the *rbcL* number of copies in a DNA extract from communities made of both diatom and non-diatom phytoplankton species differ following the ratio of diatom/non-diatom. I hypothesized that the accessibility of the DNA is different among the different phytoplankton taxonomic groups as the morphological structure is rather different.

CHOICE OF PLATFORM: USE OF ONT MINION

The Miseq platform is constrained by its short read length, allowing amplicons of less than 400 base pairs (bp). Therefore, studies have focused on the use of short barcodes, such as sections of the *rbcL* region ,e.g. the "diat.barcode" (Rimet et al., 2019) or the barcode used in Kelly et al (2018), short sections of 18S gene, or COI gene (Kermarrec et al., 2014). Notwithstanding the high quality of data generate by MiSeq sequencing method, Oxford Nanopore Technology potentially offers a lower cost alternative to the Illumina platforms with its MinION platform (Lin et al., 2021). A drawback to the MinION platform is the reportedly lower accuracy relative to the MiSeq. As biomonitoring studies are more tolerant of lower read quality (due to the use of classifiers and identification similarities percentage) (Maitland et al., 2020) than rare species detection studies (in which read accuracy is key)(Hatzenbuhler et al., 2017), it might be that the MinION platform has the potential to be a suitable option for diatom monitoring using Metabarcoding. Differences between the two platforms are tested here.

EFFECT OF LONGER AMPLICON LENGTH OF RBCL

An advantage to the MinION platform is that it can give longer DNA reads (~10-60k bp) compared to the MiSeq platform (<400 bp)(Lin et al., 2021). In Pearman et al., 2020 , the comparison of MiSeq (Illumina) and MinION (ONT) output indicates that using a long low quality barcode instead of a short high quality on might be equivalent or even better if the barcode exceed 1500bps. As the most used diatom short barcode is located on the *rbcL* gene and the full length of this gene is ~1500bps (Valegård et al., 2018), the use of MinION sequencing with the full length *rbcL* could be a good alternative to the use of the short barcode with MiSeq. This amplicon is frequently used to obtain the diatom full *rbcL* sequences destined to reference library such as diat.barcode (Rimet et al., 2019) or the UK barcoding project (Kelly et al., 2018). This protocol was tested in the thesis PhD of Glover, 2019.

Nevertheless, in the context of biomonitoring, the use of ONT MinION technology with the short *rbcL* barcode, could provide reads of adequate quality for efficient taxonomic assignment. Within this chapter I also test the use of a longer *rbcL* fragment.

USE OF MOCK COMMUNITIES TO TEST THESE BIASES

Given the unsureness associated with morphological data (which uses cell count) and the two HTS platforms, in this chapter I undertake a direct comparison of the reported community composition between three Metabarcoding methods (Illumina MiSeq and MinION ONT Metabarcoding of short amplicons, ONT MinION Metabarcoding of long *rbcL* amplicon) on samples of known community composition (mock communities) to compare these different approaches.

Using natural samples to compare methods is strongly limited uncertainties of the composition of the underlying community as well as whether the diatoms are alive or not. An additional issue is that true diatom ecological communities are often highly diverse and therefore can contain species not present in the reference library (Vasselon et al., 2017c). Thus, I use mock communities, to investigate this (Vasselon et al., 2018). This provided a controlled and homogenous dataset in which to assess the repeatability of the Metabarcoding methods.

In this chapter, I created mock communities composed of pennate diatom, centric diatoms and non-diatom phytoplankton. These choices were driven by the known different number of chloroplasts between centric and pennate diatoms, the differing biovolumes of species (small and large), and the potential difference of DNA extractability between the shelled diatoms and the less protected phytoplankton taxa. The species were selected to include common species that are known to be easily detected by Metabarcoding studies (Kelly et al., 2020, 2018; Vasselon et al., 2017c).

I sequenced DNA extracts from each mock community with both Illumina MiSeq and ONT MinION in order to compare both sequencing platform for diatom biomonitoring. Finally, I tested two different *rbcL* amplicon barcodes (a full length and a short barcode) with the ONT MinION sequencer.

MATERIALS AND METHODS

MOCK COMMUNITY

Phytoplankton cultures from the Thonon Culture Collection (INRAE-CARRTEL Thonon-les-Bains, https://www6.inrae.fr/carrtel-collection_eng/) were used to create seven different mock communities including one Centric diatom, eight Pennate diatoms, three non-diatom green phytoplankton (Chlorophyta) species and one cyanobacteria (Table 1). The conditions of culture are fully detailed in the Thonon Culture Collection website (Frédéric Rimet et al., 2018). Diatoms (Bacillariophyta) are suspended in DV medium, Cyanobacteria in Z medium, and Chlorophyta in LC medium (details in Rimet et al 2018). They were conserved at 7°C, with an artificial daily photoperiod of 12 hours.

The mixing of the cultures to create each mock community was based on the preliminary observations of each pure culture to estimate the cell concentration using a haemocytometer (Improved Neubauer – counting chamber). A fixed volume (10 µL) of each culture was counted in triplicate, and the average value used to calculate the volume of culture to add to each mock community. Due to the tendency of phytoplankton to form biofilms and in a lesser extend to conglomerate, this required careful mixing steps prior to transfer. This consisted to repetitive pipetting (~10 times) when sampling from each pure culture to disperse the conglomerate and homogenize the pure cultures.

The concentration/proportion of each taxon is detailed below (Table1), as well as their attributes. The proportions and concentrations were based on the number of cells rather than biovolume, to make the communities equivalent to the individual cell counts used in the traditional light microscopic method.

The mock communities comprised:

- Mock Community 1 (MC1): A community composed of each species in the same proportion. (Species evenness)
- Mock Community 2 (MC2): A community composed of the centric and each pennate diatom in the same proportion and each non-diatom 10 times less concentrated.
- Mock Community 3 (MC3): A community with the centric diatom and each non-diatom in the same proportion and each pennate diatom 10 times less concentrated.

- Mock Community 4 (MC4): A community with each pennate diatom and non-diatom in the same proportion and the centric diatom 10 times less concentrated.
- Mock Community 5 (MC5): A community with each species in the same proportion except one of the non-diatom species 100 times more concentrated.
- Mock Community 6 (MC6): A community with each species in the same proportion except the centric diatom species 100 times more concentrated.
- Mock Community 7 (MC7): A community with each species in the same proportion except a one pennate diatom species 100 times more concentrated.

Each mock community was created in duplicate.

The exact composition of each mock communities is presented in Table 1 below:

	Binomial name	Chlamydomonas intermedia	Planktothrix rubescens	Botryococcus protuberans	Chlorella vulgaris	Nitzschia palea	Fragilaria cf. nanoides	Gomphonema parvulum	Achnanthidium minutissimum	Sellaphora seminulum	Staurosira venter	Pinnularia viridiformis	Cocconeis pediculus	Cyclotella meneghiniana
	TCC number	TCC003	TCC013	TCC123	TCC137	TCC139-1	TCC870	TCC612	TCC679	TCC828	TCC691	TCC890	TCC931	TCC640
	Average Cell Biovolume	50	400	68	65	391	470	331.2	76	69	315	13724	2281	1356
Composition of mock communities	Species Evenness MC1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1
	Less non-diatom MC2	x0.1	x0.1	x0.1	x0.1	x1	x1	x1	x1	x1	x1	x1	x1	x1
	Less pennates MC3	x1	x1	x1	x1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x1
	Less centrics MC4	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x1	x0.1
	One dominant non diatom MC5	x0.1	x0.1	x0.1	x10	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1
	One dominant centric MC6	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x10
	One dominant Pennate MC7	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x0.1	x10	x0.1	x0.1	x0.1	x0.1	x0.1
	Group	Non-Diatom microalgae Chlorophyta and Cyanobacteria (*Planktothrix*)				Pennate Diatoms								Centric Diatoms

Table 11 Mock communities' composition. Proportion of each culture added in each community: x1 is equivalent to the same number of cells regardless of the species, as such, x10 means there is 10 times more cells of this species compared to the species evenness and inversely x0.1 means there is ten time less cells of this species in this particular community compared to the species evenness community.

MOLECULAR METHODS

DNA Extractions

DNA extraction was undertaken straight after the preparation of the mock communities in order to prevent effect of diatom population growth on the species ratio.

I follow the protocol present in Kelly et al., 2018, detailed in the Materials & Methods Chapter and relies on DNeasy Blood & Tissue Kit which includes mini spin column and Protease K. Beforehand I prepared the samples by centrifugation of a 10 mL of each mock community (3,000g for 15 minutes) and resuspension of the pellet in 1mL of distilled water.

To evaluate the repeatability of the DNA extraction the duplicate underwent two separated DNA extractions which end up in four DNA extracts per mock community. For example, MC1Aa and MC1Ab are the DNA extraction duplicates of the MC1A mock community, and MC1Ba and MC1Bb are the DNA extraction duplicates of the MC1B mock community.

PCR amplification and amplicon choice

Two different amplicons within the *rbcL* gene were used, the 'long barcode' of the full length *rbcL* (including the spacer between *rbcS* and *rbcL*) and a short *rbcL* barcode (referred to as 'Short barcode') originally designed for diatom Metabarcoding survey in the UK using the MiSeq Illumina sequencing platform (Kelly et al., 2018).

MiSeq and MinION Short barcode PCR and sequencing

The short barcode amplicon is 331bps long, more fully described in the Materials & Methods chapter. The set of primer is from (Kelly et al., 2018), and is Forward primer *rbcL*-646F: ATGCGTTGGAGAGARCGTTTC, reverse primer *rbcL*-998R: GATCACCTTCTAATTTACCWACAACTG. The PCR conditions are as used in Chapters 3, 4 and 5 and described in the Materials and Methods Chapter.

Because of a lack of space in the MiSeq sequence run, no DNA extraction duplicate is present for MiSeq Illumina data. The methodological details for the MiSeq sequencing can be found in the Materials and Methods Chapter. Details of the MinION sequencing are given below.

Full length rbcL barcode

The long barcode amplicon is approximately ~1450bps, covering the whole *rbcL* gene and the small spacer region between *rbcL* and *rbcS* (coding the small chain of the RuBisCO protein). The amplicon was targeted with a PCR using the set of primers DP*rbcL*1 (5'-AAGGAGAAATHAATGTCT-3') and DP*rbcL*7 (5'-AARCAACCTTGTGTAAGTCTC-3') (Jones et al., 2005). The PCR protocol starts with 94°C for 3 minutes, followed by 35 cycles at 94°C for 60 seconds, 55°C for 60 seconds and 72°C for 90 seconds, followed by a final extension at 72°C for 5 minutes.

MinION Sequencing.

The protocol used for both short barcode and long barcode MinION sequencing was the official ONT PCR barcoding (96) amplicons (SQK-LSK110) protocol. The library preparation includes barcoding PCR ("tag"), DNA repairing, end preparation, adapter ligation, and beads purification between each previous step. The only notable difference is the concentration of Agencourt AMPure XP beads concentration related to the DNA length targeted during PCR purification step: ~ 350bps for the short amplicon and ~1450bps for the long amplicon. I used respectively x1 and x0.5 concentration.

Controls

Negative controls were implemented all along the process to assure the reliability of the results: a DNA extraction negative control (pure distilled water), a PCR negative control (no DNA extract added), and an index negative control during the Illumina Index PCR stage.

Positive controls were implemented for the same reasons and include DNA extraction control (sample with DNA successfully extracted during another study), PCR positive control (DNA extract from previous study), and an artificial oligo with binding sites for primer pairs t during the Illumina sequencing step.

BIOINFORMATIC ANALYSIS

Illumina MiSeq Data

The diatom-izer pipeline (see Materials & Methods Chapter and Chapter 3) was used to process the data from the MiSeq output. In brief, it comprises the following steps in a R script using the package DADA2(Callahan et al., 2016):

- quality and length filtering;
- trimming to the right length;
- error rate learning, dereplication;
- denoising using the DADA2 algorithm and the learnt error rate;
- merging of the paired reads (forward and reverse);
- chimera removal;
- Naïve Bayesian taxonomic assignment using the custom (supplemented with non-diatom phytoplankton) diat.barcode reference library created in Chapter 3.

ONT MinION data

As I wanted to use the latest bioinformatic tools for ONT MinION sequences, I created a new script in order to process the raw data from the MinION output for both long and short barcodes (Appendix C).

- The script was based on the NGSpecies script (Sahlin et al., 2021) with optimised parameters for diatom data, and added Naïve Bayesian taxonomic classification (Wang et al., 2007) to take advantage of the full lineage present in the diat.barcode reference library. The pipeline includes the steps below: Basecalling with Guppy (ONT)
- Quality filtering with NGSpeciesID (Sahlin et al., 2021)
- Length filtering with NGSpeciesID
- Clustering/polishing with NGSpeciesID.
- Taxonomic assignment: Naïve Bayesian classifier with the software Mothur (Schloss et al., 2009)

A schematic of part of the pipeline is given in Figure 32. The NGSpeciesID (Sahlin et al., 2021) component of the pipeline is a python-based program containing a set of tools to cluster the reads generated by the basecaller and "polish" (equivalent to the denoising step for Illumina sequencing) the relatively low-quality reads associated with ONT sequencing. The outputs are consensus sequences created from clusters of reads, that share high similarities of sequence, and merge reverse complement consensus sequences thereafter. This step is handled by medaka (https://github.com/nanoporetech/medaka) which is provided and developed by Oxford Nanopore Technologies. It relies on neural networks which correct the individual

sequences by comparison with a draft assembly. NGSpeciesID also include a primer-removal tool, length filtering and quality filtering based on Phred score.

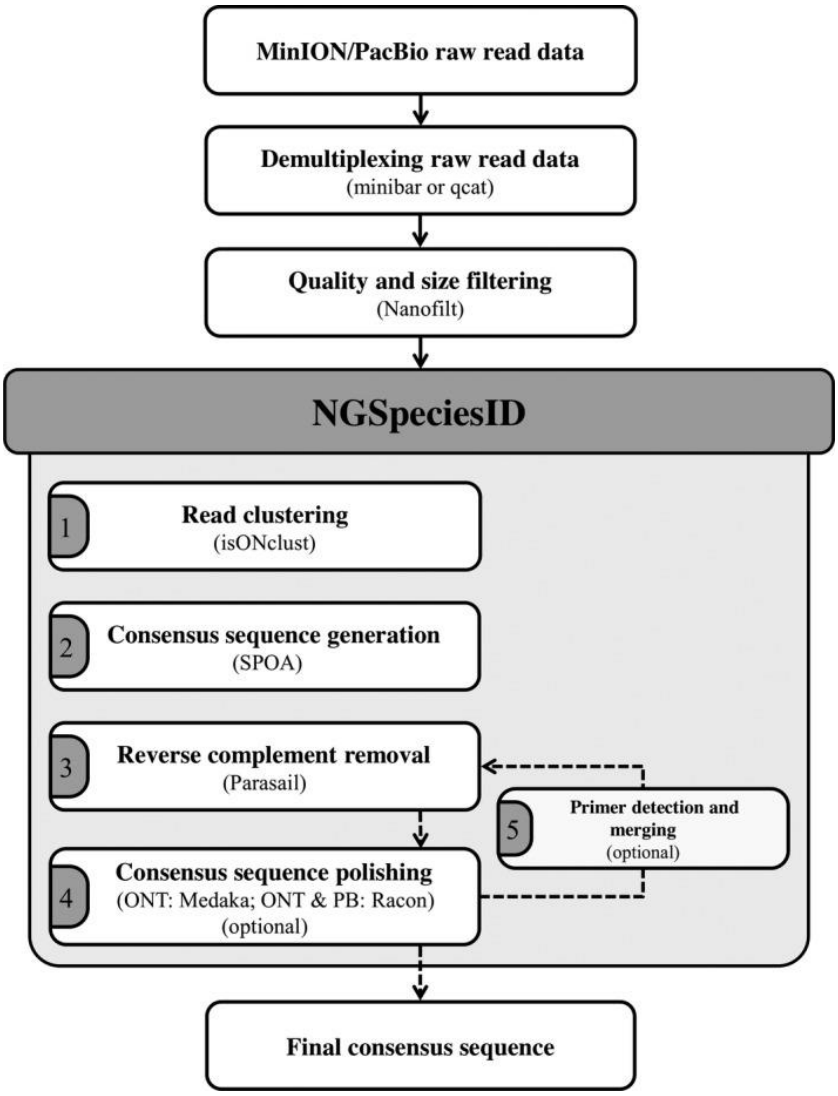


Figure 33 Steps involved in DNA barcode consensus calling of long-read data. The respective software tools used in the different steps are provided in brackets. For more details see Sahlin et al., 2021

The taxonomic assignment for ONT reads is usually performed by Basic Local Alignment Search Tool (BLAST) but I integrated an alternative assignment based on a Naïve Bayesian approach (Wang et al., 2007) . This Naïve Bayesian approach use the sequence and the lineage related to each sequence of the reference library to assign each candidate sequences to different taxa level with a confidence provided for each rank. This has the advantage of assigning sequences that are not present in the reference library at a higher taxonomic rank, e.g. a candidate read from a species not included in the reference library would be assigned by the Naïve Bayesian approach to a higher taxonomic such as Genus, provided reference library includes sufficient species from the same genus. Such functionality is not possible with

the BLAST. Nevertheless, Wang taxonomic classification needs the full taxonomy of each sequence whereas BLAST only requires species name.

The Wang taxonomic assignment was performed using the software Mothur (Schloss et al., 2009) on the galaxy server (Afgan et al., 2018) to run the assignment on each sample simultaneously. To preserve diversity information as much as possible, the abundancy ratio threshold was set very low: 0.0001, to take into account every read cluster that represents at least 0.01% of the total number of reads. A counterpart of this choice of parameter is the moderate extended processing time required.

Reference library :

A full length *rbcL* reference library was created, based on the sequences provide by the open data repository from diat.barcode (Rimet et al., 2019) with the addition of few sequences from GenBank of other phytoplankton and algae groups such as *Trebouxiophyceae*, *Cryptophyceae* or *Chrysophyceae*.

The full lineage was built to nine taxonomic levels: Domain, Kingdom, infraKingdom, Phylum, Class, Order, Family, Genus and Species.

Both short barcode analysis (MinION and MiSeq) used the same diat.barcode (Rimet et al., 2019) reference library used and detailed in Chapter 3 and the Materials and Methods Chapter 2.

STATISTICAL ANALYSIS

All outputs from the bioinformatic pipelines, which were OTU tables combined with the assigned taxonomy, were transferred to RStudio prior to analysis.

Community structure analysis: Hierarchical Clustering on Correspondence Analysis

The routinely used ecological index for diatom biomonitoring is the Trophic Diatom Index (TDI; Kelly, 1998) but this and other ecological indices do not seem like the suitable metric to compare samples in mock communities as their compositions differs greatly from the compositions in natural sites and the TDI was designed for natural environment communities. Moreover, the TDI relies on the alkalinity of the environment which is not applicable for *in*

vitro studies, especially with species culture in different culture medium. Hence, the use of community structure comparisons instead of ecological indexes such as TDI.

As the OTU dataset is a substantial dataset, multivariate data analysis was used to efficiently compare the similarities and differences among the samples to find clusters.

In Husson et al., 2010, the Hierarchical Clustering on Principal Component has been proven as an effective way to combine the three standard methods used in multivariate analyses : principal component method, hierarchical clustering and partitioning clustering. In this chapter I combined a correspondence analysis method, using the OTU table as a contingency table, with a hierarchical clustering using Ward's criterion (Murtagh and Legendre, 2011; Ward, 1963), and finally a k-means clustering to optimise the partition created with the hierarchical clustering.

The analyses were run on the R software using the FactoMineR Package (Lê et al., 2008) and the Manhattan distance. This was selected as the number of different species present makes the data have high numbers of dimensions.

Community relative abundance graphics were created using phyloseq package (McMurdie and Holmes, 2013).

RESULTS

ADDITIONAL REFERENCE SEQUENCES FOR THE OPEN ACCESS REFERENCE LIBRARY RSYST

This mock community experiment provides additional sequences for the diatom reference library, diat.barcode. Some of the phytoplankton taxa present did not have a sequence in the reference library but close related taxa sequences were present. The sequences assigned to the genus *Chlamydomonas* in the HTS data derives from the species *Chlamydomonas intermedia* which is the only *Chlamydomonas* present in the mock community. In a similar way sequences from *Botryococcus protuberans* were extracted. The sequences were sent to the study group of diat.barcode to be integrated in the next version of the open database.

DNA EXTRACTION AND AMPLIFICATION SUCCESS

All samples and duplicates were successfully extracted with the exception of one duplicate from the mock community 4 (sample name: MC4Ba), which generated no amplification at the PCR stage and was not included in the ONT MinION sequencing run. DNA extraction controls demonstrated the right execution of the extraction, the negative control was blank and the positive control demonstrated that the DNA extractions had worked.

PCR amplification was successful for all duplicates, and the negative controls (without DNA template) did not amplify DNA as intended.

SEQUENCING : NUMBER AND QUALITY OF READS

Both MiSeq Illumina and ONT MinION run were executed smoothly. Very few (<100) reads were obtained in the negative controls. The positive controls were all amplified which confirm that the different method sequencing steps were executed appropriately.

For the short barcode:

- Mean error rate for the MinION run, as estimated by the MinION platform as an output, was 8.13% which is in the expected range; this is the “raw” data before any processing. The average number of read per duplicates was 4,756.
- The MiSeq average error rate, as estimated by the MiSeq platform as an output, was <1% with an average number of read per duplicates of 173,974.

MinION method with the long amplicon (full *rbcL* + spacer ~1500 bps) produced a large number of reads (>100,000 in average). The mean error rate was 7.8%

The difference of number of reads between the two MinION runs (~5,000 vs ~100,000) are partially explained by the long barcode run only having the samples of this study while the short barcode was integrated in another run with samples from other studies (non-photosynthetic organism, to prevent contamination). The proportion of the flow cell allocated to each run was therefore different.

COMMUNITY INVENTORIES

MinION long barcode

For the long barcode, the community inventory obtained at the end of the bioinformatic pipeline was totally different from the original mock communities. It was impossible to detect and identify any non-diatom to species with the long amplicon. A significant proportion of unassigned sequences were present with less than 10 % of the total reads assigned to genus level.

Only two diatoms from the original mock communities were identified to the species level, and only in very limited numbers of the mock communities: *Sellaphora seminulum* was found in four different samples but never in the replicate from the same original mock community; *Nitzschia palea* was only found in a replicate from the Mock Community 2 (Fewer non-diatoms) and totally absent from the other replicate of the same community.

Nevertheless, 10 out of the 13 Genera initially in the mock communities were detected. The proportion were very variable and the taxa were generally present in only few duplicates.

With such an important proportion of unassigned sequences I decided not to run community structure analysis with the long reads data.

Short barcode sequencing: MinION and MiSeq

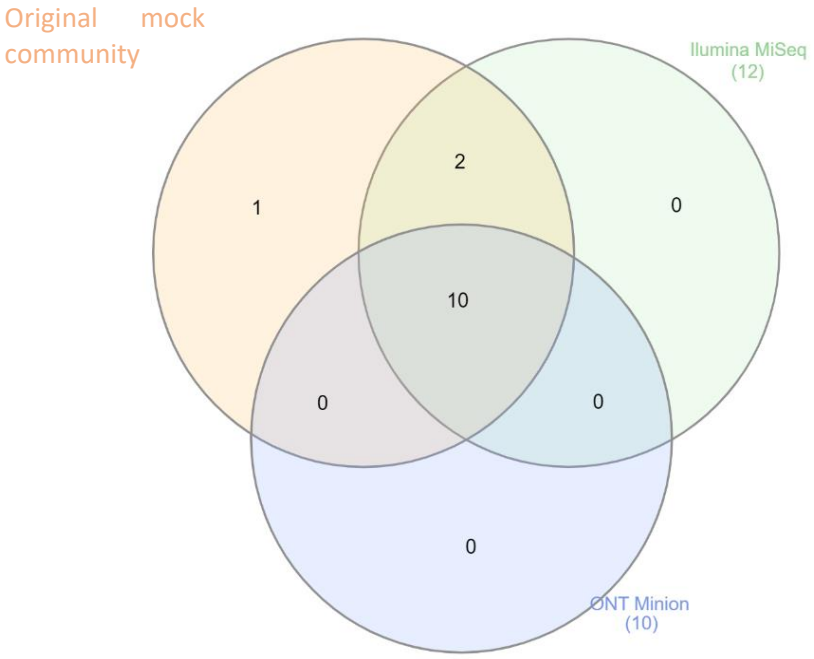


Figure 34 Venn Diagram of the proportion of species detected by the MinION and MiSeq short barcode sequencing, relative to the original mock community composition, with a relative abundance cut-off value of 1%. Bracketed figures indicate the proportion of the 13 species from the original community detected using the different method. Numbers present in the interface between several circles indicates the proportion of the 13 species detected by both method (or method and original community). The central area is the number of species detected by both platforms that were present in the original sample.

Detection of the species

The number of unassigned sequences at genus level was consistently lower than 5% in both the ONT MinION and MiSeq Illumina data.

The Illumina MiSeq method detected 12 of the original mock communities of 13 species and the ONT MinION method detected 10 species. All undetected species were non-diatom: *Chlamydomonas intermedia, Planktothrix rubescens, Botryococcus protuberans. Planktothrix rubescens* was the only species undetected on both platforms and it is the only member of the Cyanobacteria clade (Figure 33).

At Genus level both methods were able to detect 13 out of the 13 genera in the mock community.

An important proportion of the reads are assigned to a *Tetradesmus* in both the MiSeq and MinION data, which is a genus not found in the mock community. However, *Tetradesmus* is within the Chlorophyta group.

RELATIVE ABUNDANCE

The relative abundances for each mock community are given in Figures 34 & 35, with Figure 34 giving the MiSeq data and Figure 34 the MinION data.

Overall, the results for the MinION and MiSeq data gave similar results, with the same taxa over or underrepresented. The following results are seen in both MinION and MiSeq data. The compositions of repeats tend to be similar to the theoretical composition of the mock community (based on the ratio of cells added to each mock communities) in the community created with a dominant taxon (MC5, MC6 and MC7). In both the MinION and MiSeq data, the only centric diatom (*Cyclotella*) was overrepresented (24.6% of the Total MinION reads, 51.4% of the total MiSeq reads) in the "even community" (MC1), and the non-diatoms were underrepresented, except *Chlorella*, which is especially overrepresented in the MinION data (34.5% of the total reads). The pennate diatoms are not perfectly evenly represented, with a notable prevalence of the *Nitzschia* genus compared to others. The mock communities MC4 (built with fewer centric diatoms) present a surprising composition with the *Sellaphora* genus overrepresented, this particularity is present in all the duplicates of MC4 (34% of the MinION total reads, 62.1% of the MiSeq total reads). *Sellaphora* is also overrepresent in some duplicates from the MC5, MC6, and MC7. *Nitzschia* and *Sellaphora* are the most overrepresented pennate diatoms. *Sellaphora seminulum* and *Nitzschia palea* are both taxa with notably diverging proportions between the duplicates in the majority of mock communities.

MiSeq Illumina

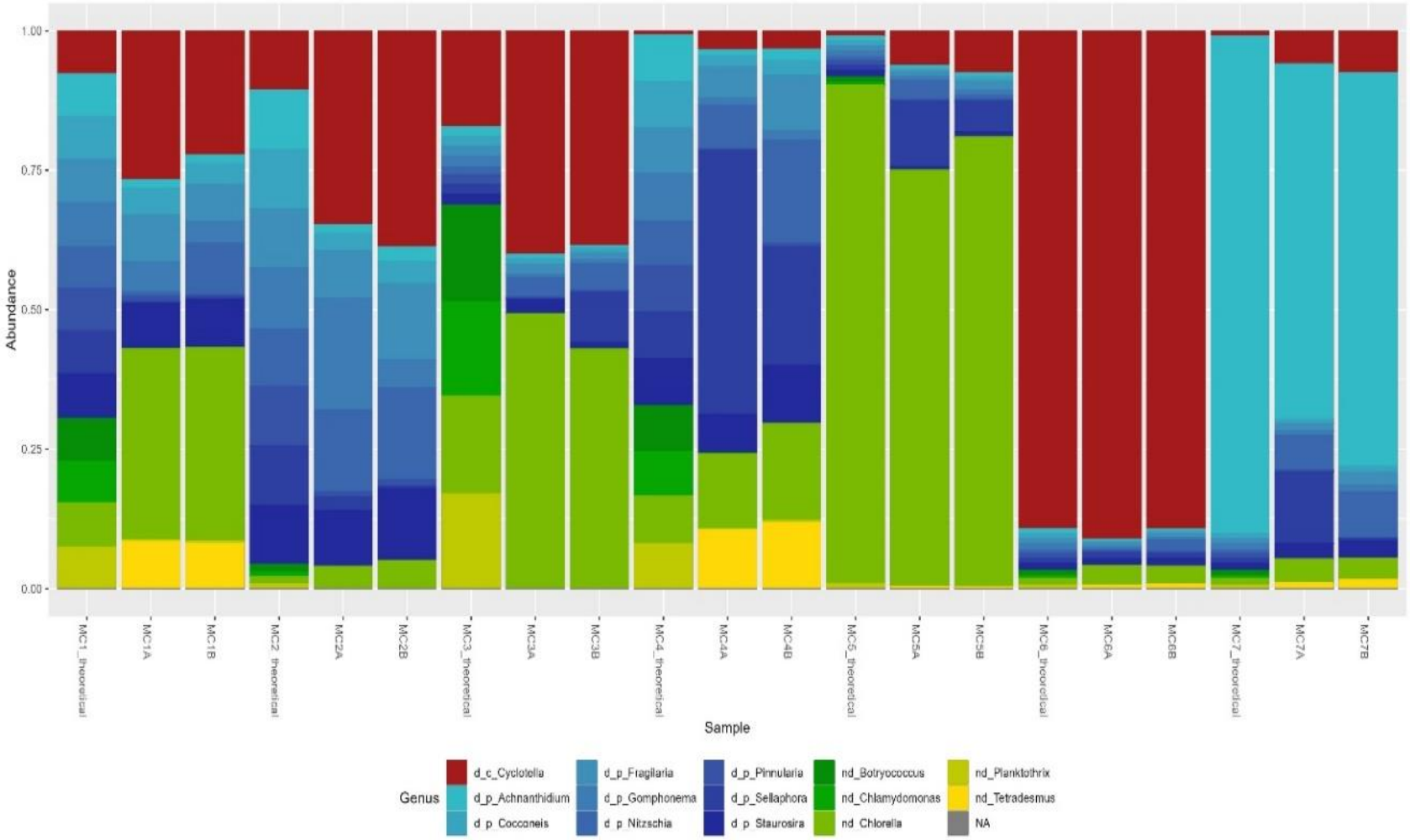


Figure 35 MiSeq Illumina data, showing the mock communities composition (measured by cell count) and metabarcoding (read count) relative abundance. The first bar for each mock community is the original community composition, two subsequent bars are the metabarcoding repeats for each MC. Centric diatom in red, pennate diatoms in blue, non-diatom phytoplankton in green.

ONT MinION

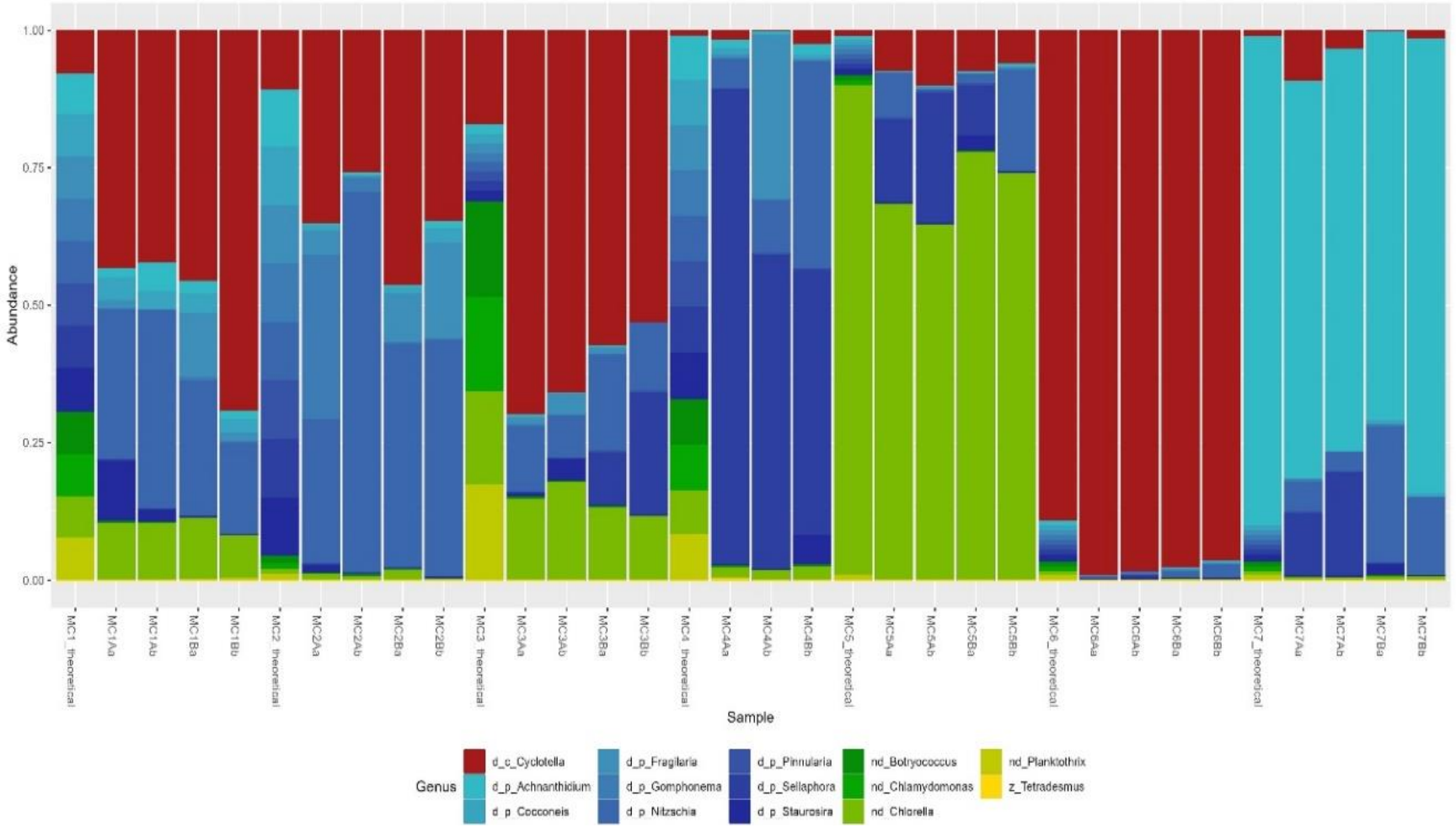


Figure 36 MinIon data, showing the mock communities composition (measured by cell count) and metabarcoding (read count) relative abundance. The first bar for the mock community is the original species composition, the subsequent three bars are the metabarcoding data. Centric diatoms in red, pennate diatoms in blue, non-diatom phytoplankton in green.

REPEATABILITY OF METABARCODING

The hierarchical clustering analysis results are presented in Figures 36 to 41. These figures are arranged to show the results of the Hierarchical clustering, used for identifying groups of similar observations in the contingency table (OTU table) of the mock community replicates as a site map (Figures 36 and 39), then the cluster dendrograms (hierarchical tree), presented to show the strength of the clustering relationships (Figures 37 and 40). These two sets of analysis are brought together in the final figures as a hierarchical tree superimposed on the site map with the final K-means clustering to improve the initial partition obtained from hierarchical clustering (Figures 38 and 41).

While the mapping is not a statistical test as such, the proportion of variance on the axis (Dimension percentage, Dim 1 and Dim 2) was notably very high in both MiSeq (Figure 36) and MinION (Figure 39), which is a signal of meaningful mapping, therefore, samples mapped together are structurally close.

For the MinION data, all the replicates from the same mock community were clustered together. The MC1 and MC2 formed a single large cluster of eight replicates (the four replicates of each mock communities) (Figures 39 to 41).

MiSeq data showed systematic clustering of the replicates from the same mock community, with the exception of the duplicates from the MC6 that were not clustered together (Figures 36 to 38)

The combination of these multiple, different methods of analyses demonstrates the robustness of these findings.

The results show that replicates (both DNA extraction and mock community mixing) from a same mock community are very close in term of community structure. Therefore, the repeatability of the method is high.

MiSeq Illumina

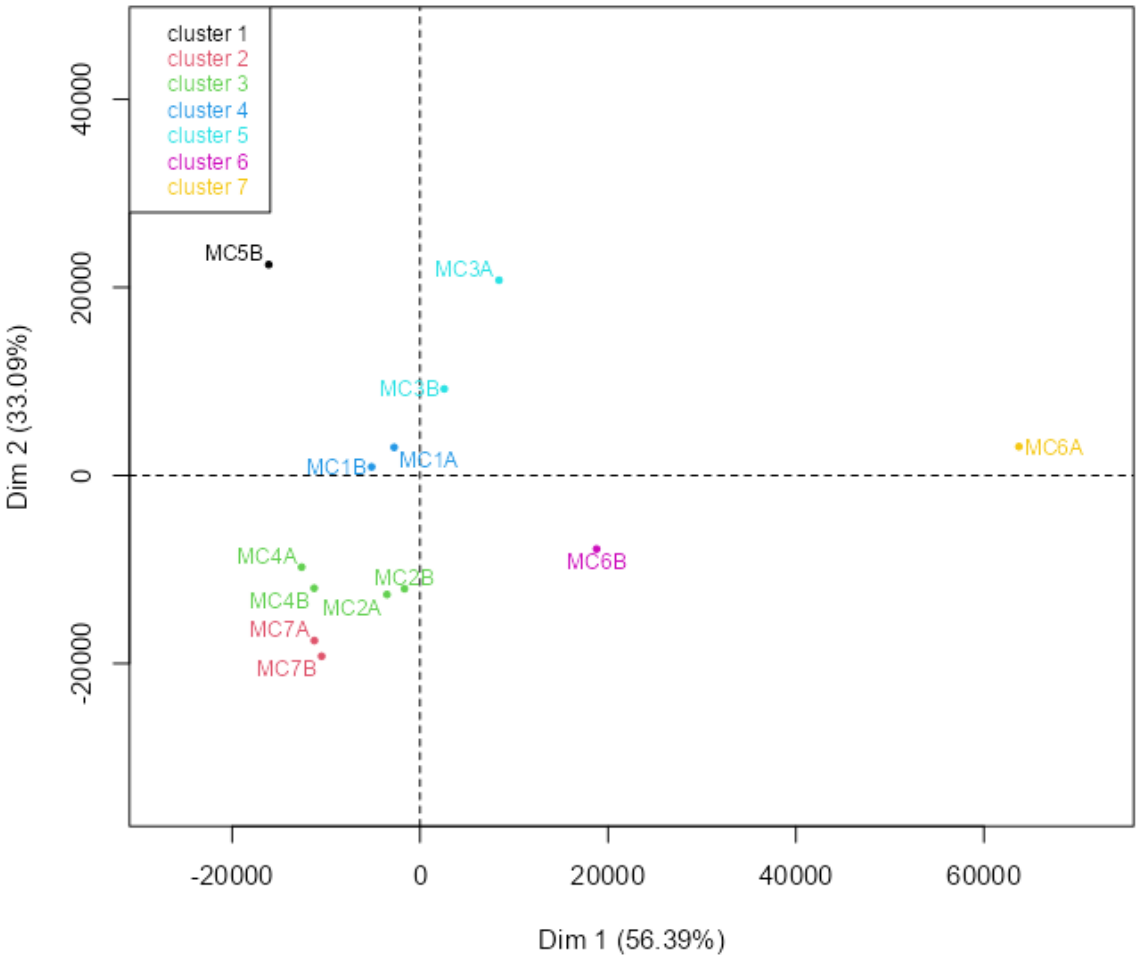


Figure 37 Site map of the Hierarchical clustering on the MiSeq Illumina data, where each Mock Community repeat is shown a single data point, and each K cluster is shown in a single colour. The proportion of variance explained by each dimension are present in the X (dimension 1) and Y (dimension 2) axis. The first two dimensions of the PCA express 89.48 % of the total dataset inertia.

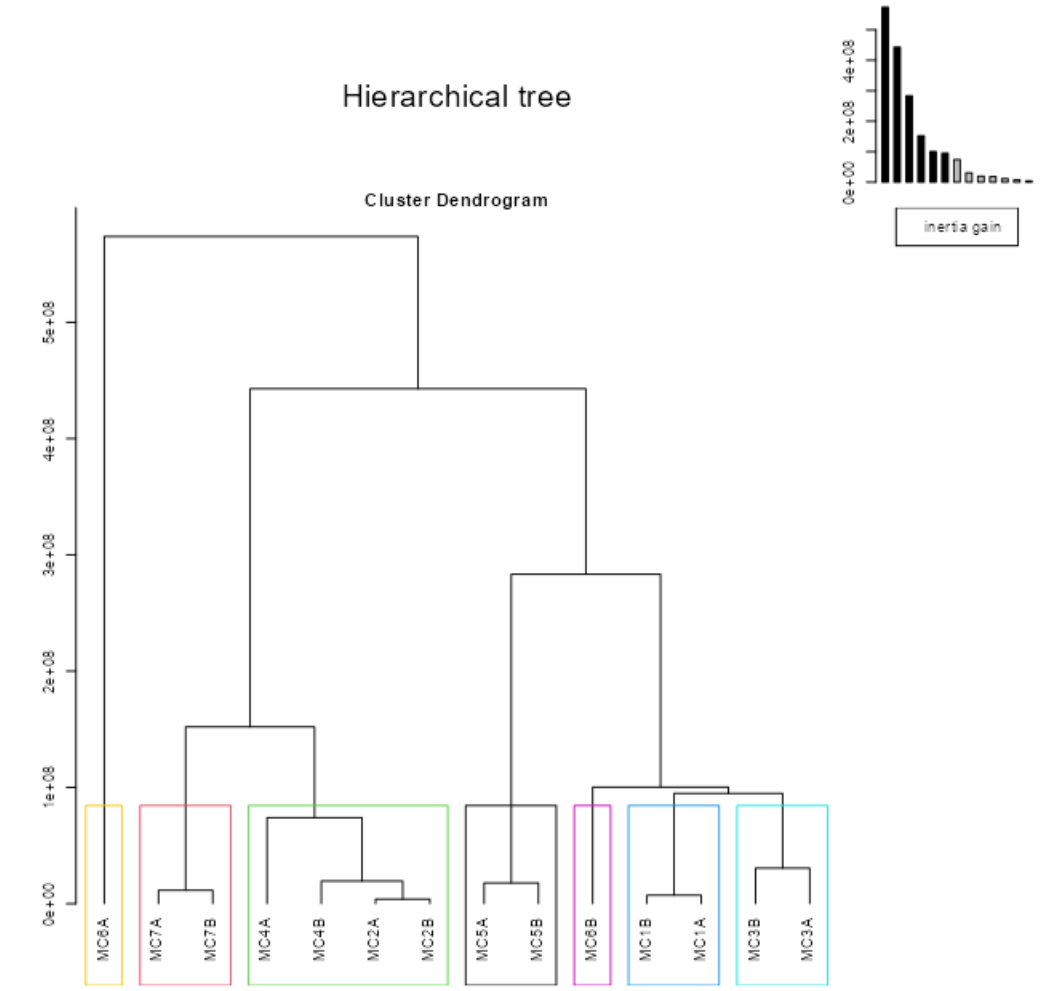


Figure 38 Dendrogram generated by the hierarchical clustering of the replicates for the different mock communities from the MiSeq Illumina sequencing data. Colours are the same clusters found in Figure 5.

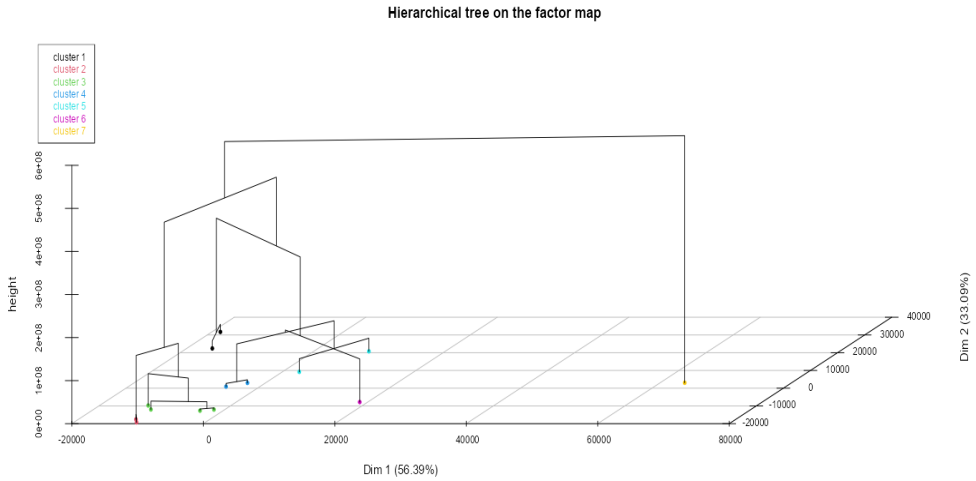


Figure 39 Three-dimensional plot combining the hierarchical clustering (figure 37) and the factorial map (figure 36) of the site map of the replicates from the MiSeq short barcode sequencing data. Each replicate is a single point, with clusters shown in the same colour

MinION ONT

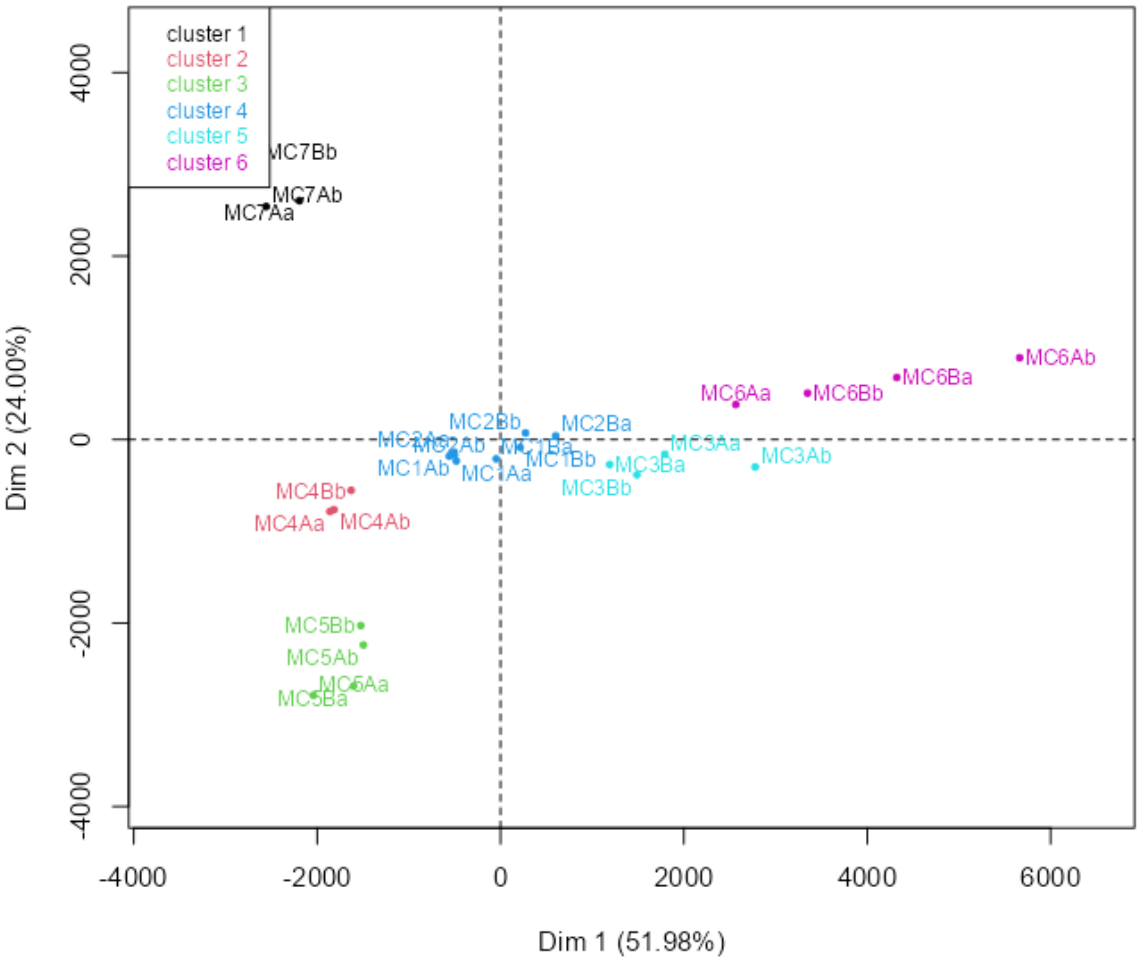


Figure 40 Site map of the Hierarchical clustering on the MinION data, where each Mock Community repeat is shown a single data point, and each K cluster is shown in a single colour. The proportion of variance explained by each Dimension are present in the X (dimension 1) and Y (dimension 2) axis. The first two dimensions of the PCA express 75.98 % of the total dataset inertia.

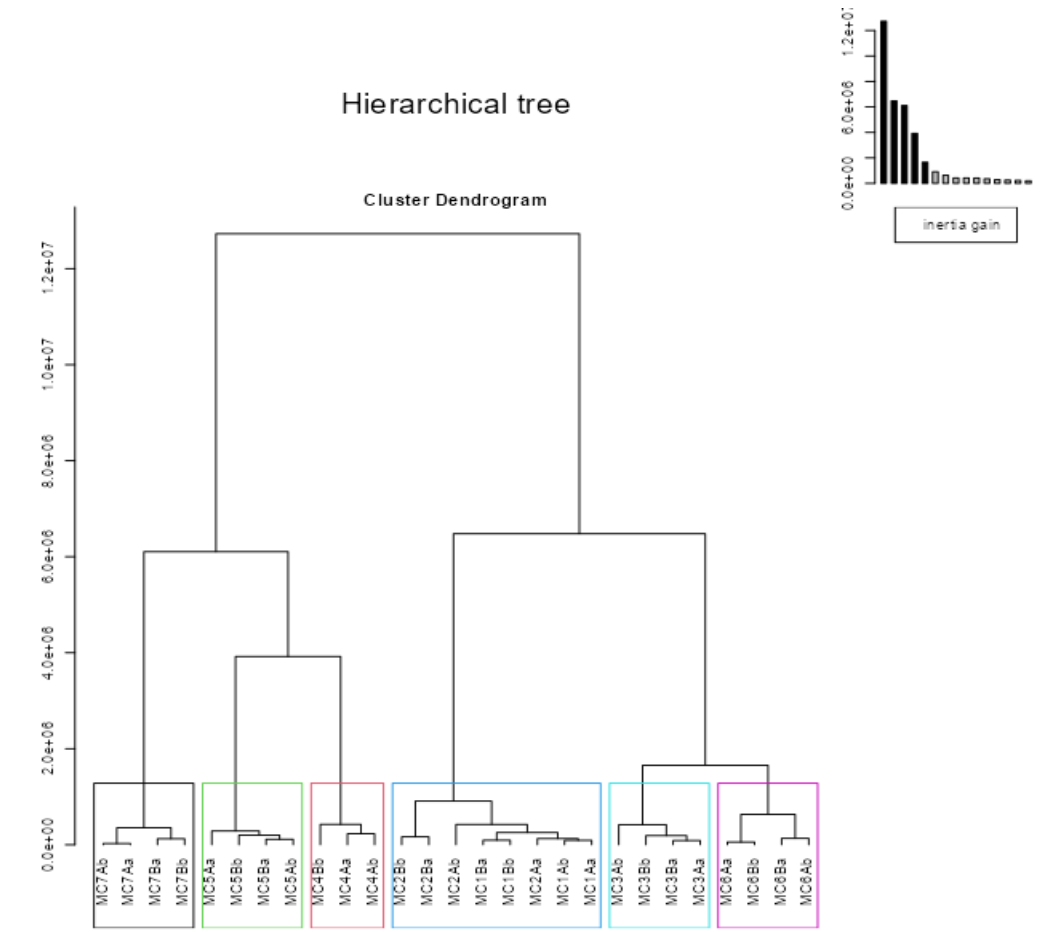


Figure 41 Dendrogram generated by the hierarchical clustering of the replicates for the different mock communities from the ONT MinION sequencing data.

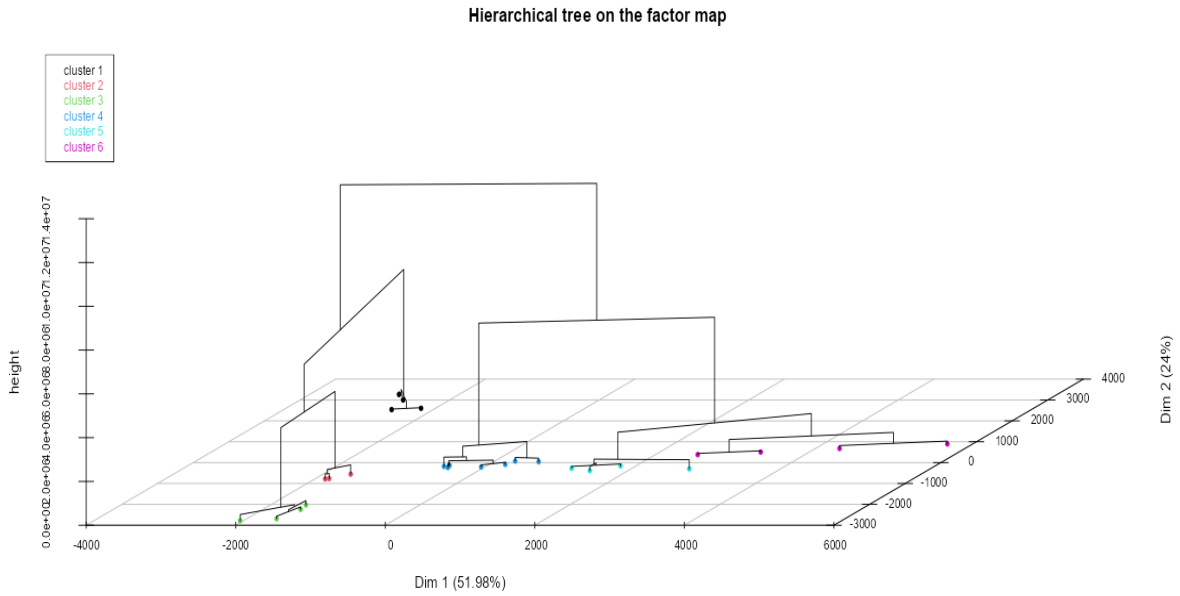


Figure 42 Three-dimensional plot combining the hierarchical clustering (Figure 40) and the factorial map (Figure 39) of the site map of the replicates from the MinION short barcode sequencing data. Each replicate is a single point, with clusters shown in the same colour.

DISCUSSION

MOCK COMMUNITY AS EXPERIMENTAL TOOL

The use of mock communities as a method to compare the two sequencing methodologies and to explore biases within these methodologies was confirmed. This study confirms the adequacy of mock communities as tools for experiment on phytoplankton. The experiments were simple to implement and provided sufficient control of the parameters tested. However, the design of the communities depends on the availability of pure culture collections, such as the Thonon culture collection, to provide high quality foundations for meaningful research. A relatively weakness in the current protocol was the mixing step as certain phytoplankton species tended to conglomerate. While the experiment went well, the use of a flow cytometer for sorting the cell could have been a great optimization to the method.

FULL LENGTH RBCL BARCODE MINION SEQUENCING

During this study I followed the method previously used in Glover, 2019, which used the primers from Jones et al., 2005, which produce the full length *rbcL* fragment. This has never been used with MinION sequencing in a published paper.

The full-length amplificon generated enough reads to extract full length sequences for additions in the reference library, however the low number of reads was insufficient to run a proper analysis. However, with the long barcode 10 the 13 non-diatom phytoplankton genera were detected with the 0.1% abundance cut-off value. The very high number of filtered reads (length and quality filtering, removal of unassigned sequences, removal of sequences assigned to contaminant such as *Solanum* and *Triticum*) compared to the low number of successfully assigned reads appears to show that the specificity of the long barcode primers is low.

The overall result with the long amplicon PCR primers shows that these were unsuitable for biomonitoring, although they were useful for extraction the full *rbcL* sequences of pure cultures. Better optimised primers for the long *rbcL* barcode may give more even and complete results. For example Valegård et al., 2018 designed a amplicon for diatoms that amplifies the whole *rbcL* gene as well as the *rbcS* gene and the spacer between the two genes.

In Hamsher et al., 2011 a 748bps long barcode of *rbcL* was designed, this could be a good alternative utilising the ability of MinION sequencing to generate longer reads than Illumina.

However, as demonstrated below, the shorter, more frequently used *rbcL* barcode region is sufficient for current river biomonitoring surveillance.

SHORT BARCODE : COMMUNITY COMPOSITION GENERATED BY METABARCODING COMPARED TO THE ORIGINAL MOCK COMMUNITY COMPOSITION.

Species detectability on the MinION and MiSeq

With the exception of the Cyanobacterium *Planktothrix*, not detected by either platform, all genera from the original mock communities were detected with Illumina. It is plausible that the Cyanobacterium *rbcL* gene is not amplified by the short barcode PCR as the primers were designed to target diatoms; Cyanobacteria and diatoms are evolutionarily distant which may result in poor primer specificity to the Cyanobacteria.

There is a consensus that the diatom evolution is marked by a first endosymbiosis (Figure 42), around 1.8 billion years ago, of a heterotrophic exosymbiont with a cyanobacterium endosymbiont (Falkowski and Knoll, 2007). The resultant proto algae was the common

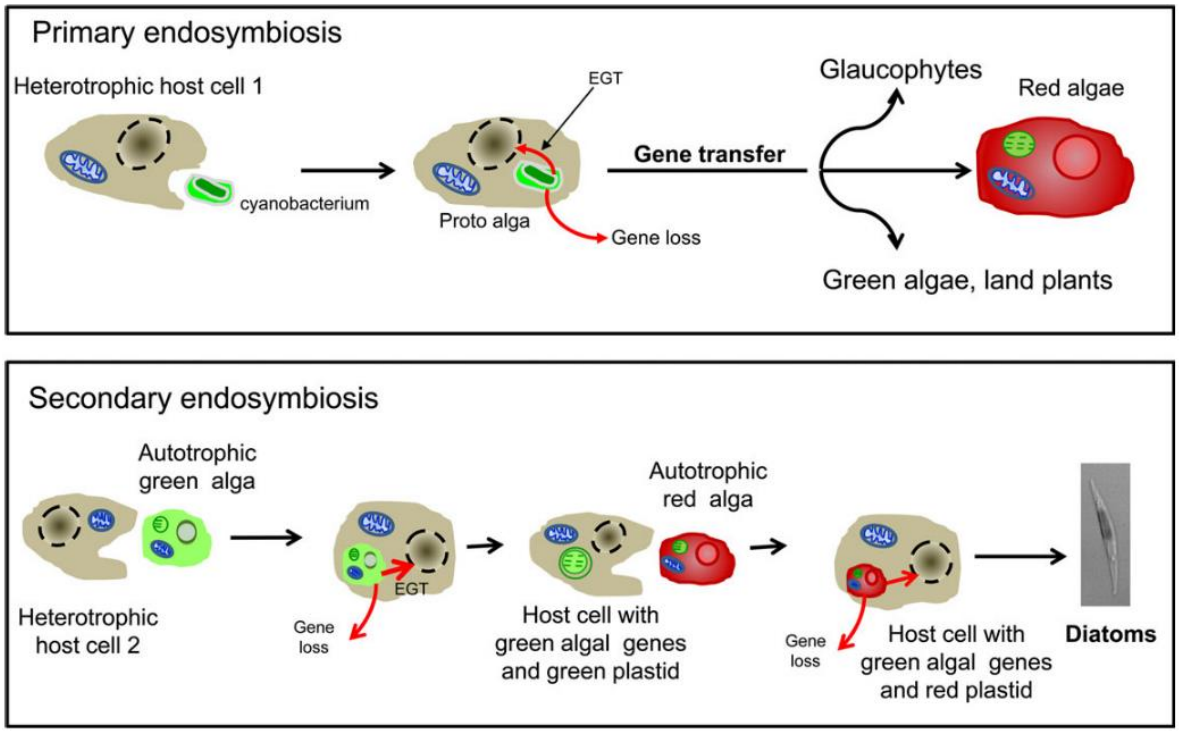


Figure 43 Schematic representation of the primary (upper panel) and secondary (lower panel) endosymbiont hypothesis of diatom evolution. From Falkowski & Knoll, 2007

ancestor of Glaucophyte, Red Algae, Green Algae and Land Plant. A second endosymbiosis apparently occurred 1.4 billion years ago (Yoon et al., 2002), when a Green Algae endosymbiont integrates another heterotrophic exosymbiont and a descendant of this evolutional event integrates, by another endosymbiosis, a Red Alga, which result in the common ancestor of diatom. This hypothesis, known as the secondary endosymbiont hypothesis (Moustafa et al., 2009), implies a more distant ancestor between Cyanobacterium(e.g. *Planktothrix*) and diatom than between Green Algae (e.g. *Chlorella*) and diatom.

While detection was efficient at genus level (Figure 35), it appears that detection was more difficult to obtain with the MinION outputs at species level, likely due to the relatively higher error rate (8% versus <1%). Moreover, a notable proportion of the MinION reads (and a minor component of the MiSeq reads) were incorrectly assigned to *Tetradesmus*, a non-diatom genus within the Chlorophyta group which is not present in the mock community. The lack of non-diatom sequences in the reference library is probably the reason of the misassignment as the Naïve Bayesian taxonomic classifier method needs a diverse and curated reference library to perform robust taxonomic assignment.

Overall, this is encouraging and shows that adding numerous common non-diatom reads (and their correct taxonomic lineage) is an easy improvement to diatom Metabarcoding results, especially if non-diatom phytoplankton are to be used for bioassessment.

VARYING REPRESENTATION IN THE METABARCODING DATA ACCORDING TO SPECIES

With the short barcode method (with both ONT and Illumina platform), the very large majority of the reads originated from the targeted amplified region, which means the PCR primers have a high specificity. Nevertheless, the read proportion of each taxon is still quite different to the proportion of cell of each taxon integrated in the mock communities. This is clearly shown in Figures 36 and 37 for the MC1 community, where the community has equal proportions of each species, yet the read proportions are highly skewed.

Overall, there is an overrepresentation of few genera for both platforms (Figures 36 & 37):

- *Chlorella* for the non-diatom phytoplankton: for example in the evenness community (MC1) this is 34.5% of the total MinION reads instead of the theoretical 7.7%.

- *Sellaphora* and *Nitzschia* for the pennate diatoms, for example in the MC4 these are 34% of the total MinION reads instead of the theoretical 8.3%.

- the centric diatom *Cyclotella*: for example in the evenness community (MC1), this is 24.6% of the total MinION reads instead of the theoretical 7.7%.

The *Chlorella vulgaris* overrepresentation is interesting as this species is reported to have only one chloroplast per cell (Wakasugi et al., 1997), indicating that copy number per cell alone is not only the explanatory variable. The reason for the over representation of this taxon may be that the chloroplast DNA from *Chlorella vulgaris* is especially easy to extract and/or well preserved, but could also be a primer bias (this could have been investigated via *in silico* methods or qPCR if the time and resources were present at the moment). As such, and as it is a high abundance species in freshwater environment (Wirth et al., 2020), *Chlorella vulgaris* DNA is likely to be over-represented relative to diatom DNA during Metabarcoding-based diatom biomonitoring studies. The non-diatom taxa are, aside *Chlorella*, under-represented and as the barcode was designed to target diatom specifically it is an expected result. This could be an explanation of the community structure similarities between MC1 and MC2 as their only difference is the lower abundancy of non-diatom taxa in MC2, which composed the overall most under-represented group in the mock communities.

While the overrepresentation of *Cyclotella* was predicted as centric diatom are known to have a large number of chloroplasts compared to pennate diatoms (Bedoshvili et al., 2009), this does not explain the overrepresentation of *Sellaphora* and *Nitzschia*. The overrepresentation of these two taxa is unexpected as in other diatom mock communities studies they did not show overrepresentation, either when the primers were the same (Kelly et al., 2018) or the alternative diat.barcode primers (Vasselon et al., 2018). This experiment used mock communities composed of a mix of individual phytoplankton rather than mock communities made of a mix of extracted DNA, as in the other studies. This suggests that the overrepresentation originates from the DNA extraction step; it could be a structural specificity of *Sellaphora* and *Nitzschia* that makes their DNA easier to extract. *Nitzschia* is a long and frangible diatom, as such it is plausible that its DNA is easily released during DNA extraction. In Chapter 4, the diatom community results show the same an overrepresentation of *Nitzschia* genus in the Metabarcoding community compared to the LM identification

community, again suggesting that the overrepresentation of *Nitzschia* in is linked to the DNA extraction step. *Sellaphora* genus does not show the same tendency in the Chapter 1&3 data. The overrepresentation of *Sellaphora* is not explained by literature or my results.

When comparing MinION data and MiSeq data, I can note that the groups that are dominant and overestimated are even more overestimated in the MinION data than in the MiSeq data. Some differences in the two method could explain this, for example a high number of PCR cycles are known to overestimate the larger groups and underestimate the minor groups and therefore lower the detected diversity of the community(Kelly et al., 2019).

Overall, these data show that Metabarcoding does not accurately reflect community abundance generated by LM. This may have implications for water quality index calculation such as TDI. However, the new versions of the TDI take into account the difference of community abundance specific to the Metabarcoding method (see Chapter 4).

REPEATABILITY OF PHYTOPLANKTON METABARCODING WITH SHORT READ MISEQ AND MINION PLATFORMS

The MinION data (Figures 41, 42 and 43) show a clear clustering of replicates and, at the same time, separate samples from different mock communities. Only samples from the MC1 (evenness/equity) and from MC2 (Less-non diatom) were not separated by the clustering and were too close structurally to be differentiated in different groups, despite the underlying differences in the original mock communities.

For the MiSeq data, most replicates were grouped together while being efficiently separated from one mock community to the others. In this dataset, the samples from MC2 (Less non-diatom) and MC4 (less centric) could not to be grouped separately, and the replicates from MC6 (One dominant centric) were both groups alone in a separate group.

The conclusion of the comparison of both Hierarchical clustering indicates that the results generated by both sequencing technologies are very similar and the duplicates of each mock community generally share similarities in term of community structure. It is important to note that the MiSeq data would have more difficulties to group replicates together because they are composed of only two replicates per mock communities while MinION data is composed of twice the number of replicates per mock community.

The overall comparison shows a close similarity between each pair of duplicates, both DNA extraction duplicates and PCR duplicates. Only the PCR duplicates from MC6 with MiSeq platform were not clustered together, although being alone in their own cluster. Moreover, while abundancy of each taxon was slightly different among the replicates of each mock community, the detection of species was always the same (Figures 36 & 37). The precedent points show a great repeatability and corroborate with other studies comparing mock communities replicates (Vasselon et al., 2018) and PCR replicates from river samples (Kelly et al., 2018).

However, few taxa proportions change significantly between the duplicates and the official biofilm sampling method (Kelly et al., 2018), which involves pooling at least 3 environment samples per sampling sites, should address this problem quite well. Pooling replicates of DNA extraction should be able to provide even better repeatability at the cost of extra sampling time (collecting three times more samples per sites).

MINION SEQUENCING FOR DIATOM AND PHYTOPLANKTON BIOMONITORING

This new bioinformatic pipeline for MinION data from diatom samples run efficiently for all samples and provide high quality data for biomonitoring. I have coupled MinION sequencing with read clustering and consensus forming (NGSpecies) and these bioinformatic tools are crucial to improve the quality of the output by “polishing” the reads. MiSeq sequences have also been processed in order to improve their quality with a “denoizer” (DADA2)(Callahan et al., 2016) which is a similar bioinformatic approach.

Although sequencing quality was considered as the major factor of choice for Metabarcoding diatom biomonitoring, the bioinformatic tools, and especially sequences “denoiser” and “polisher”, are other significant factors to make both sequencing platform adequate for biomonitoring. Nevertheless, biomonitoring is a field of science that is rather tolerant to small sequencing errors or small nucleotide changes. For mutation analysis, subspecies identification, or even detection of very rare species, the use of high-fidelity sequencing is required, and as very specific low occurrence sequence patterns could be erased by the denoising/polishing approach.

Previous studies propose that MinION sequencing could be use with long reads (>1500 bps) (Laver et al., 2015) but the new insights of this study indicate that even short reads (<400bps) generate similar outputs to MiSeq sequencing in term of community structure and biomonitoring usability. The ONT MinION platform has showed clear potential for diatom biomonitoring survey based on Metabarcoding. Moreover, the bioinformatic tools have a major impact on the suitability of the data. This new bioinformatic pipeline for ONT MinION reads for diatom biomonitoring worked smoothly and showed that there is alternative to the more costly Illumina sequencing platform. To confirm the utility of MinION data, further environmental studies that use ecological and water quality indexes calculation will need to be done.

CHAPTER 6 POSITIVE SELECTION IN DIATOMS ASSOCIATED WITH SPECIES MORPHOLOGY AND ECOLOGY

INTRODUCTION

Positive selection is the process that drives the increase in prevalence of advantageous genetic variants (traits) in a population (Anisimova et al., 2001). It is the natural selection that has been described by Darwin as the force promoting the spread of beneficial alleles. Its counterpart is the purifying/negative selection which purges the deleterious traits on the fitness of the individual in a population (Massingham and Goldman, 2005). The study of both positive and negative selection is necessary to estimate the contribution of natural selection to molecular evolution. These forms of selection influence the conservation or the removal of sequences patterns in accordance with the population history and interaction with the environment. Conversely, both phylogenetic and DNA barcoding at species level studies should ideally be run using DNA sequences that are selectively neutral, so that they reflect the population history and taxonomy of a particular organism, rather than the selective history of specific gene region (Deagle et al., 2014). A better understanding of the forces affecting the evolution of the genes used in phylogenetic studies will improve the understanding of the evolution history of the gene and, by extension, the evolution history of taxa. Moreover, the recent use of conservative regions for barcode identification of taxa relies upon the barcode region being selectively neutral, as positive or negative selection on the barcode region may affect whether it discriminates between species (for a counter example, see Percy *et al* 2014). A good understanding of the selective forces acting on a DNA barcode region is therefore necessary.

The *rbcL* gene is a coding region that encodes the large subunit of the ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo) which is arguably the most abundant protein on Earth (Erb and Zarzycki, 2018). This enzyme is responsible for practically all the Carbon fixation occurring on Earth as it is involved in photosynthetic CO_2 assimilation and photorespiratory Carbon oxidation. Due to the conservative and universal nature of the *rbcL* gene, it has been widely used in phylogenetic studies from phytoplankton to land plant (Bailet et al., 2020). Nevertheless, molecular analysis found positive selection in *rbcL* of the majority of the land plants (Kapralov and Filatov, 2007).

Diatoms (Bacillariophyta) are known to be one of the most productive photosynthetic organisms with the marine diatoms responsible for 40% of the total marine productivity (Tréguer et al., 2017). RuBisCO protein is formed of 8 long chains (coded by the *rbcL* gene), which contain the active site of the enzyme, and 8 small chains (coded by the *rbcS* gene). Diatoms produce a specific form of RuBisCO called form ID (see Figure 43). Studies have found that the diatom *rbcL* gene has evolved under a positive selection and especially followed the history of Carbon dioxide concentration fluctuations (Kapralov and Filatov, 2007; Young et al., 2012).

In this study, I explore the different characteristics (morphology and ecology), that potentially drove the evolution of the *rbcL* gene in the Bacillariophyta clade: salinity environment, morphological symmetry and pyrenoid structure. These variables are directly linked to the intracellular Carbon concentration as well as quantity and shape of chloroplasts, which are all limiting factor of the diatoms primary production.

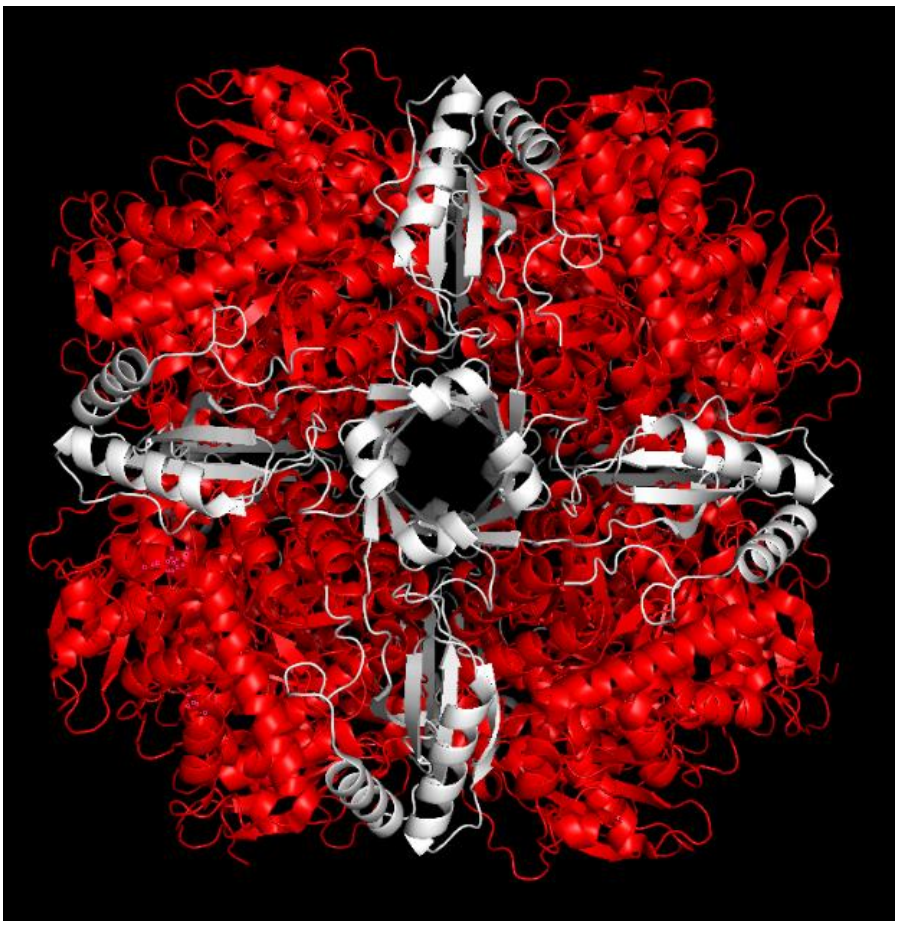


Figure 44 3D view of the overall structure of Rubisco (form I D diatom) from Thalassiosira hyalina. The large subunits are in red and the small subunits in white.

CENTRIC VS PENNATE

The most evident difference in structure within the diatom clade is between the centric and the pennate diatoms (Figure 44). The centric diatoms present a radial symmetry while most of the pennate diatoms present a bilateral symmetry. Moreover, the centric diatoms are more primitive and integrate multiple small chloroplasts compared to the few (if not single) larger chloroplasts specific to pennate diatoms morphology.

Due to these important differences of structure that also affect the chloroplasts, I hypothesize that the evolution of the *rbcL* gene has potentially been under a different positive selection during the evolution of these clades.

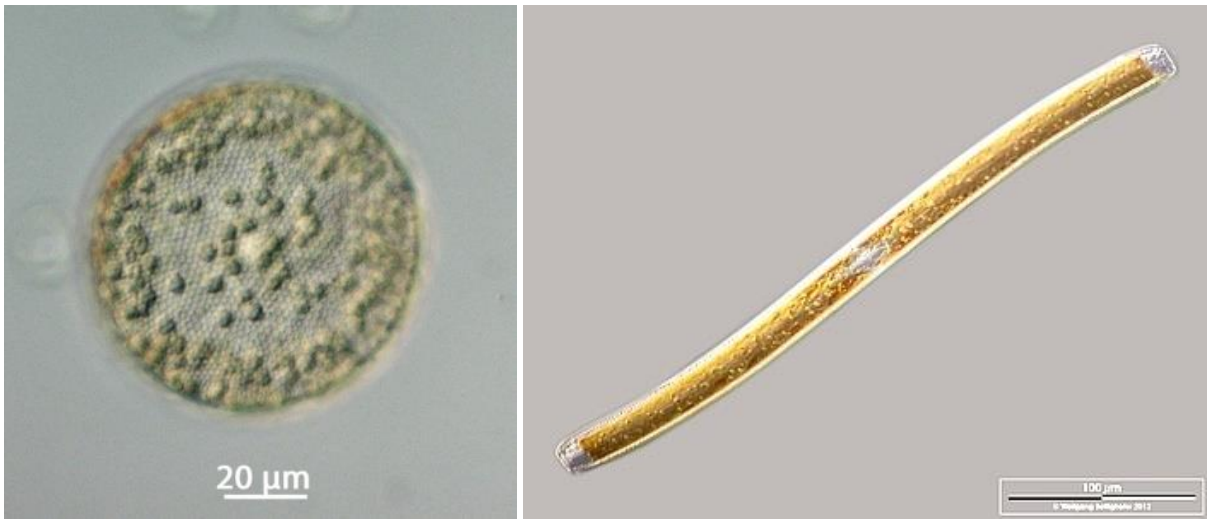


Figure 45 Optical microscopy photographs of a centric diatom (*Roperia tesselata*) (left) and a pennate diatom *(Nitzchia sigmoides*) (right). The several small chloroplasts in centric diatom are opposed to the few large chloroplast in the pennate diatom. Plankton*Net Data Provider at the Alfred Wegener Institute for Polar and Marine Research hdl: 10013/de.awi.planktonnet

MARINE VS FRESHWATER

As a widespread clade, diatoms are present in all waterbodies including a large range of salinity. Marine diatoms are known to be responsible for around 40% of the total ocean oxygen production which directly originates from the activity of the RuBisCO protein.

Despite global ocean acidification that is currently occurring (Hönisch et al., 2012), the last 100 million years have presented an increase in the pH of the ocean, and the ocean remains alkaline with an average pH of 8.2 (Marion et al., 2011; Tyrrell and Zeebe, 2004). The pH in

the majority of fresh water bodies lies between 6 and 8, and the average pH of UK rivers is 7.44 (*River Water Quality Monitoring 1990 to 2018 - PH - Data.Gov.UK*, n.d.). In a high alkalinity environment such as an ocean, the main form of Carbon is HCO_3^- which leaves a low concentration of free CO_2 in the water available to diatoms.

In addition to this limited access to Carbon, the nutrient limitation in the ocean, especially of Nitrogen and Phosphorus, leads marine species to have selected larger species among the diatom evolution compare to freshwater species (Litchman et al., 2009). This is mainly driven by the need for a wider exchange surface with the environment to compensate the low availability of both Carbon and nutrient.

Hence the reason for selecting the comparison between saline and freshwater diatoms as my hypothesise is there is a different evolutionary signature affecting their *rbcL* genes according to the salinity of their environment.

CLADE COMPARISON: PYRENOID STRUCTURE

The pyrenoid is a single or multiple microcompartment present in the chloroplast and composed of condensed RuBisCO. It is the main centre of Carbon dioxide fixation for most algae clades (Badger et al., 1998; Raven, 2010). The dense accumulation of RuBisCO combined with the effect of inorganic Carbon pumps and Carbonic anhydrases lead to Carbon dioxide concentration near the pyrenoid. The diversity of shapes and numbers of pyrenoids are specific to each diatom species and might be the result of a long evolutionary process to adapt to the environment of each diatom, and especially of the Carbon dioxide bioavailability.

I decided to focus on the effect of three different pyrenoids shapes within clades of diatoms on the *rbcL* gene evolution. My hypothesis is there would be an evolutionary signature related to pyrenoid structure due to the link between pyrenoid structure and its ability to concentrate CO_2.

Clade models for phylogenetical studies allow differences in site-specific selective constraints among clades in the tree (Bielawski & Yang, 2004; Forsberg & Christiansen, 2003).

The diatom species have been grouped in three clades regarding their particular pyrenoid structure described in Mann, 1989:

- *Navicula* and its allies (*Seminavis, Hippodonta, Trachyneis, Pleurosigma, Gyrosigma*) which have one or more bar-like pyrenoids (appears strictly rectangular in face view) that is not invaginated.
- *Placoneis, Cymbella, Gomphonema* and other genera with pyrenoids that form a bridge between the two halves of a chloroplast: in this case the pyrenoid seems 'exposed' rather than embedded inside the chloroplast.
- *Pinnularia* and *Caloneis:* all have invaginated pyrenoids: the body of the pyrenoid is penetrated by tubular invagination lines of cytoplasm.

The overall objectives of this study are to determine if positive selection occurred during evolutionary diatom history and to identify the drivers of these hypothetical positive selections.

MATERIALS & METHODS

PAML DESCRIPTION (MAXIMUM LIKELIHOOD)

Historically two methods have been created and used to detect positive selection in homologous protein coding sequences: a parsimony method from SuzUKi & Gojobori (1999), and a likelihood method based on the work of Nielsen & Yang (1998). Here I use a Maximum Likelihood (PAML, Yang, 2007) phylogenetic analysis to reveal specific evolutionary signatures within diatom clades, a modified version of the methods used by Nielsen & Yang (1998). It enabled me to compare distinct diatom groups and search for difference among taxonomic groups attributable to positive selection. The assessment of positive selection is based on an analysis of the ratio of the number of non-synonymous substitutions (amino acid altering substitutions) to the number of synonymous substitutions (dN/dS or ω ratio) to estimate the balance between negative or purifying selection ($\omega < 1$), neutral selection ($\omega = 1$) and positive selection ($\omega > 1$). In general, when positive selection has been significantly detected in a protein, the adaptative evolution is driven by only a few amino acid sites (Hughes and Nei, 1988; Yang, 2007)

I used the software EasycodeML (Gao et al., 2019) to run the different codon-based models as it is an updated version of the original PAML program that integrates a graphical user interface, multi-threading, and performs the likelihood ratio test.

In a multiple sequence alignment, the site model assumes that the ratio of nonsynonymous to synonymous substitution rates (ω ratio) is constant between the branches of the phylogenetic tree but different among sites in the aligned sequences. There are several codon substitutions models as the 4 different nucleotides allow 64 ($=4^3$) possible codons and the non-synonymous/amino acid-altering substitutions are under a more restrictive selective pressure. I used some of the codon-substitution models M0-13 from the work of Yang et al., 2000:

- M0 (one-ratio)
- M1a (nearly neutral)
- M2a (positive selection)
- M3 (discrete)
- M7 (beta)
- M8 (beta and ω > 1)
- M8a (beta and ω = 1)

The complete list of codon substitutions and the description of their parameters are present in the annex section.

I used a likelihood-ratio test to compare the fit of the different models to the sequence data. Evidence of positive selection can be revealed by a better fit with M2a over M1a, or with M8 over M7 or M8a. (Anisimova et al., 2001; Swanson et al., 2003; Wong et al., 2004; Yang and Nielsen, 2002).

In the M8vsM8a test, a new LRT is implemented to determine if the d(N)/d(S) ratio is significantly greater than one. This is a more refined test of positive selection than the previous LRTs which only identified if there was a class of sites with a d(N)/d(S) ratio >1 but did not test if that ratio was significantly greater than one.

Consequently, the M8vsM8a is considered as the gold standard test. Nevertheless, the M8vsM7 test and, to a lesser extent, the M2avsM1a test are still consider as robust tests.

As mentioned earlier, the diatoms have been grouped using the following 3 criteria:

- Diatom primary shape (centric or pennate)
- Salinity preferences

- Pyrenoid structure/morphology

In order to perform the analysis, datasets were created from the full *rbcL* sequences provided in the diat.barcode open access database (created and curated by INRAE-CARRTEL Thonon (Rimet et al., 2019)).

RESULTS

CENTRIC VS PENNATE

A dataset composed of 46 centric diatom species and 78 pennate diatom species, 124 sequences in total from the open dataset DIAT.BARCODE (INRAE- UMR CARTEL Thonon-les-Bains- France; Rimet et al., 2019) was used to generate the phylogenetical tree (Maximum Likelihood with 100 replications, Figure 45). The dataset and the phylogenetical tree were used conjointly to run the site model in EasyCodeML.

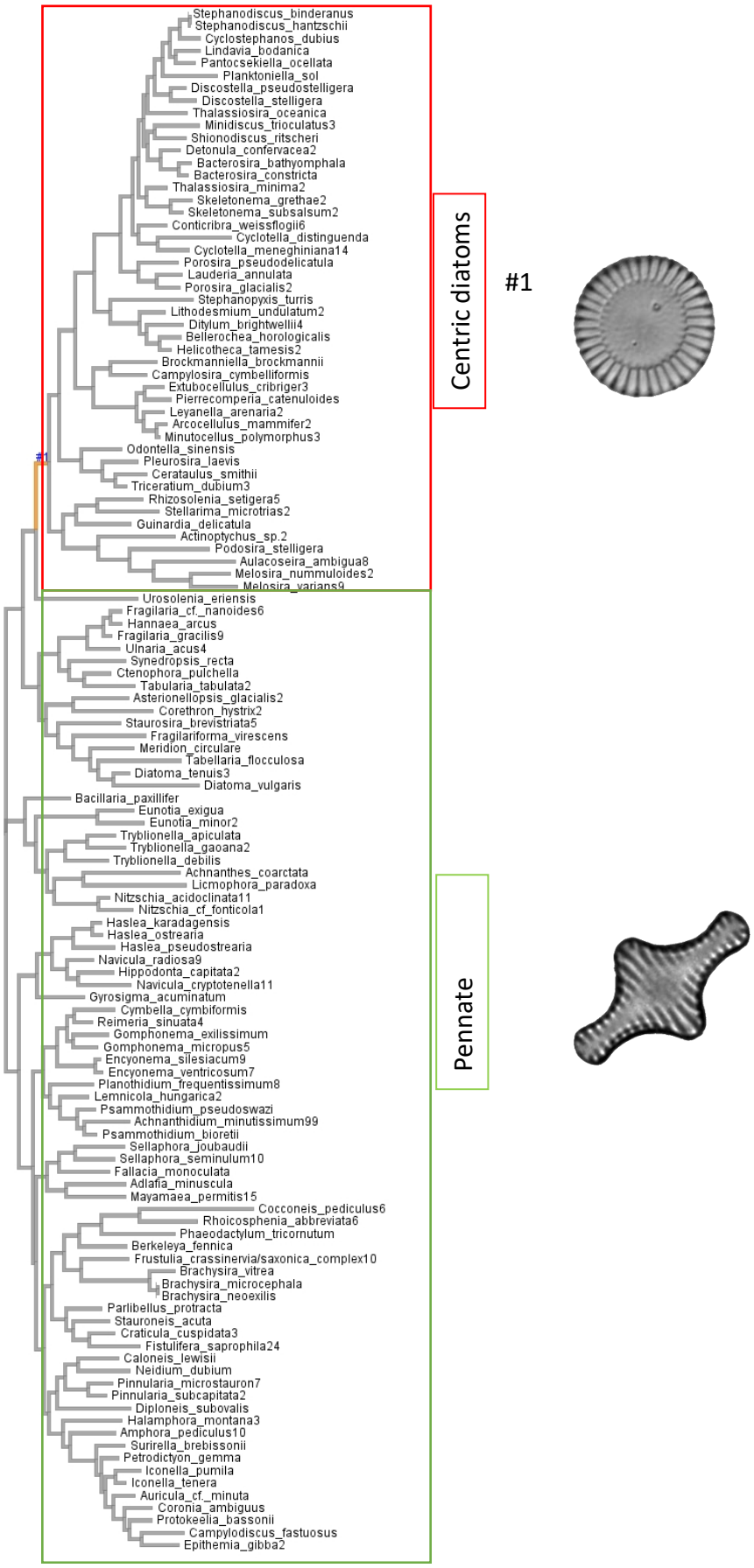


Figure 46 Phylogenetic tree used for the site model comparing centric (orange branched labelled #1) and pennate diatoms.

The likelihood test used to compare the models shows a better fit of the M2a model over M1a and a better fit of the M8 model over the M7 model (Table 12). This is strong evidence of the presence of positive selection driven by the primary structure of the diatoms (centric vs pennate diatom) during the evolution of the *rbcL* gene. This supports my hypothesis that the chloroplast structural differences of these two clades (small multiple chloroplasts for centric diatoms and single up to a few large chloroplasts for pennate diatoms) has been specifically selected according to their specification/function and are not neutral mutations that persist in the genotype. Moreover, the better fit of M8 and M7 over M2a and M1a tend to show that there is a beta distribution of dN/dS classes. The M8-M8a test has non-significant value and it is the more robust test for positive selection as it determines if the dN/dS ratio is significantly greater than 1 instead of just identifying if there is a class of site with dN/dS ratio greater than 1. Therefore, we cannot say that the dN/dS ratio is significantly greater than 1 even though the M7 and M8 model comparison seems to show the presence of positive sites. The relatively small dataset could be the reason why the M8-M8a did not generate significant likelihood-ratio test values as this kind of test is more likely to be significant with large size datasets.

Site model (SM)						Comparison of models		
Model	log-likelihood	Estimates of parameters				Model compared	likelihood ratio test P-value	Positive sites
M3	-30372.5	p:	0.82	0.17	0.01	M0 vs. M3	<0.001*	
		ω:	0.0040	0.33	7.1			
M0	-32351.393816	ω0:	0.055					
M2a	-31473.903427	p:	0.48	0.38	0.14	M1a vs. M2a	<0.001*	
		ω:	0.014	1.0	2.3			
M1a	-30776.737900	p:	0.89	0.12				
		ω:	0.014	1.0				
M8	-30254.264803	p0=0.96920	p=0.14	q=1.1		M7 vs.M8	<0.001*	282 I 0.960*
		(p1= 0.03080)	ω= 1.0					
M7	-30330.073347	p=	0.20	q=0.95				
M8a	-30254.262330	p0=0.96908	p=0.14	q=1.1		M8a vs.M8	0.94	
			(p1= 0.030)	ω= 1.0				

Table 12. Site Model results for centric vs pennate diatoms

The site model comparisons show a strong positive selection signal with significant model fitting differences between: M0 vs. M3, M7 vs.M8 and M8a vs. M8. One site (282), located at the interface of the *RbcL* dimer (Iida et al., 2009), has been shown to be under positive selection by the Bayes Empirical Bayes (BEB) analysis implemented in PAML (Table 12). These results are clear evidence of selective pressure driven by the primary shape of the diatom and confirms my assumption of different adaptation of the *rbcL* gene between those two diatom morphologies.

MARINE VS FRESHWATER

27 sequences from the open dataset DIAT.BARCODE (INRAE- UMR CARTEL Thonon-les-Bains-France) have been selected to create a dataset. The relatively small size of the dataset is due to the fact that only centric diatoms were selected to prevent confounding effects of the positive selection influence of the primary shape of the diatoms (pennate vs centric) on the result. As a matter of fact, the vast majority of saline diatoms are centric.

The dataset as been used to create the phylogenetical tree (Maximum Likelihood with 100 replications, Figure 47). The dataset and the phylogenetical tree have been conjointly used to run the Clade Model in EasyCodeML.

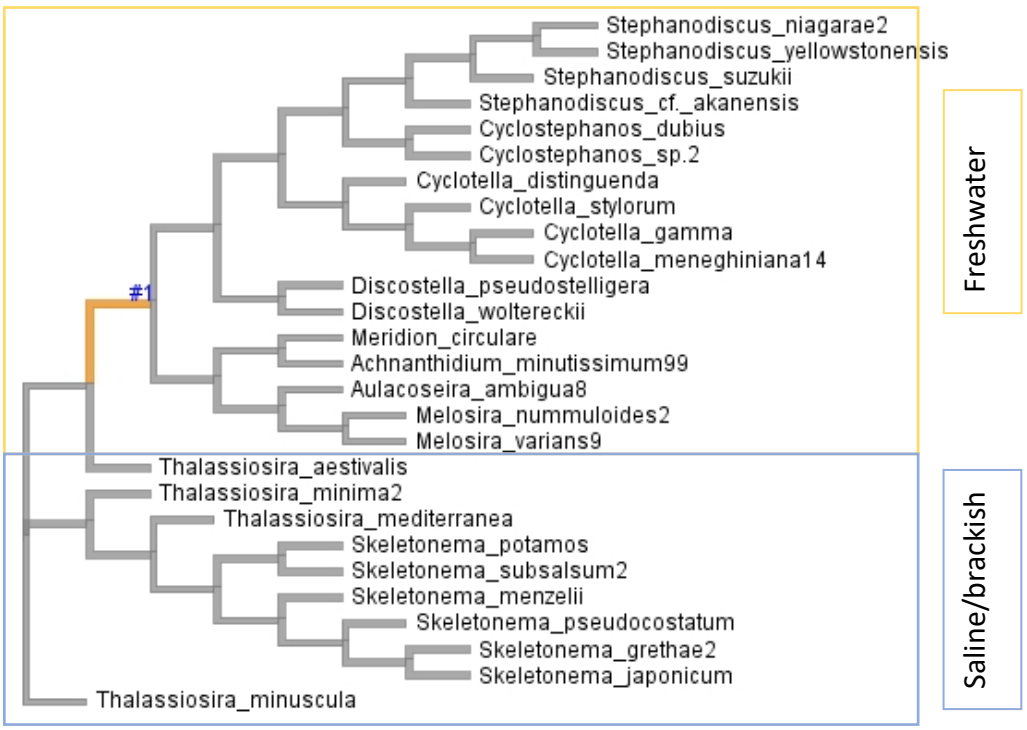


Figure 48 Phylogenetical tree used for the site model between saline and freshwater diatoms (yellow branch labelled #1).

Site model (SM)						Comparison of models		
Model	log-likelihood	Estimates of parameters				Model compared	likelihood ratio test P-value	Positive sites
M3	-6717.3	p: ω:	0.85725 0.00543	0.14275 0.40423	0.00000 39.45386	M0 vs. M3	<0.001*	
M0	-6947.709058	ω0:	0.05292					
M2a	-6758.829158	p: ω:	0.90686 0.01542	0.09314 1.00000	0.00000 93.27602	M1a vs. M2a	0.999	
M1a	-6758.828892	p: ω:	0.90686 0.01542	0.09314 1.00000				
M8	-6727.293208	p0=0.96898	p=0.03423	q=0.23717		M7 vs.M8	<0.001*	34 A 0.515, 254 I 0.808, 262 E 0.702, 284 I 0.933, 353 A 0.518, 362 Y 0.693, 375 K 0.529, 437 A 0.963*, 442 N 0.974*
M7	-6747.267661	(p1= 0.03102) p=	ω= 1.00000 0.03622	q=	0.21430			
M8a	-6707.534978	p0=0.97786 (p1= 0.02214)	p=0.07663 ω= 1.00000	q=1.47137		M8a vs.M8	<0.001*	

Table 13. Site model results for the saline vs freshwater analysis

In this analysis, the M7-M8 and the M8-M8a comparison (Table 2) both shows a significant sign of positive selection. Furthermore, two sites (437,442) have been significantly detected as positive sites and seven other sites (34, 254, 262, 284, 353, 362, 375) have been predicted as potential positive sites with the M7-M8 comparison test. We can confidently say that the dN/dS ratio is following a beta distribution and the ratio is significantly greater than 1 which is a strong proof of significant positive selection occurrence in the *rbcL* gene. These significant tests result give strong support of a different evolutionary signature in the *rbcL* gene between Marine and Freshwater diatoms.

CLADE COMPARISON: PYRENOID STRUCTURE

285 full length sequences from the open dataset DIAT.BARCODE were used to create a diatom phylogenetical tree (Maximum Likelihood with 100 replications).

The clade model C (CmC) on EasyCodeML was generated in order to test if the *rbcL* gene of those three clades has evolved differently. As other clade models, site-specific selective constraint differences are possible among clades in the phylogenetical tree. It started by

estimating a different ω ratio for each clade (three in this experiment) and then the model has been compared against a null model (M2a_rel) that estimates a unique fixed ω ratio among clades (Weadick & Chang, 2012).

The results show a significant difference between the ω ratio of each clade, indicating a different evolutionary signature linked to the pyrenoid structure. The trees generated in the analysis are shown in Figure 48. As we can see the three clades are phylogenetically close, and also share some morphological features such as shape and size, and the pyrenoid structure is one of their greatest differences (Mann, 1989). Due to the involvement of the pyrenoid in the Carbon dioxide fixing, the pyrenoid structure is directly linked to the adaptation of a diatom to the Carbon content of its environment. It is then unsurprising to find that the pyrenoid structure seems to be an evolutionary driver of the *rbcL* gene by applying a selection force of the diatoms that I linked to the Carbon dioxide concentration in the environment.

Model	Number of parameters	log-likelihood	Model compared	Likelihood ratio test P-value
CmC	575	-48812	M2a_rel vs CmC	0.000017332*
M2a_rel	572	-48824		

Table 14 Clade model (CmC) result for the pyrenoid structure analysis. The likelihood-ratio test p value is significant with α=0.05

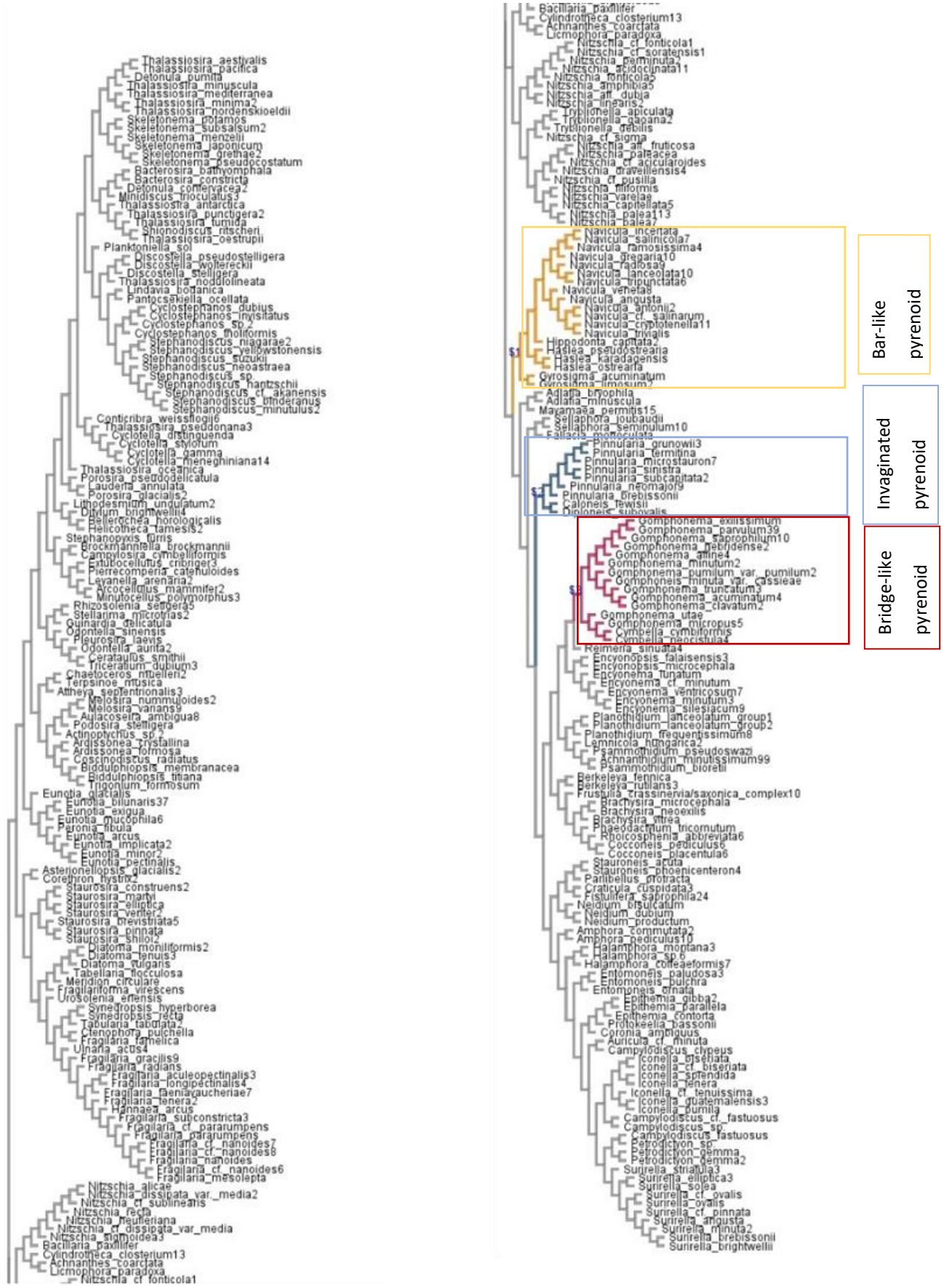


Figure 49 Phylogenetical tree used in the clade model comparison. Yellow: bar -like pyrenoid, Blue: pyrenoids penetrated by tubular invagination, Red: Pyrenoid that forms a bridge

SELECTIVE SITES MAPPING ON THE 3D RBCL PROTEIN

Mapping the positively selected residues on the RuBisCO tertiary structure revealed that they are located in important regions for dimer-dimer, intradimer, large subunit-small subunit and RuBisCO-RuBisCO activase interactions, and that some of the positively selected residues are close to the active site (Figures 50 & 51). Within the positively selected residues I can highlight the one located at position 282, which is as mentioned before the interface of the *RbcL* dimer

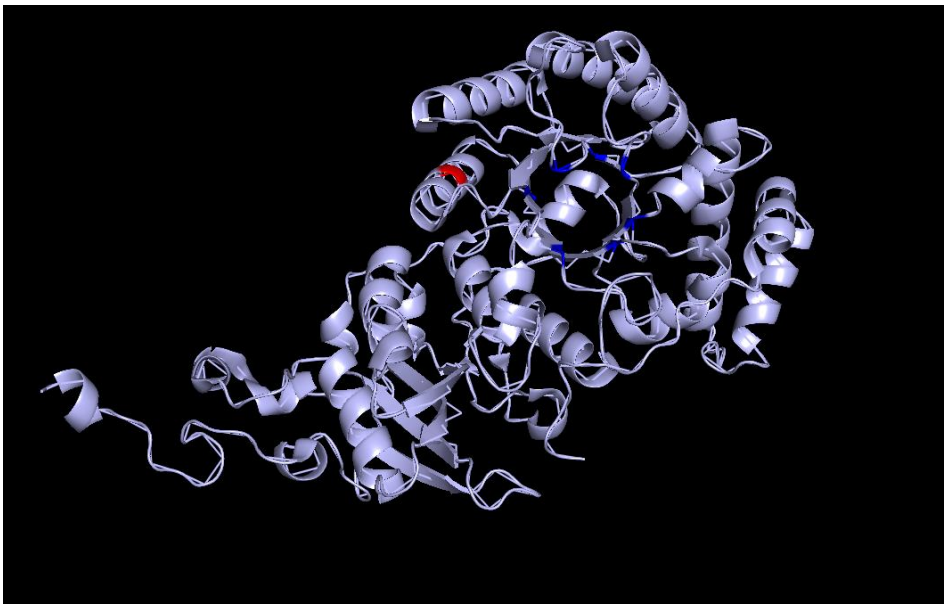


Figure 51 3-dimensions view of a single Rubisco long chain from Thalassiosira antarctica. Blue = Active sites, Red = significant positive sites for centric vs pennate diatom model.

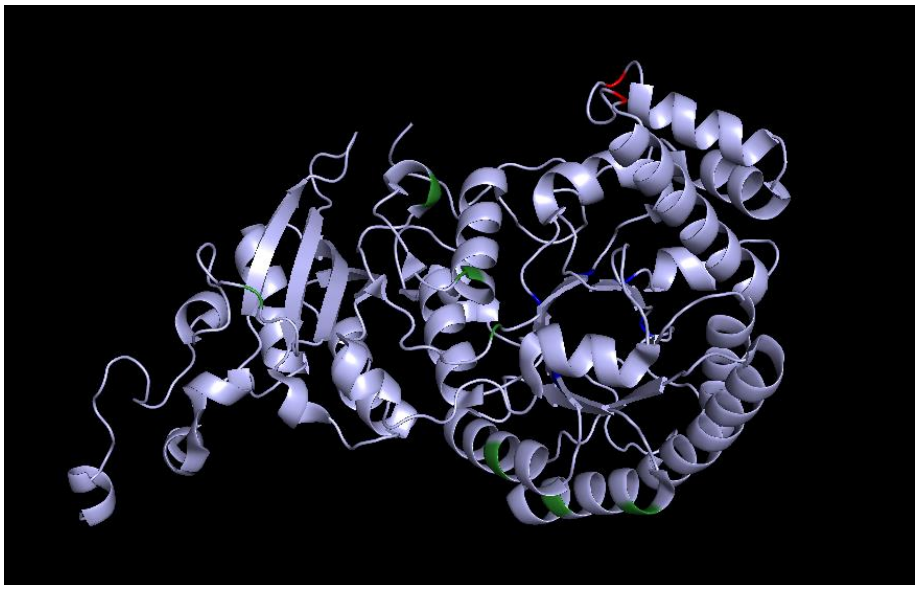


Figure 52 3-dimensions view of a single Rubisco long chain from Thalassiosira antarctica. Blue = Active sites, Red = significant positive sites for saline vs freshwater diatom model. Green= potential sites

(Iida et al., 2009). This gives support to the fact that this residue is positively selected as a mutation at this locus is likely to change the interaction link between the dimer which is a main driver of the overall protein structure.

DISCUSSION

The comparison of the different codon models between the diatom clades confirmed the initial hypothesis that the evolution of the large subunit of RuBisCO has been driven by the structure and the environment of the species during their history. In this case I focused on the characteristics that influence the availability of Carbon dioxide as the RuBisCO protein is directly involved in the first step of Carbon fixation and, therefore, CO_2 concentration is a main limiting factor of the RuBisCO activity.

The model fitting comparison between the different clades enabled us to reveal significant drivers of the evolution of RuBisCO: morphological symmetry (pennate/centric), salinity of the environment and pyrenoid shape. Moreover, the 282 codon (at the interface of the *RbcL* dimer (Iida et al., 2009)) has been detected as a positive site for the morphological symmetry and the 437 and 442 codons are significantly positive sites for the salinity preferences.

The *rbcL* gene from diatoms seems to have evidence of positive selection, as opposed to the low rates present in higher plants (Yao et al., 2019) . Since the CO_2 diffuses through biological membranes, diatoms and other microalgae are facing an additional challenging problem as the diffusion of CO_2 is enhanced by the single-to-few cell organization that increases the exchange surface with the environment (Moroney and Somanchi, 1999). This could explain the greater influence of characteristics related to CO_2 concentration on the evolution of the diatoms *rbcL* gene. Rather than RuBisCO positive selection, higher plants are known to evolve in specialisation of cells to manage separately the different steps of the fixation, such as C_4 plants, which concentrate PEP carboxylase in leaf mesophyll cells and concentrate RuBisCO in bundle-sheath cells (Gao et al., 2014). Unicellular organisms such as microalgae are indeed not able to specialise like multicellular entities and this is coherent with the higher rates of positive selection of the *rbcL* gene of diatoms.

Previous work has shown that the kinetic diversity of diatom form ID RuBisCO is greater than the one from plant form I RuBisCO (Young et al., 2016). Furthermore, the Bacillariophyta clade

present a high diversity in strength of the Carbon concentrating mechanisms (CCM) compared to plant groups. The diatom CCM are also noticeably more efficient (Young et al., 2016). This is also logical with the higher dependency of RuBisCO evolution with diatom while superior plants can rather organize and specialized cells to divide the different steps of carboxylation.

Young et al.,2012 suggested that within the Bacillariophyta clade, positive selection showed a constant occurrence during period of falling Phanerozoic CO_2 and seems to show the development of Carbon-concentrating mechanisms. Consequently, the declines in atmospheric CO_2 fuelled the positive selection in RuBisCO for a sizeable proportion of diatom species.

In the era of Metabarcoding-based ecological assessment, a better understanding of the forces influencing the evolution of *rbcL* should help to create adequate short barcodes within the *rbcL* gene as several sites has been identified as positive sites and can therefore be used to distinguish specific clades confidently. However, a barcode is required to be present in a conserved region to provide robust taxonomic assignment (Deagle et al., 2014). This short barcode located on the *rbcL* gene seems to be a good compromised with few positive sites along a conserved coding region.

An extension to this study would be to analyse the positive selection present in the *rbcS* gene as it codes for the small subunit of the RuBisCO protein and has been shown to be involved in the control of the Pyrenoid structure in the green algae genus *Chlamydomonas* (Meyer et al., 2012). This would require a substantial task to create the dataset as *rbcS* has been less studied that *rbcL* and imply, therefore, a lower availability of sequences for molecular analysis.

In conclusion the analyses show that the three criteria chosen (Diatom primary shape, Salinity preferences and Pyrenoid structure) all reveal proof of positive selection on the *rbcL* gene. This corroborates the insights of the recent studies that consider the *rbcL* gene as a site under positive selection in the diatom genome, this corroborate with Kapralov and Filatov, 2007b and Young et al., 2012. Notwithstanding the relatively evolutionary activeness of this locus, recent studies have showed the effectiveness of barcode in the *rbcL* to distinguish even at species level diatoms during Metabarcoding analysis (Vasselon et al., 2017a). Finally, we can say that my study validates that environment preferences such as salinity, and morphology of both chloroplast and frustule drove the evolution of the *rbcL* gene in the diatom clade.

CHAPTER 7 GENERAL DISCUSSION AND PERSPECTIVES

SUMMARY

This thesis project is part of ongoing research into improvement of diatom Metabarcoding tools with the goal of integrating them in the routine diatom biomonitoring survey for freshwater quality surveillance.

It follows and completes the work initiated by Kelly et al. (2018) and Glover (2019) in the UK, and Zimmermann et al. (2015), Kermarrec (2012) and Vasselon (2018) in the rest of Europe. These different projects set the basis of the method across each methodological step: sampling, nucleic acid extraction, PCR, sequencing, but also bioinformatic processing and ecological index calculation.

In particular, these studies highlighted the choice of the *rbcL* gene, among others, as a suitable barcode region and pushed forward the development of reference libraries based on this barcode region. Furthermore they investigated the suitability of the sequencing technologies available at the time : 454 Pyrosequencing (Kermarrec, 2012; Zimmermann et al., 2015), PGM Ion torrent (Vasselon, 2018) and MiSeq Illumina (Kelly et al., 2018; Vasselon, 2018).

The aim of this thesis was to confirm and expand upon these proposed approaches and to optimize the most limiting points to make Metabarcoding a viable alternative to the LM traditional method.

Firstly, I updated the bioinformatic pipelines with a particular focus on the latest denoising and polishing algorithms that drastically reduce the time and computational power requirements (Chapter 1 for Illumina pipeline and 3 for Illumina and ONT Pipeline). This moved from the current QIIME1-based bioinformatic pipeline to DADA2 (Illumina) and NGSpecies (ONT) based scripts, which simplifies the use of bioinformatic pipelines without compromising on the latest bioinformatic tools used.

A section of the project deals with the benchmark of the reference libraries, especially on the current reference library from Kelly et al. (2018) and the diat.barcode European reference library (Chapter 4). The addition of non-diatom sequences improved the taxonomic assignment result in terms of reducing unassigned sequences and increasing the reliability of

the detection (Chapter 2 and 3). In this study, the Illumina MiSeq, which is the sequencing platform of choice of the last decade, was directly compared to the more recent lower cost and more portable MinION device from ONT in order to test this technology (Chapter 3). The *rbcL* barcodes reliability for diatom biomonitoring survey was also analysed by comparing different length of rbcL barcode sequences with the two different technologies: short barcode for Miseq and MinION and full length *rbcL* with MinION (Chapter 3). Moreover, evolution of the *rbcL* gene was studied to identify the main drivers of its evolution, and especially identify the different positive selection event to confirm that the *rbcL* gene is a rather conserved region as it is a requirement for a good barcode region (Chapter 6).

This study has the interest to integrate comparisons and experimentations that were operated in a variety of environments, samples from the UK and from France (Chapter 1), Yorkshire rivers and mesocosm runnels and even in vitro mock communities. This enabled a better understanding of the limitation of Metabarcoding and LM bioindication.

The following discussion aims to summarize the findings of this thesis and to integrate these results in the current context of diatom biomonitoring for water quality surveillance.

OPTIMISATION OF THE METABARCODING METHOD

SAMPLING

The absence of natural stone in the mesocosm motivated the test of the tile-based sampling approach, adapted from Kelly et al. (1998). My experiments clearly show its efficiency in different environments, from rivers to mesocosms (Chapter 4), and with an interesting addition of standardization by using the same material and size. The communities identified from this method are similar to the ones originating from cobble-based sampling (Yorkshire river in Chapter 4). It appears to be an improved method for benthic biofilm sampling.

Nevertheless, tile-based sampling can be more difficult to implement as it requires preliminary work to place out the tiles and then collect them. There is also a possibility of the tiles being lost, either as they become covered with mud, taken away by the stream or moved by passers-by. A large-scale experiment would clarify the ratio of collected to lost tiles and better quantify the feasibility of this method for river survey. For mesocosm and other artificial shallow water bodies it appears to be the most convenient method.

BARCODE / PCR

The use of both *rbcL* short-barcode (Chapter 3) has been shown to be reliable and efficient for TDI calculation. The UK short-barcode (Kelly et al., 2018) has been successfully used with both Illumina and ONT technology with a noticeable suitability for taxonomic assignment for biomonitoring. The phylogenetic study of the *rbcL* gene (Chapter 6) highlights the link between the evolution of the *rbcL* gene and the particular structure and ecological preferences of diatom groups. While these evolution patterns have been highlighted, the overall gene region is conserved enough to be a suitable barcode and enable a robust identification of taxa and distinction at species and genus level.

As the suitability interchangeability of diat.barcode primers and the current UK *rbcL* barcode primers has been proven for at least UK samples, this study confirms the possibility to use the diat.barcode reference library as standard because the generated EQR (TDI ecological classes) are significantly the same and the diat.barcode is a more curated and well reference database (see Chapter 3). Another main benefit would be the better flexibility in terms of sample location as the diat.barcode was built with diatom and microalgae from all over the world, enabling comparison of results from different continents.

Although my experiment did not find any benefit to use the longer and full length *rbcL* barcode we can only conclude on the unsuitability of this particular set of primers for biomonitoring (Chapter 4). Nevertheless, the issues occurred clearly before the sequencing steps, and very likely during the PCR step, which lead to the low sequencing depth. Longer amplicons still have the potential to improve the taxonomic resolution but my experiment of the full *rbcL* barcode from Glover (2019) method suggests that an amplification efficiency is a must, in both specificity to the targeted group and quantity of DNA amplified. I conclude that the use of this full length *rbcL* barcode method is suitable for pure culture that needs to be referenced in a taxonomic library but not for Metabarcoding studies. There is a need for comparison of long *rbcL* barcodes for diatom survey, for example the 748 bp from Hamsher et al. (2011) that could be coupled with long read sequencing from ONT and has the potential to generate more accurate taxonomic assignments without using the MiSeq platform.

SEQUENCING PLATFORM ONT SEQUENCING

This project is the first to investigate the potential of the MinION platform for diatom biomonitoring using the *rbcL* short (and potentially long) barcodes with a direct comparison with the Illumina technology (Chapter 4). The need for such an experiment was evident as attention on this new portable and lower cost sequencing device has recently risen (Krehenwinkel et al., 2019). This device and technology has the potential to create a shift in Metabarcoding studies by enabling a more accessible and affordable method for sequencing. MinION device price (~1000$) is much more affordable/accessible than MiSeq device (~30000$) which can enable more laboratories to access HTS and therefore to integrate Metabarcoding in their standardized survey method.

The study demonstrated the usability of the MinION platform (Mock Community Chapter 5) and I removed the biases of the use of different primers that could be difficult to interpret.

ONT technology offers a variety of platforms and the MinION experiment opens up the possibility of using other ONT platforms that run multiple flowcells simultaneously (five flowcells with the GridION and 48 with the PromethION; (https://nanoporetech.com/products) to upscale delivery, or even smaller flowcells (Flongle) for lower cost, smaller surveys that require less sequencing dept. My bioinformatic pipeline is usable for experiments of virtually any size and the low computational power requirement makes it manageable to any research group.

BIOINFORMATIC PIPELINE

The greatest improvement in the diatom Metabarcoding method is the new adapted bioinformatic pipelines for both MiSeq and MinION data, which generated fit for purpose results while being run on a traditional research laptop for less than a day (Chapters 3, 4, 5). This contrasts to the previous commonly used MiSeq bioinformatic pipeline (Kelly et al., 2018) which was based on QIIME1 that took several days to run on servers yet produced less suitable results (i.e. it produced results less similar to the LM method in both community structure and TDI values; Chapter 3). This confirms the necessity of an efficient and reliable bioinformatic pipeline for every Metabarcoding study because of the substantial amount of data these generate and the accuracy needed for reliable taxonomic assignments.

A high variability of river ecological conditions were present in the river and mesocosm sampling locations in this project (Chapter 4), confirming the robustness of these newly adapted bioinformatic pipelines (based on DADA2 for MiSeq and NGSpecies for ONT) for routine water quality surveys.

Read Denoiser / Polisher

The optimisation generated by denoiser and polisher deal successfully with the problem caused by the sizeable number of reads generated by each sequencing run, as well as the low, but present, sequencing errors (deletions, insertions and substitutions) that are difficult to distinguish from real genetic variation.

Increasingly, bioinformatic pipelines are moving from traditional OTU clustering to denoising/polisher algorithms (e.g. in Barnes et al., 2020; Liu et al., 2023) and the experiments done here further confirmed the adequacy of these new tools to, firstly, generate reads useable for diatom biomonitoring and secondly, reduce the time and computational power required to process the data of a whole sequencing run (Chapter 3).

Chimera filtering

The use of DADA2 coupled with an efficient chimera removal step successfully processed the MiSeq sequencing outputs from all the samples, regardless of the river and mesocosm locality and the *rbcL* barcode used (barcode from Kelly et. al 2018 or diat.barcode; Chapter 3).

While chimeras represent only a minor proportion of the reads (<2%) and, therefore, cannot affect the TDI calculation in a major way, the community evenness index was clearly more similar to the LM data with the use of a pipeline that integrates a chimeral removal step. Therefore, the presence of chimeras artificially inflates the calculated diversity by generating sequences that have no biological origin and are artifacts from PCR barcode amplification (Chapter 3). The integration of chimera filtering in the routine bioinformatic pipeline has the benefit of producing a better estimation of the community structure without adding significant complexity to the bioinformatic process.

Naive Bayesian classifier

This study is one of the first to integrate Naive Bayesian taxonomic assignment (Wang et al., 2007) to a ONT bioinformatic pipeline. The integration of the Naive Bayesian classifier in both ONT and Illumina bioinformatic pipelines was successful and increased the taxonomic

resolution as well as having the advantage of locating the unassigned reads to higher taxonomic ranks (Chapter 3 & 5), which is useful to discard obvious contaminants (e.g. both potatoes and banana sequences were commonly found in samples; Chapters 3 & 4) or to integrate them in the index calculator if a genus-level ecological preference is present (as is the case for some Diatom genera; diat.barcode Rimet 2019).

A drawback of the Naive Bayesian classifier is the need to have the taxonomic hierarchy associated with each read in the reference library (Callahan et al., 2016; Wang et al., 2007). This was done by adapting the Kelly 2018 reference library, which was time consuming but manageable (Chapter 4).

The integration of Naive Bayesian taxonomic assignment instead of the traditional BLAST method in the majority of the new Metabarcoding bioinformatic pipeline seems a logical evolution due to the obvious advantages. Notwithstanding the importance of the taxonomic assignment method, it relies on the presence of a reliable reference library to reveal its potential. Moreover, the bioinformatic tools are not yet perfectly adapted to Naive Bayesian classifiers, hence the use of Mothur and the galaxy server to handle the multithread Wang assignment step in my ONT bioinformatic pipeline.

Taxonomy Reference library

During this study I corrected the reference library from Kelly et al. (2018), which had a 14 bp deletion across all reference sequences, affecting results that used this library. Other stakeholders (e.g. EA) were informed as providing them with the corrected library. This 14bp represented more than 4% of the total barcode region, creating a significant issue for the taxonomic assignment, especially with the use of BLAST with 95% similarity threshold combined with OUT clustering at 97% similarity threshold.

This project is the first to compare the corrected reference library with the diat.barcode reference library (Rimet 2019f) and to show their interchangeability (Chapter 3 & 4). Moreover, the addition of non-diatom microalgae to the reference library has been proven to improve the taxonomic resolution (Chapter 4). This shows it is crucial to have access to a taxonomic reference made of diverse taxonomic groups with an important number of sequences and taxa, especially in the context of the increasing use of Naive Bayesian classifiers that rely on this type of data more than the BLAST assignment.

INTERCHANGEABILITY OF LM AND METABARCODING

This study enabled a large-scale comparison of the Metabarcoding latest updates to the traditional LM by executing the two approaches simultaneously on the same samples. This was done particularly in Chapter 1 and 2.

RELATIVE ABUNDANCE

There is a clear correlation between the number of *rbcL* copies, which are directly linked to the number of chloroplasts per cell, and the read abundance in Metabarcoding generated communities (Chapter 5). The mock community experiments (Chapter 5) show that centric diatoms are over represented, as is expected from their natural high number of chloroplasts and I only corrected the number of each diatom added in each community with the biovolume average of each species (close to the biovolume-based correction factor used in Vasselon et al., 2018). A correction factor based on the average number of chloroplasts of each organism seems like a better choice and would be an interesting tool for the future of diatom biomonitoring using Metabarcoding, as ecological indices (e.g. TDI, IPS and IBD) rely on relative abundance of each taxon. In the Chapter 5, uncommon taxa were more abundant when the LM method was used and confirm the possibility of using LM to deal with the rare ambiguous samples that contains uncommon taxa.

The mock communities experiment (Chapter 5) was the first to use non-diatoms in the assemblage, and to use living organisms instead of DNA extracts. This is closer to natural conditions, as it integrates the different accessibility of DNA of each organism as well as the interaction of the different groups. The mock community approach permitted to create an unprecedented experiment with direct comparison of the community that I composed and the associated HTS community. I did not compare two measurements (LM counts vs HTS reads) or index calculation, instead focussed on the Metabarcoding method independently from the LM method and TDI calculation biases.

UPDATE OF TROPHIC DIATOM INDEX : NGS TDI5

The new version of TDI adapted for NGS data (TDI5) was tested in a limited number of the experiments (see Chapter 4). From the results of the small-scale experiment, the new TDI5 version performed worse compared to the previous TDI4 version. However, during the update

of the TDI5 (Kelly et al., 2020), the authors identified that there was little improvement in terms of LM correlation, but that the recalibrated TDI5 version gave a better linear fit than the original TDI4 with Metabarcoding data where the relationship was clearly curved. The larger data used in Kelly et al. 2020 to analyse the recalibrated TDI5 provides higher confidence in their conclusions rather than my more limited experiments, but further assessment of this may be necessary.

COMMUNITY STRUCTURE ANALYSIS : MANTEL TEST, SPECIES EVENNESS, NMDS

Diatom Metabarcoding methods use thousands of DNA sequence reads instead of hundreds of visual counts from the LM method, which tends to overestimate the diversity of the surveyed site compared to the LM method (Bailet et al., 2020). Aspects of this may be due to artifacts introduced by the sequencing methods. The data in Chapter 3 proved that adding the chimera filtering step to the bioinformatic pipeline reduces some of the overestimation of diversity. In particular, the measured species evenness became similar between LM and Metabarcoding communities (Chapter 3). Similarly, the denoiser/polisher algorithms, by correcting each read to ASV, also decrease the number of singletons ASVs and the overestimation of the diversity.

Similarities of structure between LM and Metabarcoding community have been highlighted during this study with the use of Mantel tests (Chapters 3 and 4). Moreover, the NMDS analysis of the communities from the Yorkshire rivers and the mesocosm sites (Chapter 4) drew the same conclusion in terms of clustering regardless of the method used.

This thesis clearly shows an interchangeability between LM and Metabarcoding for community structure analysis, and this has been possible thanks to the new bioinformatic tools.

CONVENIENCE OF USE

Light Microscopy identification is specialist, time-consuming and relies on the analysis of one sample at a time, Metabarcoding enables the simultaneous analysis of large numbers of samples, limited by the number of indexing tags in the sequencing step and by the number of reads generated per sequencing run. The time requirement is also reduced (Chapter 3) and does not need extended diatom identification training. As such it confirms that the

Metabarcoding method, when optimized, is more accessible and convenient to environment managers than diatom LM identification. Nevertheless, the price per sample is only lower with Metabarcoding when at least tens of samples are analysed together. However, the price of both sequencing run reagent and instrument has decreased year after year while the number of reads per run increased (Stefan et al., 2022; Stevens et al., 2023). At the same time LM cost is more constant as the method and the instrument are not being improved significantly.

FUTURE PERSPECTIVES

There is an active community researching the implementation and optimisation of Metabarcoding in ecological surveys. New ideas are emerging and the optimisation shown in my thesis project can decrease the technical limitations and enable the experimentation of this new idea for more complete, accurate and manageable water quality assessments. Below I discuss further options to improve or better utilise phytoplankton Metabarcoding data.

WHOLE PHYTOPLANKTON COMMUNITY BIOMONITORING

Metabarcoding is free from some of limitations of LM, and in particular has the potential to target any organism in the community as long as the DNA is available and a reference barcode sequence is available for it, whereas LM is limited to observed living organism or residue that persists (such as frustule or Chrysophycean cysts). Hence the new trend to use the whole phytoplankton community with Metabarcoding in order to create more complete ecological assessments that are not limited to diatoms (Hering et al., 2018; Huo et al., 2020). This creates new challenges that were investigated primarily in the mock communities experiment (Chapter 3) but also in the reference library update (Chapter 2) by quantifying the part of the non-diatom phytoplankton targeted by the *rbcL* short-barcode.

Moving away from diatom-only surveys needs the integration of non-diatom phytoplankton taxa ecological preferences in the Trophic Indices based on the whole phytoplankton community. This may be time consuming but numerous of these common organisms have been studied for centuries and their ecological preferences are well documented (CEMAGREF, 1982; Descy and Coste, 1991; Kelly and Whitton, 1995).

RNA METABARCODING

While eDNA Metabarcoding is now widely used for biomonitoring and other species detection surveys (Rishan et al., 2023; Schenekar, 2023), studies into eRNA Metabarcoding have started to get attention (e.g. Veilleux et al., 2021). The use of RNA is thought to be limited, mainly due to its higher degradation in the environment linked to its single-stranded structure (Kagzi et al., 2022). However, targeting eRNA rather than eDNA could avoid some pitfalls, for example by avoiding upstream contamination as it was hypothesized a eRNA higher degradation rate (Pochon et al., 2017; Yates et al., 2021) may prevent nucleic material from dispersing too far from its place of synthesis.

River algae and arthropods eRNA Metabarcoding has been tested against eDNA Metabarcoding for ecological surveys and water in Miyata et al., 2022, which targets river algae and arthropods. This shows high potential with a very low false positive ratio but also presents the limitations such as the low sensitivity. There is room for improvement but this study is a very robust proof of concept. Diatom Metabarcoding can directly benefit from this approach with the updates in sampling, PCR barcode, Sequencing platform and bioinformatic that I benchmarked.

As RNA is a marker of the expression of a gene, when the barcode is the *rbcL* gene it can be an estimator of the primary production of each organism. As such it could enable us to measure the part each organism plays in the overall primary production of an ecosystem. The use of RNA may also exclude dead organisms and the dormant cells that are present but not active in the ecosystem. If eDNA Metabarcoding is an indicator of the presence of an organism, then eRNA can be an indicator of which organism is active and how much is active (Pochon et al., 2017). It could hypothetically reveal e.g. that major taxa are not the most primary producers.

TAXONOMY-FREE APPROACH

A taxonomy-free approach aims to classify OTUs/ASVs in environmental samples without relying on predefined taxonomic databases. It presents an interesting alternative that does not require the extensive work of investigating the ecological optimum of each phytoplankton taxa (Descy and Coste, 1990; Kelly, 1998), but instead could use machine learning to train on sequencing output from previous environmental and sequencing records. This would ensure

the continuity of the work of Feio et al., 2020 which is a diatom only taxonomy free approach, and we can imagine to extend this approach to all the phytoplankton community. The use of a less diatom-specific PCR primer for the *rbcL* barcode such as the UK barcode (Kelly et al., 2018) instead of the diat.barcode primers (Rimet et al., 2019) would be interesting and can use the previous sequencing run from the UK river water quality survey. The limitation of this technique is the difficulty to correct, identify and discard artifacts or contaminant, from other samples but also from any other possible sources, because of the lack of identification step.

MACHINE LEARNING

The current ecological indices such as the TDI have been designed to only focus on nutrient preferences (Nitrogen and Phosphorus) because of the extensive preliminary work required to analyse each taxon and understand its nutrient concentration preferences. Therefore, it would be difficult to integrate several pollution and environmental characteristics inside an easy to interpret metric such as the TDI and the IPS. Nevertheless, new machine learning technology coupled with metadata on water chemistry and microalgae communities from numerous sites could be used to create new and more informative indices.

Machine learning could also be used to improve the LM method by creating more advanced automatic identification tools coupled with flow cytometry, already studied with algae and the use of the Artificial neural network model in Balfoort et al., 1992 . These new machine learning tools and the technical progress in computational power have the potential to follow this experimentation and conceive a bioinformatic tool for biomonitoring. This would form an interesting combination of tools with the Metabarcoding.

GENERAL CONCLUSION

This PhD thesis provides important optimisation of the Diatom biomonitoring survey with the use of Metabarcoding. Almost every aspect of the method, from sampling to ecological assessment, has been tested and most of the time improved. The project outputs can directly be used and enable better results and facilitate the process, making the Diatom Metabarcoding more accessible. Moreover, the adaptability and versatility of the method, using different primers for the *rbcL* barcode and sequencing platforms, open the possibilities of moving away from a diatom-only biomonitoring method and to integrate other microalgal taxa into the ecological assessment to provide a better understanding of the variables conditions of the water bodies.

The use of Metabarcoding is unfortunately often associated with the end of the morphological identification approach, which is not the case. Metabarcoding, due to its capability for rapid assessment of several samples can facilitate the routine biomonitoring process, but the LM approach is an interesting approach to handle the rare, unusual sites, and should be used jointly to Metabarcoding to have the best of both worlds. It seems logical to imagine a water quality assessment team composed of molecular ecologists, bioinformaticians and diatom identification experts.

In conclusion, this PhD thesis project proved the adequacy of using Metabarcoding for diatom biomonitoring routine survey. The optimisation of the Method was able to confirm the complementarity of this method with the traditional Light Microscopy approach.

APPENDIX

CHAPTER 3 DADA2 R SCRIPT

```
##### Packages activation #####
library("DADA2")
library("stringr")
##### Setting of the fastq path directory #####
path<-"FASTQ_DIRECTORY"

list.files(path)


# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq.gz and SAMPLENAME_R2_001.fastq.gz


fnFs <- sort(list.files(path, pattern="_R1_001.fastq.gz", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq.gz", full.names = TRUE))

##Alternative formats

fnFs <- sort(list.files(path, pattern="_L001_R1.fastq.gz", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_L001_R2.fastq.gz", full.names = TRUE))

fnFs <- sort(list.files(path, pattern="_L001_R1_001.fastq.gz", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_L001_R2_001.fastq.gz", full.names = TRUE))

fnFs <- sort(list.files(path, pattern=".R1.fastq.gz", full.names = TRUE))
fnRs <- sort(list.files(path, pattern=".R2.fastq.gz", full.names = TRUE))


# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)


##### Plot quality #####

plotQualityProfile(fnRs[1:8])
plotQualityProfile(fnFs[1:8])

#this plot may be useful to check the quality of your reads, the quality of the reverse read usually decline sooner than the forward quality

# Place filtered files in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))

##### Trimming & Filtering #####
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,200),trimLeft =c(21,27),  ##c(27,22) for European(diat.barcode) Primers and c(21,27) for UK primers
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE, #These argument values should work for most MiSeq Runs but can be changed , for more info check the DADA2 website
```

```
compress=TRUE, multithread=16, verbose=TRUE) # On Windows set multithread=FALSE

#ratio of filtered seq
out<-as.data.frame(out)

out$ratio<-(out[,2]/out[,1])*100

head(out)
mean(out$ratio)

##### Errors rate learning #####

errF <- learnErrors(filtFs, multithread= T, verbose=TRUE, randomize = TRUE)

errR <- learnErrors(filtRs, multithread=F,  verbose=TRUE, randomize = TRUE)

plotErrors(errF, nominalQ=TRUE)

plotErrors(errR, nominalQ=TRUE)

##### Dereplication #####
derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)

# Name the derep-class objects by the sample names
names(derepFs) <- sample.names
names(derepRs) <- sample.names

##### Denoising using the DADA2 algorythm #####
dadaFs <- dada(derepFs, err=errF, multithread=F,verbose=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=FALSE,verbose = TRUE)

dadaFs[[1]]

##### Merging paired reads  ####
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE)
# Inspect the merger data.frame from the first sample
head(mergers[[1]])


seqtab <- makeSequenceTable(mergers) # ASV table
dim(seqtab)

# Inspect distribution of sequence lengths
table(nchar(getSequences(seqtab)))

##### Chimera removal #####

seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=F, verbose=TRUE) # If not using windows Multithreading could be used but might crash

dim(seqtab.nochim)

sum(seqtab.nochim)/sum(seqtab)

write.csv(seqtab.nochim,file=str_glue("ASV_table_nochime_{basename(path)}.csv"))# export ASV table cleaned of chimera
```

```
#write.csv(seqtab.nochim,file="ASV_table_nochim_INRA_Run2.csv")# export ASV table cleaned of chimera
#write.csv(seqtab.nochim,file="ASV_table_nochim_B2B6R_check.csv") # export ASV table cleaned of chimera
#write.csv(seqtab.nochim,file="ASV_table_nochim_INRA_both_runs.csv") # export ASV table cleaned of chimera
#write.csv(seqtab.nochim,file="ASV_table_nochime_RiverMeso.csv")

###### Tracking file #####

getN <- function(x) sum(getUniques(x))
track <- cbind(out[1:2], sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.nochim))
# If processing a single sample, remove the sapply calls: e.g., replace sapply(dadaFs, getN) with getN(dadaFs)
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
rownames(track) <- sample.names
head(track)
write.csv2(track,file=str_glue("TrackFile_{basename(path)}.csv"))

##### Taxonomic assignment #####

## There we use 3 different classifiers : One created with the UK GoldStandard (corrected), one from diat.barcode (rsyst) and one custom one based on diat.barcode with the addition of non-diatom taxa.

#taxa_GS_corrected <- assignTaxonomy(seqtab.nochim, "GoldStandard_UK_diatoms_rbcL_DADA2.fasta", taxLevels = c("Class","Genus", "Species","ID","clone"),outputBootstraps = FALSE, verbose = TRUE, multithread=FALSE, minBoot=60)
#write.csv(taxa_GS_corrected ,file=str_glue("taxa_{basename(path)}_GS_corrected_2019.csv"))



taxa_diatbarcode <- assignTaxonomy(seqtab.nochim, "Rsyst__1401seqs_312bp_taxonomy_CLASSIFIER_DADA2.fasta", taxLevels = c("Domain", "Kingdom","infraKingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species","Clone"), multithread=F,outputBootstraps = FALSE, verbose = TRUE, minBoot=60)
write.csv(taxa_diatbarcode ,file=str_glue("taxa_{basename(path)}_diatbarcode.csv"))



taxa_custom <- assignTaxonomy(seqtab.nochim,"Diat.barcode_CLASSIFIER_DADA2_08_03_2022_version.fasta", taxLevels = c("Domain", "Kingdom","infraKingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species","Clone"),outputBootstraps = FALSE, verbose = TRUE, multithread=F, minBoot=60)
write.csv(taxa_custom2021 ,file=str_glue("taxa_{basename(path)}_custom.csv"))


taxa.print <- taxa_XXXXX # Removing sequence rownames for display only, change taxa_XXXX by the name of the taxa file you want to display.

rownames(taxa.print) <- NULL
head(taxa.print)

write.csv(taxa.print,file="taxa_{basename(path)}.csv")


##alternatively : Without chimera removing, facultative step in order to see the effect of chimera removal on the assignment
```

```
#taxa1 <- assignTaxonomy(seqtab, "GoldStandard_UK_diatoms_rbcL_DADA2.fasta", taxLevels = c("Domain", "Kingdom","infraKingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species","Clone"), multithread=16)

#taxa.print1 <- taxa1 # Removing sequence rownames for display only
#rownames(taxa.print1) <- NULL
#head(taxa.print1)

##### Graphical representations using Phyloseq #####

library(phyloseq); packageVersion("phyloseq")
library(ggplot2); packageVersion("ggplot2")


##### Phyloseq object creation #####

#sample file made of sample names, facultative for most phyloseq object but needed for Krona plot later
s_data<-data.frame(row.names=sample.names,SampleID=sample.names,Pool="1")

##choose one of them
taxtable<-taxa_GS_corrected_dash
taxtable<-taxa_diatbarcode
taxtable<-taxa_custom
basename(taxa_custom)

ps <- phyloseq(sample_data(s_data),otu_table(seqtab.nochim, taxa_are_rows=FALSE),tax_table(taxa_custom2022)) #change taxa with the taxonomic assignment you prefer

ps

#rarefying step (facultative)
ps.rare<-rarefy_even_depth(ps, sample.size = 6000)
write.csv(tax_table(ps.rare),str_glue("{basename(path)}fragi_rare_taxa.csv"))

# Extract abundance matrix from the phyloseq object
OTU1 = as(otu_table(ps.rare), "matrix")
# transpose if necessary
if(taxa_are_rows(ps.rare)){OTU1 <- t(OTU1)}
# Coerce to data.frame
OTUdf = as.data.frame(OTU1)

write.csv(OTU1,file=str_glue("{basename(path)}OTU1.csv"))
write.csv(OTUdf,file=str_glue("{basename(path)}OTUdf.csv"))



# Transform data to proportions as appropriate for Bray-Curtis distances
ps.prop <- transform_sample_counts(ps, function(otu) otu/sum(otu))
ord.nmds.bray <- ordinate(ps.prop, method="NMDS", distance="bray")
plot_ordination(ps.prop, ord.nmds.bray, title="Bray NMDS")

top50 <- names(sort(taxa_sums(ps), decreasing=TRUE))[1:50]

allnames <- names(sort(taxa_sums(ps), decreasing=TRUE))
sort(taxa_sums(ps))

ps.top50 <- transform_sample_counts(ps, function(OTU) OTU/sum(OTU))
ps.top50 <- prune_taxa(top50, ps.top50)
```

```
ps.all <- transform_sample_counts(ps, function(OTU) OTU/sum(OTU))
ps.all <- prune_taxa(allnames, ps.all)



## Assigned taxa : greyscale, Unassigned taxa : red
plot_bar(ps.all,fill="Species")+scale_fill_grey()+scale_color_grey()+
  geom_bar(aes(color= Species,fill=Species), stat="identity", position="stack") +
  theme(legend.position ="bottom")## Species


ggsave(str_glue("{basename(path)}_Species_custom2022.jpeg"),width=60,height=60,units="cm") #export jpeg
ggsave(str_glue("{basename(path)}_Species_custom2022.pdf"),width=60,height=60,units="cm") #export pdf



plot_bar(ps.all,fill="Genus")+scale_fill_grey()+scale_color_grey()+
  geom_bar(aes(color= Genus,fill=Genus), stat="identity", position="stack") +
  theme(legend.position ="bottom")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))## Genus


ggsave(str_glue("{basename(path)}_Genus_custom2022.jpeg"),width=60,height=60,units="cm") #export jpeg
ggsave(str_glue("{basename(path)}_Genus_custom2022.pdf"),width=60,height=60,units="cm") #export pdf


plot_bar(ps.all,fill="Class")+scale_fill_grey()+scale_color_grey()+
  geom_bar(aes(color= Class,fill=Class), stat="identity", position="stack") +
  theme(legend.position ="bottom")## Class

ggsave(str_glue("{basename(path)}_Genus_custom2022.jpeg"),width=60,height=60,units="cm") #export jpeg
ggsave(str_glue("{basename(path)}_Genus_custom2022.pdf"),width=60,height=60,units="cm") #export pdf


##alternative palette
plot_bar(ps.all,fill = "Genus")+scale_fill_manual(values=getPalette(XXXXXXXX)) ## change the get_palette() value to the number of genera in your data
## you can change "Genus" for "Species" or "Phylum", or any other taxonomic rank present in your taxa file.

plot_bar(ps.top50,fill = "Genus")+scale_fill_manual(values=getPalette(40))


#####Species richness

phylorichness<-plot_richness(ps, measures=c("Shannon", "Simpson"),shape = "bar")
write.csv(phylorichness$data,str_glue("{basename(path)}_richness.jpeg"))


###### KRONA graphs #####

library("psadd")

plot_krona(ps.all,output=str_glue("krona_{basename(path)}_taxa_custom2022_pool"),variable="Pool")


plot_krona(ps.all,output=str_glue("krona_{basename(path)}_taxacustom2022_all"),variable="SampleID")
```

```
plot_bar(ps.all,fill="Genus")+scale_fill_manual(values = palette_diatom)+scale_color_manual(values = palette_diatom)+
  geom_bar(aes(color= Genus,fill=Genus), stat="identity", position="stack") +
  theme(legend.position ="bottom")## Genus
```

QIIME2 PIPELINE USING DADA2 : DIATOM-IZER

######## Sequences must be stored in a file called "A" inside the working directory; the metadata is called "meta-diatoms.tsv" It could #####
######## just made of a column with the name of each sequence.#####

QIIME2 VERSION 2018.8
source activate QIIME2-2018.8

TRAIN THE CLASSIFIER
QIIME tools import \
  --type 'FeatureData[Sequence]' \
  --input-path Rsyst_230218_align_1401seqs_312bp.fasta \
  --output-path ref-seqs.qza

QIIME tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path Rsyst_230218_align_1401seqs_312bp_vf.txt \
  --output-path ref-taxonomy.qza

QIIME feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs.qza \
  --i-reference-taxonomy ref-taxonomy.qza \
  --o-classifier diatoms_classifier_classic.qza

##Custom

QIIME tools import \
  --type 'FeatureData[Sequence]' \
  --input-path custom_QIIME2.fasta \
  --output-path ref-seqs-custom.qza

QIIME tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path custom_june.txt \
  --output-path ref-taxonomy-custom.qza

QIIME feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs-custom.qza \
  --i-reference-taxonomy ref-taxonomy-custom.qza \
  --o-classifier diatoms-classifier-custom.qza

##Gold Standard

QIIME tools import \
  --type 'FeatureData[Sequence]' \
  --input-path GOLD_standard_UK_mothur.fasta \
  --output-path ref-seqs-goldstandard.qza

QIIME tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path GOLD_standard_UK_mothur.txt \
  --output-path ref-taxonomy-goldstandard.qza

```
QIIME feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs-goldstandard.qza \
  --i-reference-taxonomy ref-taxonomy-goldstandard.qza \
  --o-classifier diatoms-classifier-goldstandard.qza

##Hybrid

QIIME tools import \
  --type 'FeatureData[Sequence]' \
  --input-path custom_QIIME2.fasta \
  --output-path ref-seqs-custom.qza

QIIME tools import \
  --type 'FeatureData[Taxonomy]' \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path custom_june.txt \
  --output-path ref-taxonomy-custom.qza

QIIME feature-classifier classify-hybrid-vsearch-sklearn\
  --i-reference-reads ref-seqs-custom.qza \
  --i-reference-taxonomy ref-taxonomy-custom.qza \
  --i-classifier diatoms-classifier-custom.qza \
  --o-classification diatoms-classifier-hybrid.qza

#################  IMPORT FROM ILLUMINA ####################

QIIME tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path both_run \
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \
  --output-path demux.qza

QIIME demux summarize \
  --i-data demux.qza \
  --o-visualization demux.qzv

#################### DENOISING DADA2 WITH TRIMMING OF THE RSYST/DIAT.BARCODE PRIMERS ####################

QIIME DADA2 denoise-paired \
          --i-demultiplexed-seqs demux.qza \
          --p-trim-left-f 21 \
          --p-trim-left-r 27 \
          --p-trunc-len-f 240 \
          --p-trunc-len-r 200 \
          --p-chimera-method consensus \
          --p-max-ee 2 \
          --p-trunc-q 2 \
          --o-representative-sequences rep-seqs-DADA2.qza \
          --o-table table-DADA2.qza \
          --o-denoising-stats stats-DADA2.qza \
          --p-n-threads 15 \
          --verbose

# you can adapt the settings with your primers, for example  --p-trim-left-f 27 -p-trim-left-r 22 \ if you are using the diat.barcode primer instead of the UK set of primers
```

```
QIIME metadata tabulate \
  --m-input-file stats-DADA2.qza \
  --o-visualization stats-DADA2.qzv


QIIME feature-table summarize \
  --i-table table-DADA2.qza \
  --o-visualization table-DADA2.qzv \
  --m-sample-metadata-file metadata_INRA.tsv


QIIME feature-table tabulate-seqs \
  --i-data rep-seqs-DADA2.qza \
  --o-visualization rep-seqs-DADA2.qzv


################Chimera#################

QIIME vsearch uchime-denovo \
          --i-table table-DADA2.qza \
          --i-sequences rep-seqs-DADA2.qza \
          --output-dir uchime-dn-out

QIIME metadata tabulate \
  --m-input-file uchime-dn-out/stats.qza \
  --o-visualization uchime-dn-out/stats.qzv

QIIME feature-table filter-features \
  --i-table table-DADA2.qza \
  --m-metadata-file uchime-dn-out/nonchimeras.qza \
  --o-filtered-table uchime-dn-out/table-nonchimeric-wo-borderline.qza
QIIME feature-table filter-seqs \
  --i-data rep-seqs-DADA2.qza \
  --m-metadata-file uchime-dn-out/nonchimeras.qza \
  --o-filtered-data uchime-dn-out/rep-seqs-nonchimeric-wo-borderline.qza
QIIME feature-table summarize \
  --i-table uchime-dn-out/table-nonchimeric-wo-borderline.qza \
  --o-visualization uchime-dn-out/table-nonchimeric-wo-borderline.qzv

QIIME feature-table filter-features \
  --i-table table-DADA2.qza \
  --m-metadata-file uchime-dn-out/chimeras.qza \
  --p-exclude-ids \
  --o-filtered-table uchime-dn-out/table-nonchimeric-w-borderline.qza
QIIME feature-table filter-seqs \
  --i-data rep-seqs-DADA2.qza \
  --m-metadata-file uchime-dn-out/chimeras.qza \
  --p-exclude-ids \
  --o-filtered-data uchime-dn-out/rep-seqs-nonchimeric-w-borderline.qza
QIIME feature-table summarize \
  --i-table uchime-dn-out/table-nonchimeric-w-borderline.qza \
  --o-visualization uchime-dn-out/table-nonchimeric-w-borderline.qzv

######## TAXONOMIC ASSIGNMENT #########

#without the removal of the chimera
```

```
QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms-classifier-custom.qza \
  --i-reads rep-seqs-DADA2.qza \
  --o-classification taxonomy-custom.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-custom.qza \
  --o-visualization taxonomy-custom.qzv

QIIME taxa barplot \
  --i-table table-DADA2.qza \
  --i-taxonomy taxonomy-custom.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-custom.qzv

#Chimera with borderlines#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms-classifier-custom.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-w-borderline.qza \
  --o-classification taxonomy-nochimera-WB-custom.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WB-custom.qza \
  --o-visualization taxonomy-nochimera-WB-custom.qza

QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-w-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WB-custom.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WB-custom.qzv

#Chimera without borderline#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms-classifier-custom.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-wo-borderline.qza \
  --o-classification taxonomy-nochimera-WoB-custom.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WoB-custom.qza \
  --o-visualization taxonomy-nochimera-WoB-custom.qza

QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-wo-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WoB-custom.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WoB-custom.qzv

##GOLD_standard (Kelly  et al. 2018)

#Chimera with borderlines#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms_classifier_GS_aligned.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-w-borderline.qza \
  --o-classification taxonomy-nochimera-WB-GS-aligned.qza
```

```
QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WB-GS-aligned.qza \
  --o-visualization taxonomy-nochimera-WB-GS-aligned.qza

QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-w-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WB-GS-aligned.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WB-GS-aligned.qzv

#Chimera without borderline#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms_classifier_GS_aligned.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-wo-borderline.qza \
  --o-classification taxonomy-nochimera-WoB-GS-aligned.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WoB-GS-aligned.qza \
  --o-visualization taxonomy-nochimera-WoB-GS-aligned.qza

QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-wo-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WoB-GS-aligned.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WoB-GS-aligned.qzv

##Rsyst_Classic

#Chimera with borderlines#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms_classifier_classic.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-w-borderline.qza \
  --o-classification taxonomy-nochimera-WB-classic.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WB-classic.qza \
  --o-visualization taxonomy-nochimera-WB-classic.qza

QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-w-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WB-classic.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WB-classic.qzv

#Chimera without borderline#

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms_classifier_classic.qza \
  --i-reads uchime-dn-out/rep-seqs-nonchimeric-wo-borderline.qza \
  --o-classification taxonomy-nochimera-WoB-classic.qza

QIIME metadata tabulate \
  --m-input-file taxonomy-nochimera-WoB-classic.qza \
  --o-visualization taxonomy-nochimera-WoB-classic.qza
```

```
QIIME taxa barplot \
  --i-table uchime-dn-out/table-nonchimeric-wo-borderline.qza \
  --i-taxonomy taxonomy-nochimera-WoB-classic.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-nochimera-WoB-classic.qzv

QIIME feature-classifier classify-sklearn \
  --i-classifier diatoms-classifier-custom.qza \
  --i-reads rep-seqs-DADA2.qza \
  --o-classification taxonomy.qza

QIIME metadata tabulate \
  --m-input-file taxonomy.qza \
  --o-visualization taxonomy.qzv

QIIME taxa barplot \
  --i-table table-DADA2.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots.qzv

QIIME metadata tabulate \
  --m-input-file taxonomy-custom.qza \
  --o-visualization taxonomy-custom.qzv

QIIME taxa barplot \
  --i-table table-DADA2.qza \
  --i-taxonomy taxonomy-custom.qza \
  --m-metadata-file metadata_INRA.tsv \
  --o-visualization taxa-bar-plots-custom.qzv


############## ALTERNATIVE STEP ######################################

########## REMOVING ADAPTER #################

QIIME DADA2 denoise-paired \
        --i-demultiplexed-seqs demux_nobarcode.qza \
        --p-trunc-len-f 240 \
        --p-trunc-len-r 180 \
        --p-adapter-f [ForwardPrimer] \
        --p-adapter-r [ReversePrimer] \
        --p-chimera-method consensus \
        --p-max-ee 2 \
        --p-trunc-q 2 \
        --o-representative-sequences rep-seqs-DADA2.qza \
        --o-table table-DADA2.qza \
        --o-denoising-stats stats-DADA2.qza \
        --p-n-threads 25 \
        --verbose
```

NGSPECIESID SCRIPT

Bioinformatic pipeline script for MinION ONT using NGSpeciesID(Sahlin et al., 2021) polisher, and Mothur(Schloss et al., 2009) script naïve Bayesian classifier assignment (Wang et al., 2007).

```
#####Basecalling#####

Guppy_basecaller -I fast5/ -s basecalled/ -r --device auto -q 0 --disable_qscore_filtering -c DNA_r10.4.1_400bps_sup.cfg --compress_fastq

####Demultiplexing#####

guppy_barcoder -t 10 --device auto -i basecalled/ -s barcoded/ -r -q 0 --compress_fastq


##### Long barcode #####

for file in *.fastq; do
bn=`basename $file .fastq`
NGSpeciesID --ont --consensus --sample_size 500 --m 800 --s 100 --medaka --primer_file primers.txt --fastq $file --outfolder ${bn}

##### Short barcode #####

for file in *.fastq; do
bn=`basename $file .fastq`
do NGSpeciesID --ont --m 331 --s 100 --abundance_ratio 0.001 --fastq $file --outfolder ${bn} --primer_file Primer_UK_rbcLshort.fa --consensus –medaka
```

Resulting sequences were classified using Mothur and the diat.barcode reference library (custom version with added non-diatom sequences) on the Galaxy server in order to parallelize the process on each fasta file :

```
##### Long barcode #####
classify.seqs(fasta=*.fasta,reference=full_length_diatbarcode_custom.fasta, taxonomy= full_length _diatbarcode_custom.tax)
##### Short barcode #####
 classify.seqs(fasta=*.fasta, reference=short_diatbarcode_custom.fasta, taxonomy= short_diatbarcode_custom.tax)
```

MODELS OF VARIABLE Ω RATIOS AMONG SITES

(Yang and Nielsen, 2002)

Model code	p	Parameters	Notes
M0 (one-ratio)	1	ω	One ω ratio for all sites
M1 (neutral)	1	p_0	$p_1 = 1 - p_0$, $\omega_0 = 0$, $\omega_1 = 1$
M2 (selection)	3	p_0, p_1, ω_2	$p_2 = 1 - p_0 - p_1$, $\omega_0 = 0$, $\omega_1 = 1$
M3 (discrete)	$2K - 1$	p_0, p_1, … , p_{K-2},	$p_{K-1} = 1 - p_0 - p_1 - \ldots - p_{K-2}$
	($K = 3$)	ω_0, ω_1, . . ω_{K-1}	
M4 (freqs)	$K-1$ ($K = 5$)	p_0, p_1, … , p_{K-2}	The ω_k are fixed at 0, ⅓>, ⅔, 1, and 3
M5 (gamma)	2	α, β	From G (α, β)
M6 (2gamma)	4	p_0, α_0, β_0, α_1	p_0 from G (α_0, β_0) and $p_1 = 1 - p_0$ from G (α_1, α_1)
M7 (beta)	2	p, q	From B (p, q)
M8 (beta&ω)	4	p_0, p, q, ω	p_0 from B (p, q) and $1 - p_0$ with ω
M9 (beta&gamma)	5	p_0, p, q, α, β	p_0 from B (p, q) and $1 - p_0$ from G (α, β)
M10 (beta&gamma + 1)	5	p_0, p, q, α, β	p_0 from B (p, q) and $1 - p_0$ from 1 + G (α, β)
M11 (beta&normal>1)	5	p_0, p, q, μ σ	p_0 from B (p, q) and $1 - p_0$ from N (μ σ2), truncated to ω > 1
M12 (0&2normal>1)	5	p_0, p_1, μ_2, σ_1, σ_2	*p0* with $\omega_0 = 0$ and $1 - p_0$ from the mixture: p_1 from N (1, $\sigma_1$2), and $1 - p_1$ from N (μ, $\sigma_2$2), both normals truncated to ω > 1
M13 (3normal>0)	6	p_0, p_1, μ_2, σ_0, σ_1, σ_2	***p0*** from N (0, $\sigma_0$2), p_1 from N (1, $\sigma_1$2), and ***P2*** = $1 - p_0$ - Pl from N (μ_2, $\sigma_2$2), all normals truncated to ω > 1

p, number of parameters in the ω distribution.

REFERENCES

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/NAR/GKY379

Aguirre, L.E., Ouyang, L., Elfwing, A., Hedblom, M., Wulff, A., Inganäs, O., 2018. Diatom frustules protect DNA from ultraviolet light. Sci. Rep. 8. https://doi.org/10.1038/S41598-018-21810-2

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A., Knight, R., 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2. https://doi.org/10.1128/msystems.00191-16

Anisimova, M., Bielawski, J.P., Yang, Z., 2001. Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. Mol. Biol. Evol 18, 1585–1592.

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

Bach, L.T., Riebesell, U., Gutowska, M.A., Federwisch, L., Schulz, K.G., 2015. A unifying concept of coccolithophore sensitivity to changing carbonate chemistry embedded in an ecological framework. Prog. Oceanogr. 135, 125–138. https://doi.org/10.1016/j.pocean.2015.04.012

Badger, M.R., Andrews, T.J., Whitney, S.M., Ludwig, M., Yellowlees, D.C., Leggat, W., Price, G.D., 1998. The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae. Can. J. Bot. 76, 1052–1071.

https://doi.org/10.1139/b98-074

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

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

Balfoort, H.W., Snoek, J., Smiths, J.R.M., Breedveld, L.W., Hofstraat, J.W., Ringelberg, J., 1992. Automatic identification of algae: neural network analysis of flow cytometric data. J. Plankton Res. 14, 575–589. https://doi.org/10.1093/PLANKT/14.4.575

Barnes, C.J., Rasmussen, L., Asplund, M., Knudsen, S.W., Clausen, M.L., Agner, T., Hansen, A.J., 2020. Comparing DADA2 and OTU clustering approaches in studying the bacterial communities of atopic dermatitis. J. Med. Microbiol. 69, 1293–1302. https://doi.org/10.1099/jmm.0.001256

Bedoshvili, Y.D., Popkova, T.P., Likhoshway, Y. V., 2009. Chloroplast structure of diatoms of different classes. Cell tissue biol. 3, 297–310. https://doi.org/10.1134/s1990519x09030122

Bendich, A.J., 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? BioEssays 6, 279–282.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., others, 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857.

Bresciani, M., Bolpagni, R., Braga, F., Oggioni, A., Giardino, C., 2012. Retrospective assessment

of macrophytic communities in southern Lake Garda (Italy) from in situ and MIVIS (Multispectral Infrared and Visible Imaging Spectrometer) data. J. Limnol. 71, 180–190. https://doi.org/10.3274/JL12-71-1-05

Brooks, S.J., Bennion, H., Birks, H.J.B., 2001. Tracing lake trophic history with a chironomid-total phosphorus inference model. Freshw. Biol. 46, 513–533. https://doi.org/10.1046/j.1365-2427.2001.00684.x

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

Carbiener, R., Trémolières, M., Mercier, J.L., Ortscheit, A., 1990. Aquatic macrophyte communities as bioindicators of eutrophication in calcareous oligosaprobe stream waters (Upper Rhine plain, Alsace). Vegetatio 86, 71–88. https://doi.org/10.1007/BF00045135

Carter, J., Walling, D.E., Owens, P.N., Leeks, G.J.L., 2006. Spatial and temporal variability in the concentration and speciation of metals in suspended sediment transported by the River Aire, Yorkshire, UK. Hydrol. Process. 20, 3007–3027. https://doi.org/10.1002/HYP.6156

CEMAGREF, 1982. Etude des methodes biologiques dappreciation quantitative de la qualite des eaux.

CEN, 2014. Water quality — Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers.

Cermeño, P., Falkowski, P.G., Romero, O.E., Schaller, M.F., Vallina, S.M., 2015. Continental erosion and the Cenozoic rise of marine diatoms. Proc. Natl. Acad. Sci. U. S. A. 112, 4239–4244. https://doi.org/10.1073/PNAS.1412883112

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

Chepurnov, V.A., Mann, D.G., Sabbe, K., Vyverman, W., 2004. Experimental Studies on Sexual

Reproduction in Diatoms. Int. Rev. Cytol. 237, 91–154. https://doi.org/10.1016/S0074-7696(04)37003-8

Cooper, A.R., Infante, D.M., Wehrly, K.E., Wang, L., Brenden, T.O., 2016. Identifying indicators and quantifying large-scale effects of dams on fishes. Ecol. Indic. 61, 646–657. https://doi.org/10.1016/j.ecolind.2015.10.016

Danilov, R.A., Ekelund, N.G.A., 2001. Phytoplankton communities at different depths in two eutrophic and two oligotrophic temperate lakes at higher latitude during the period of ice cover. Acta Protozool. 40, 197–201.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. Biol. Lett. 10, 2–5. https://doi.org/10.1098/rsbl.2014.0562

Del Carmen Pérez, M., Maidana, N.I., Comas, A., 2009. Phytoplankton composition of the Ebro River estuary, Spain. Acta Bot. Croat. 68, 11–27.

Descy, J.-P., Coste, M., 1991. A test of methods for assessing water quality based on diatoms. SIL Proceedings, 1922-2010 24, 2112–2116. https://doi.org/10.1080/03680770.1989.11899905

Descy, J.., Coste, M., 1990. Utilisation des diatomées benthiques pour l'évaluation de la qualité des eaux courantes. Contrat CEE B-71 -23. Rapport final, Facultés Universitaires N.D. d e la Paix, Namur - CEMAGREF.

Descy, J.P., Leitao, M., Everbecq, E., Smitz, J.S., Delige, J.F., 2012. Phytoplankton of the river loire, France: A biodiversity and modelling study. J. Plankton Res. 34, 120–135. https://doi.org/10.1093/plankt/fbr085

Duleba, M., Földi, A., Micsinai, A., Várbíró, G., Mohr, A., Sipos, R., Szabó, G., Buczkó, K., Trábert, Z., Kiss, K.T., Bíró, T., Vadkerti, E., Ács, É., 2021. Applicability of diatom metabarcoding in the ecological status assessment of Hungarian lotic and soda pan habitats. Ecol. Indic. 130. https://doi.org/10.1016/j.ecolind.2021.108105

Edgar, R., 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 081257. https://doi.org/10.1101/081257

Edgar, R.C., Flyvbjerg, H., 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31, 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

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

Emiliani, M.O.G., 1997. Effects of water level fluctuations on phytoplankton in a river-floodplain lake system (Paraná River, Argentina). Hydrobiologia 357, 1–15. https://doi.org/10.1023/a:1003149514670

Environment Agency, 2016. Climate change and eutrophication risk in English rivers, Report – SC140013/R.

Environment Agency, n.d. Environment Agency water quality data Water Quality Archive [WWW Document]. URL https://environment.data.gov.uk/water-quality

Erb, T.J., Zarzycki, J., 2018. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. Curr. Opin. Biotechnol. 49, 100–107. https://doi.org/10.1016/j.copbio.2017.07.017

European Commission, 2021. Report on the implementation of Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources based on Member State reports for the period 2016–2019. Eur. Commun.

European Committee for Standardization, 2003. Water quality — Guidance standard for the routine sampling and pre- treatment of benthic diatoms from rivers, prEN 13946:2002 (E).

Eyre, B., Twigg, C., 1997. Nutrient Behaviour During Post-flood Recovery of the Richmond River Estuary Northern NSW, Australia. Estuar. Coast. Shelf Sci. 44, 311–326.

Falasco, E., Ector, L., Wetzel, C.E., Badino, G., Bona, F., 2021. Looking back, looking forward: a review of the new literature on diatom teratological forms (2010–2020), Hydrobiologia. Springer International Publishing. https://doi.org/10.1007/s10750-021-04540-x

Falkowski, P.G., Knoll, A.H., 2007. Evolution of Primary Producers in the Sea, Academic Press. Elsevier Academic Press.

Fasching, C., Wilson, H.F., D'Amario, S.C., Xenopoulos, M.A., 2019. Natural Land Cover in Agricultural Catchments Alters Flood Effects on DOM Composition and Decreases Nutrient Levels in Streams. Ecosystems 22, 1530–1545. https://doi.org/10.1007/s10021-019-00354-0

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

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science (80-.). 281, 237–240. https://doi.org/10.1126/SCIENCE.281.5374.237)

Fife, M.G., Walls, P.J., 1981. The River Foss: From Yearsley Village to York Its History and Natural History.

Gao, F., Chen, C., Arab, D.A., Du, Z., He, Y., Ho, S.Y.W., 2019. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol. Evol. 9, 3891–3898. https://doi.org/10.1002/ece3.5015

Gao, X., Wang, C., Cui, H., 2014. Identification of bundle sheath cell fate factors provides new tools for C3-to-C4 engineering. Plant Signal. Behav. 9, 12–14. https://doi.org/10.4161/psb.29162

Glover, R., 2019. Biomonitoring and surveillance with short-and long-read metabarcoding 1–304.

Goss, R., Wilhelm, C., Jakob, T., 2020. Photosynthesis in diatoms, Handbook of Algal Science, Technology and Medicine. Elsevier Inc. https://doi.org/10.1016/B978-0-12-818305-2.00013-9

Grizzetti, B., Lanzanova, D., Liquete, C., Reynaud, A., Cardoso, A.C., 2016. Assessing water ecosystem services for water resource management. Environ. Sci. Policy 61, 194–203.

https://doi.org/10.1016/j.envsci.2016.04.008

Gross, M., 2012. The mysteries of the diatoms. Curr. Biol. 22, R581–R585. https://doi.org/10.1016/J.CUB.2012.07.041

Guiry, M.D., 2012. How many species of algae are there? J. Phycol. 48, 1057–1063. https://doi.org/10.1111/j.1529-8817.2012.01222.x

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

Hamsher, S.E., Evans, K.M., Mann, D.G., Poulíčková, A., Saunders, G.W., 2011. Barcoding Diatoms: Exploring Alternatives to COI-5P. Protist 162, 405–422. https://doi.org/10.1016/J.PROTIS.2010.09.005

Hatzenbuhler, C., Kelly, J.R., Martinson, J., Okum, S., Pilgrim, E., 2017. Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. Sci. Rep. 7, 1–10. https://doi.org/10.1038/srep46393

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

Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., Kelly, M., 2018. Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. Water Res. 138, 192–205. https://doi.org/10.1016/j.watres.2018.03.003

Herringer, J.W., Lester, D., Dorrington, G.E., Rosengarten, G., 2019. Can diatom girdle band pores act as a hydrodynamic viral defense mechanism? J. Biol. Phys. 213–234. https://doi.org/10.1007/s10867-019-09525-5

Hönisch, B., Ridgwell, A., Schmidt, D.N., Thomas, E., Gibbs, S.J., Sluijs, A., Zeebe, R., Kump, L., Martindale, R.C., Greene, S.E., Kiessling, W., Ries, J., Zachos, J.C., Royer, D.L., Barker, S.,

Marchitto, T.M., Moyer, R., Pelejero, C., Ziveri, P., Foster, G.L., Williams, B., 2012. The geological record of ocean acidification. Science (80-.). 335, 1058–1063. https://doi.org/10.1126/SCIENCE.1208277

Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170. https://doi.org/10.1038/335167A0

Huo, S., Li, X., Xi, B., Zhang, H., Ma, C., He, Z., 2020. Combining morphological and metabarcoding approaches reveals the freshwater eukaryotic phytoplankton community. Environ. Sci. Eur. 32. https://doi.org/10.1186/s12302-020-00321-w

Husson, F., Josse, A.J., Jérôme, A., Agrocampus, P., 2010. Technical Report-Agrocampus Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data? Appl. Math. Dep. 1–17.

Iida, S., Miyagi, A., Aoki, S., Ito, M., Kadono, Y., Kosuge, K., 2009. Molecular Adaptation of rbcL in the Heterophyllous Aquatic Plant Potamogeton. PLoS One 4, e4633. https://doi.org/10.1371/JOURNAL.PONE.0004633

Jacquet, S., Briand, J.F., Leboulanger, C., Avois-Jacquet, C., Oberhaus, L., Tassin, B., Vinçon-Leite, B., Paolini, G., Druart, J.C., Anneville, O., Humbert, J.F., 2005. The proliferation of the toxic cyanobacterium Planktothrix rubescens following restoration of the largest natural French lake (Lac du Bourget). Harmful Algae 4, 651–672. https://doi.org/10.1016/j.hal.2003.12.006

Jones, H.M., Simpson, G.E., Stickle, A.J., Mann, D.G., 2005. Life history and systematics of Petroneis (Bacillariophyta), with special reference to British waters. Eur. J. Phycol. 40, 61–87. https://doi.org/10.1080/09670260400024675

Juggins, S., Kelly, M., 2018. darleq3: User Guide (Version 0.8.4).

Kagzi, K., Hechler, R.M., Fussmann, G.F., Cristescu, M.E., 2022. Environmental RNA degrades more rapidly than environmental DNA across a broad range of pH conditions. Mol. Ecol. Resour. 22, 2640–2650. https://doi.org/10.1111/1755-0998.13655

Kallis, G., Butler, D., 2001. The EU water framework directive: measures and implications,

Water Policy

Kapralov, M. V., Filatov, D.A., 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. BMC Evol. Biol. 7. https://doi.org/10.1186/1471-2148-7-73

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., Bouchez, A., 2017. Freshwater biomonitoring in the Information Age. Front. Ecol. Environ. https://doi.org/10.1002/fee.1490

Kelly, M.G., Juggins, S., Bennion, H., Burgess, A., Yallop, M., Hirst, H., Jamieson, J., Guthrie, R., Rippey, B., 2014. DARLEQ : Diatom Assessment of River and Lake Ecological Quality User Guide Software for Freshwater Status Classification using benthic diatoms 1–11.

Kelly, M., Boonham, N., Juggins, S., Glover, R., 2020. Further development of a DNA based metabarcoding approach to assess diatom communities in rivers (C160014/R).

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D.G., Pass, D., Sapp, M., Sato, S., Glover, R., 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers, Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH.

~~Kelly, M., Ector, L., Goldsmith, B.J., 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. https://doi.org/10.1023/A~~

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

Kelly, M.G., 1998. Use of the trophic diatom index to monitor eutrophication in rivers. Water Res. 32, 236–242. https://doi.org/10.1016/S0043-1354(97)00157-7

Kelly, M.G., Birk, S., Willby, N.J., Denys, L., Drakare, S., Kahlert, M., Karjalainen, S.M., Marchetto, A., Pitt, J.A., Urbanič, G., Poikane, S., 2016. Redundancy in the ecological assessment of lakes: Are phytoplankton, macrophytes and phytobenthos all necessary? Sci. Total Environ. 568, 594–602. https://doi.org/10.1016/j.scitotenv.2016.02.024

Kelly, M. G., Cazaubon, A., Coring, E., Dell'Uomo, A., Ector, L., Goldsmith, B., Guasch, H., Hurlimann, J., Jarlman, A., Kawecka, B., Kwandrans, J., Laugaste, R., Lindstrom, E.A.,

Leitao, M., Marvan, P., Padisaka, J., Pipp, E., Prygiel, J., Rott, E., Sabater, S., Van Dam, H., Vizinet, J., 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. J. Appl. Phycol. 10, 215–224. https://doi.org/10.1023/A:1008033201227

Kelly, M G, Cazaubon, A., Coring, E., Uomo, A.D., Ector, L., Goldsmith, B., Guasch, H., Wasserforschung, I., Wielenbach, D.-, 1998. Recomendations the routine sampling of diatoms for quality in Europe. 215–224.

Kelly, M.G., Whitton, B.A., 1995. The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. J. Appl. Phycol. 7, 433–444. https://doi.org/10.1007/BF00003802

Kelly, R.P., Shelton, A.O., Gallego, R., 2019. Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. Sci. Rep. 9, 1–14. https://doi.org/10.1038/s41598-019-48546-x

Kenkel, N.C., Orloci, L., 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67, 919–928. https://doi.org/10.2307/1939814

Kermarrec, L., 2012. Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques et pour l'étude de leur taxonomie297 p.

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

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

Konur, O., 2020. Chapter 7 - The scientometric analysis of the research on the algal genomics, in: Konur, O. (Ed.), Handbook of Algal Science, Technology and Medicine. Academic

Press, pp. 105–125. https://doi.org/https://doi.org/10.1016/B978-0-12-818305-2.00007-3

Krehenwinkel, H., Pomerantz, A., Prost, S., 2019. Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: Current uses and future directions. Genes (Basel). 10. https://doi.org/10.3390/genes10110858

Kuczynska, P., Jemiola-Rzeminska, M., Strzalka, K., 2015. Photosynthetic pigments in diatoms. Mar. Drugs 13, 5847–5881. https://doi.org/10.3390/md13095847

Lane, C.R., 2007. Assessment of isolated wetland condition in Florida using epiphytic diatoms at genus, species, and subspecies taxonomic resolution. Ecohealth 4, 219–230. https://doi.org/10.1007/s10393-007-0098-0

Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., Kelly, M.G., 2017. Freshwater benthic diatoms of Central Europe: Over 800 common species used in ecological assessment (Vol. 942). M. G. Kelly (Ed.). Schmitten-Oberreifenberg: Koeltz Botanical Books. 942, 942.

Lange, K., Liess, A., Piggott, J.J., Townsend, C.R., Matthaei, C.D., 2011. Light, nutrients and grazing interact to determine stream diatom community composition and functional group structure. Freshw. Biol. 56, 264–278. https://doi.org/10.1111/j.1365-2427.2010.02492.x

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J., 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. https://doi.org/10.1016/j.bdq.2015.02.001

Lavoie, I., Dillon, P.J., Campeau, S., 2009. The effect of excluding diatom taxa and reducing taxonomic resolution on multivariate analyses and stream bioassessment. Ecol. Indic. 9, 213–225. https://doi.org/10.1016/j.ecolind.2008.04.003

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25, 1–18. https://doi.org/10.18637/JSS.V025.I01

Lenoir, A., Coste, M., 1996. Development of a practical diatom index of overall water quality applicable to the French National Water Board Network, in: International Symposium,

Volksbildungsheim Grilhof Vill, AUT, 17-19 September 1995. pp. 29–43.

Lepori, F., Barbieri, A., Ormerod, S.J., 2003. Effects of episodic acidification on macroinvertebrate assemblages in Swiss Alpine streams. Freshw. Biol. 48, 1873–1885. https://doi.org/10.1046/j.1365-2427.2003.01121.x

Li, L., Zheng, B., Liu, L., 2010. Biomonitoring and bioindicators used for river ecosystems: Definitions, approaches and trends. Procedia Environ. Sci. 2, 1510–1524. https://doi.org/10.1016/j.proenv.2010.10.164

Lin, B., Hui, J., Mao, H., 2021. Nanopore technology and its applications in gene sequencing. Biosensors 11. https://doi.org/10.3390/bios11070214

Litchman, E., Klausmeier, C.A., Yoshiyama, K., 2009. Contrasting size evolution in marine and freshwater diatoms. Proc. Natl. Acad. Sci. 106, 2665–2670. https://doi.org/10.1073/PNAS.0810891106

Liu, G., Li, T., Zhu, X., Zhang, X., Wang, J., 2023. An independent evaluation in a CRC patient cohort of microbiome 16S rRNA sequence analysis methods: OTU clustering, DADA2, and Deblur. Front. Microbiol. 14. https://doi.org/10.3389/fmicb.2023.1178744

Lund, J.W.G., 1972. Eutrophication. Proc. R. Soc. London. Ser. B. Biol. Sci. 180, 371–382.

Maitland, V.C., Robinson, C.V., Porter, T.M., Hajibabaei, M., 2020. Freshwater diatom biomonitoring through benthic kick-net metabarcoding. PLoS One 15, 1–18. https://doi.org/10.1371/journal.pone.0242143

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

Mann, D.G., 1989. The diatom genus Sellaphora: Separation from Navicula. Br. Phycol. J. 24, 1–20. https://doi.org/10.1080/00071618900650011

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

Marion, G.M., Millero, F.J., Camões, M.F., Spitzer, P., Feistel, R., Chen, C.T.A., 2011. pH of

seawater. Mar. Chem. 126, 89–96. https://doi.org/10.1016/J.MARCHEM.2011.04.002

Massingham, T., Goldman, N., 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169, 1753–1762. https://doi.org/10.1534/genetics.104.032144

McMurdie, P.J., Holmes, S., 2013. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One 8. https://doi.org/10.1371/journal.pone.0061217

Meyer, M.T., Genkov, T., Skepper, J.N., Jouhet, J., Mitchell, M.C., Spreitzer, R.J., Griffiths, H., 2012. Rubisco small-subunit α-helices control pyrenoid formation in Chlamydomonas. Proc. Natl. Acad. Sci. U. S. A. 109, 19474–19479. https://doi.org/10.1073/PNAS.1210993109/-/DCSUPPLEMENTAL

Mir, Z.A., Arafat, M.Y., Bakhtiyar, Y., 2021. Benthic Macroinvertebrates as Bioindicators of Water Quality in Freshwater Bodies. Freshw. Pollut. Aquat. Ecosyst. 165–184. https://doi.org/10.1201/9781003130116-7

Miyata, K., Inoue, Y., Amano, Y., Nishioka, T., Nagaike, T., Kawaguchi, T., Morita, O., Yamane, M., Honda, H., 2022. Comparative environmental RNA and DNA metabarcoding analysis of river algae and arthropods for ecological surveys and water quality assessment. Sci. Rep. 12, 1–13. https://doi.org/10.1038/s41598-022-23888-1

Montresor, M., Vitale, L., D'Alelio, D., Ferrante, M.I., 2016. Sex in marine planktonic diatoms: insights and challenges. Perspect. Phycol. 3, 61–75. https://doi.org/10.1127/PIP/2016/0045

Moog, O., Schmutz, S., Schwarzinger, I., 2018. Biomonitoring and Bioassessment BT - Riverine Ecosystem Management: Science for Governing Towards a Sustainable Future, in: Schmutz, S., Sendzimir, J. (Eds.), . Springer International Publishing, Cham, pp. 371–390. https://doi.org/10.1007/978-3-319-73250-3_19

Mora, D., Abarca, N., Proft, S., Grau, J.H., Enke, N., Carmona, J., Skibbe, O., Jahn, R., Zimmermann, J., 2019. Morphology and metabarcoding: A test with stream diatoms from Mexico highlights the complementarity of identification methods. Freshw. Sci. 38,

448–464. https://doi.org/10.1086/704827

Moroney, J. V, Somanchi, A., 1999. How do algae concentrate CO 2 to increase the efficiency of to increase the efficiency of photosynthetic carbon fixation? https://doi.org/10.1104/pp.119.1.9

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., Bhattacharya, D., 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science (80-.). 324, 1724–1726. https://doi.org/10.1126/science.1172983

Murtagh, F., Legendre, P., 2011. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm.

Muscutt, A.D., Withers, P.J.A., 1996. The phosphorus content of rivers in England and Wales. Water Res. https://doi.org/10.1016/0043-1354(95)00290-1

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

Nielsen, R., Yang, Z., 1998. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene.

Nonoyama, T., Kazamia, E., Nawaly, H., Gao, X., Tsuji, Y., Matsuda, Y., Bowler, C., Tanaka, T., Dorrell, R.G., 2019. Metabolic innovations underpinning the origin and diversification of the diatom chloroplast. Biomolecules 9. https://doi.org/10.3390/BIOM9080322

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F., Weedon, J., 2022. vegan: Community Ecology Package.

Okwuosa, O.B., Eyo, J.E., E., O.E., 2019. Role Of Fish as Bioindicators: A Review - IRE Journals. IRE Journals 2, 354–368.

Pandey, L.K., Bergey, E.A., Lyu, J., Park, J., Choi, S., Lee, H., Depuydt, S., Oh, Y.T., Lee, S.M., Han, T., 2017. The use of diatoms in ecotoxicology and bioassessment: Insights, advances and challenges. Water Res. 118, 39–58. https://doi.org/10.1016/j.watres.2017.01.062

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

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

Pearman, W.S., Freed, N.E., Silander, O.K., 2020. Testing the advantages and disadvantages of short- And long- read eukaryotic metagenomics using simulated reads. BMC Bioinformatics 21, 1–15. https://doi.org/10.1186/s12859-020-3528-4

Pearman, W.S., Freed, N.E., Silander, O.K., . The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition. https://doi.org/10.1101/650788

Pelusi, A., Santelia, M.E., Benevenuto, G., Godhe, A., Montresor, M., 2020. The diatom Chaetoceros socialis: spore formation and preservation. Eur. J. Phycol. 55, 1–10. https://doi.org/10.1080/09670262.2019.1632935

Pielou, E.C., 1966. The measurement of diversity in different types of biological collections. J. Theor. Biol. 13, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Pochon, X., Zaiko, A., Fletcher, L.M., Laroche, O., Wood, S.A., 2017. Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. PLoS One 12, 1–19. https://doi.org/10.1371/journal.pone.0187636

Prihoda, J., Tanaka, A., De Paula, W.B.M., Allen, J.F., Tirichine, L., Bowler, C., 2012. Chloroplast-mitochondria cross-talk in diatoms. J. Exp. Bot. 63, 1543–1557. https://doi.org/10.1093/jxb/err441

Prygiel, J., Carpentier, P., Almeida, S., Coste, M., Druart, J.C., Ector, L., Guillard, D., Honoré, M.A., Iserentant, R., Ledeganck, P., Lalanne-Cassou, C., Lesniak, C., Mercier, I., Moncaut, P., Nazart, M., Nouchet, N., Peres, F., Peeters, V., Rimet, F., Rumeau, A., Sabater, S., Straub, F., Torrisi, M., Tudesque, L., Van de Vijver, B., Vidal, H., Vizinet, J., Zydek, N., 2002. Determination of the biological diatom index (IBD NF T 90-354): Results of an intercomparison exercise. J. Appl. Phycol. 14, 27–39. https://doi.org/10.1023/A:1015277207328

Prygiel, J., Coste, M., 2000. Guide méthodologique pour la mise en œuvre de l'Indice Biologique Diatomées NF T 90-354.

R Core Team, 2022. R: A Language and Environment for Statistical Computing.

Rabalais, N.N., 2002. Nitrogen in aquatic ecosystems. Ambio 31, 102–112. https://doi.org/10.1579/0044-7447-31.2.102

~~Ramachandra, T. V, 2010. Protocols for Collection , Preservation and Enumeration of Diatoms from Aquatic Habitats for Water Quality Monitoring in India ! rroottooccoollss ffoorr CCoollllleeccttiioonn ,, ! rreesseerrvvaattiioonn aanndd EEnnuummeerraattiioonn ooff DDiiaattoommss ffr.~~

Rauwolf, U., Golczyk, H., Greiner, S., Herrmann, R.G., 2010. Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. Mol. Genet. Genomics 283, 35–47. https://doi.org/10.1007/s00438-009-0491-1

Raven, J.A., 2010. Inorganic carbon acquisition by eukaryotic algae: four current questions. Photosynth. Res. 2010 1061 106, 123–134. https://doi.org/10.1007/S11120-010-9563-7

Richard, C., Mitbavkar, S., Landoulsi, J., 2017. Diagnosis of the Diatom Community upon Biofilm Development on Stainless Steels in Natural Freshwater. https://doi.org/10.1155/2017/5052646

Riley, W.D., Potter, E.C.E., Biggs, J., Collins, A.L., Jarvie, H.P., Jones, J.I., Kelly-Quinn, M.,

Ormerod, S.J., Sear, D.A., Wilby, R.L., Broadmeadow, S., Brown, C.D., Chanin, P., Copp, G.H., Cowx, I.G., Grogan, A., Hornby, D.D., Huggett, D., Kelly, M.G., Naura, M., Newman, J.R., Siriwardena, G.M., 2018. Small Water Bodies in Great Britain and Ireland: Ecosystem function, human-generated degradation, and options for restorative action. Sci. Total Environ. 645, 1598–1616. https://doi.org/10.1016/j.scitotenv.2018.07.243

Rimet, F., Bouchez, A., 2012. Biomonitoring river diatoms: Implications of taxonomic resolution. Ecol. Indic. 15, 92–99. https://doi.org/10.1016/j.ecolind.2011.09.014

Rimet, Frédéric, Chardon, C., Lainé, L., Bouchez, A., Domaizon, I., Guillard, J., Jacquet, S., 2018. Thonon Culture Collection -TCC- a freshwater microalgae collection [WWW Document]. https://doi.org/https://doi.org/10.15454/UQEMVW

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

~~Rimet, F., Ector, L., Cauchie, H., Hoffmann, L., 2009. Changes in diatom-dominated biofilms during simulated improvements in water quality : implications for diatom-based monitoring in rivers Changes in diatom-dominated biofilms during simulated improvements in water quality : implications for diatom based mo 0262. https://doi.org/10.1080/09670260903198521~~

Rimet, Frederic, Gusev, E., Kahlert, M., Kelly, M., Kulikovskiy, M., Maltsev, Y., Mann, D., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2018. Diat.barcode, an open-access barcode library for diatoms. https://doi.org/https://doi.org/10.15454/TOMBYZ

Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M., Maltsev, Y., Mann, D.G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2019. Diat.barcode, an open-access curated barcode library for diatoms. Sci. Reports 2019 91 9, 1–12. https://doi.org/10.1038/s41598-019-51500-6

Rishan, S.T., Kline, R.J., Rahman, M.S., 2023. Applications of environmental DNA (eDNA) to detect subterranean and aquatic invasive species: A critical review on the challenges and limitations of eDNA metabarcoding. Environ. Adv. 12, 100370.

https://doi.org/10.1016/j.envadv.2023.100370

River Water Quality Monitoring 1990 to 2018 - pH - data.gov.uk [WWW Document], n.d. URL https://data.gov.uk/dataset/07c39402-9eae-4d6b-adb5-2625d230e002/river-water-quality-monitoring-1990-to-2018-ph (accessed 2.4.22).

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

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016. https://doi.org/10.7717/peerj.2584

Round, F.E., Crawford, R.M., Mann, D.G., 1990. Diatoms: biology and morphology of the genera. Cambridge university press

Sahlin, K., Lim, M.C.W., Prost, S., 2021. NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. Ecol. Evol. 11, 1392–1398. https://doi.org/10.1002/ece3.7146

Salmaso, N., 1996. Seasonal variation in the composition and rate of change of the phytoplankton community in a deep subalpine lake (Lake Garda, Northern Italy). An application of nonmetric multidimensional scaling and cluster analysis. Hydrobiologia 337, 49–68. https://doi.org/10.1007/BF00028506

Salzberg, S.L., Wood, D.E., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15. https://doi.org/10.1126/science.1093857

Sánchez, C., Cristóbal, G., Bueno, G., 2019. Diatom identification including life cycle stages through morphological and texture descriptors. PeerJ 2019, 1–24. https://doi.org/10.7717/peerj.6770

Sanyal, A., Larsson, J., van Wirdum, F., Andrén, T., Moros, M., Lönn, M., Andrén, E., 2022. Not dead yet: Diatom resting spores can survive in nature for several millennia. Am. J. Bot. 109, 67–82. https://doi.org/10.1002/ajb2.1780

Sbihi, K., Cherifi, O., Bertrand, M., El Gharmali, A., 2014. Biosorption of metals (Cd, Cu and Zn)

by the freshwater diatom Planothidium lanceolatum: A laboratory study. Diatom Res. 29, 55–63. https://doi.org/10.1080/0269249X.2013.872193

Schenekar, T., 2023. The current state of eDNA research in freshwater ecosystems: are we shifting from the developmental phase to standard application in biomonitoring? Hydrobiologia 850, 1263–1282. https://doi.org/10.1007/s10750-022-04891-z

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

Schneider, S.C., 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

Shams El-Din, N.G., El-Sheekh, M.M., El-Kassas, H.Y., Essa, D.I., El-Sherbiny, B.A., 2022. Biological indicators as tools for monitoring water quality of a hot spot area on the Egyptian Mediterranean Coast. Arab. J. Geosci. 15. https://doi.org/10.1007/s12517-022-10701-6

Smol, J.P., 1985. The ratio of diatom frustules to chrysophycean statospores: A useful paleolimnological index. Hydrobiologia 123, 199–208. https://doi.org/10.1007/BF00034378

Smyth, R.P., Schlub, T.E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M.P., Mak, J., 2010. Reducing chimera formation during PCR amplification to ensure accurate genotyping. Gene 469, 45–51. https://doi.org/10.1016/J.GENE.2010.08.009

Stefan, C.P., Hall, A.T., Graham, A.S., Minogue, T.D., 2022. Comparison of Illumina and Oxford Nanopore Sequencing Technologies for Pathogen Detection from Clinical Matrices Using Molecular Inversion Probes. J. Mol. Diagnostics 24, 395–405. https://doi.org/10.1016/j.jmoldx.2021.12.005

Stevens, B.M., Creed, T.B., Reardon, C.L., Manter, D.K., 2023. Comparison of Oxford Nanopore

Technologies and Illumina MiSeq sequencing with mock communities and agricultural soil. Sci. Rep. 13, 1–11. https://doi.org/10.1038/s41598-023-36101-8

Stevenson, J., 2014. Ecological assessments with algae: a review and synthesis. J. Phycol. 50, 437–461. https://doi.org/10.1111/jpy.12189

Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16, 1315–1328. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026042

Swanson, W.J., Nielsen, R., Yang, Q., 2003. Pervasive Adaptive Evolution in Mammalian Fertilization Proteins. Mol. Biol. Evol 20, 18–20.

Taberlet, Pierre, Coissac, E., Hajibabaei, M., Rieseberg, L.H., 2012. Environmental DNA. Mol. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P, Coissac, E., Pompanon, F., Brochmann, C., Willerslec, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding - TABERLET - 2012 - Molecular Ecology - Wiley Online Library. Mol. Ecol. 21, 2045–2050.

Tarkowska-Kukuryk, M., Mieczan, T., 2017. Submerged macrophytes as bioindicators of environmental conditions in shallow lakes in eastern Poland. Ann. Limnol. 53, 27–34. https://doi.org/10.1051/limn/2016031

Tréguer, P., Bowler, C., Moriceau, B., Dutkiewicz, S., Gehlen, M., Aumont, O., Bittner, L., Dugdale, R., Finkel, Z., Iudicone, D., Jahn, O., Guidi, L., Lasbleiz, M., Leblanc, K., Levy, M., Pondaven, P., 2017. Influence of diatom diversity on the ocean biological carbon pump. Nat. Geosci. 111 11, 27–37. https://doi.org/10.1038/s41561-017-0028-x

Trobajo, R., Mann, D.G., Clavero, E., Evans, K.M., Vanormelingen, P., McGregor, R.C., 2010. The use of partial cox1, rbcL and LSU rDNA sequences for phylogenetics and species identification within the nitzschia palea species complex (bacillariophyceae). Eur. J. Phycol. 45, 413–425. https://doi.org/10.1080/09670262.2010.498586

Tsarenko, P.M., Bilous, O.P., Kryvosheia-Zakharova, O.M., Lilitska, H.H., Barinova, S., 2021. Diversity of algae and cyanobacteria and bioindication characteristics of the alpine lake nesamovyte (Eastern carpathians, ukraine) from 100 years ago to the present. Diversity

13. https://doi.org/10.3390/d13060256

Tyrrell, T., Zeebe, R.E., 2004. History of carbonate ion concentration over the last 100 million years. Geochim. Cosmochim. Acta 68, 3521–3530. https://doi.org/10.1016/J.GCA.2004.02.018

Valegård, K., Andralojc, P.J., Haslam, R.P., Pearce, F.G., Eriksen, G.K., Madgwick, P.J., Kristoffersen, A.K., van Lun, M., Klein, U., Eilertsen, H.C., Parry, M.A.J., Andersson, I., 2018. Structural and functional analyses of Rubisco from arctic diatom species reveal unusual posttranslational modifications. J. Biol. Chem. 293, 13033. https://doi.org/10.1074/JBC.RA118.003518

Vasselon, V., 2018. Barcoding et bioindication : développement du metabarcoding des diatomées pour l ’évaluation de la qualité des cours d ’eau.

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

~~Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A., 2017a. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? https://doi.org/10.1086/690649~~ DUPLICAT OF ONE BELOW. MIGHT NEED TO RENUMBER YOUR REFS (2017A, B, C..)

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

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

Veilleux, H.D., Misutka, M.D., Glover, C.N., 2021. Environmental DNA and environmental RNA: Current and prospective applications for biological monitoring. Sci. Total Environ. 782, 146891. https://doi.org/10.1016/j.scitotenv.2021.146891

Vercruysse, K., Grabowski, R.C., Hess, T., Lexartza-Artza, I., 2020. Linking temporal scales of suspended sediment transport in rivers: towards improving transferability of prediction. J. Soils Sediments 20, 4144–4159. https://doi.org/10.1007/S11368-020-02673-5

Visco, J.A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., Pawlowski, J., 2015. Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. Environ. Sci. Technol. https://doi.org/10.1021/es506158m

Wakasugi, T., Nagai, T., Kapoor, M., Sugita, M., Ito, M., Ito, S., Tsudzuki, J., Nakashima, K., Tsudzuki, T., Suzuki, Y., Hamada, A., Ohta, T., Inamura, A., Yoshinaga, K., Sugiura, M., 1997. Complete nucleotide sequence of the chloroplast genome from the green alga Chlorella vulgaris: The existence of genes possibly involved in chloroplast division. Proc. Natl. Acad. Sci. U. S. A. 94, 5967–5972. https://doi.org/10.1073/pnas.94.11.5967

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–7. https://doi.org/10.1128/AEM.00062-07

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845

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

Westcott, S.L., Schloss, P.D., 2017. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere 2, 1–11. https://doi.org/10.1128/mspheredirect.00073-17

Whitton, B.A., Ellwood, N.T.W., Kawecka, B., 2009. Biology of the freshwater diatom

Didymosphenia. Hydrobiologia 630, 1–37. https://doi.org/10.1007/s10750-009-9753-5

Wirth, R., Pap, B., Böjti, T., Shetty, P., Lakatos, G., Bagi, Z., Kovács, K.L., Maróti, G., 2020. Chlorella vulgaris and Its Phycosphere in Wastewater: Microalgae-Bacteria Interactions During Nutrient Removal. Front. Bioeng. Biotechnol. 8, 1–15. https://doi.org/10.3389/fbioe.2020.557572

Wong, W.S.W., Yang, Z., Goldman, N., Nielsen, R., 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168, 1041–1051. https://doi.org/10.1534/GENETICS.104.031153

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591. https://doi.org/10.1093/MOLBEV/MSM088

Yang, Z., Nielsen, R., 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol. Biol. Evol 19, 908–917.

Yang, Z., Nielsen, R., Goldman, N., Mette, A.-, Pedersen, K., 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics. 155,431-49

Yao, X., Tan, Y.H., Yang, J.B., Wang, Y., Corlett, R.T., Manen, J.F., 2019. Exceptionally high rates of positive selection on the rbcL gene in the genus Ilex (Aquifoliaceae). BMC Evol. Biol. 19, 1–13. https://doi.org/10.1186/S12862-019-1521-1/FIGURES/3

Yates, M.C., Derry, A.M., Cristescu, M.E., 2021. Environmental RNA: A Revolution in Ecological Resolution? Trends Ecol. Evol. 36, 601–609. https://doi.org/10.1016/j.tree.2021.03.001

Yoon, H.S., Hackett, J.D., Pinto, G., Bhattacharya, D., 2002. The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. U. S. A. 99, 15507–15512. https://doi.org/10.1073/pnas.242379899

Young, J.N., Heureux, A.M.C., Sharwood, R.E., Rickaby, R.E.M., Morel, F.M.M., Whitney, S.M., 2016. Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms. J. Exp. Bot. 67, 3445–3456. https://doi.org/10.1093/JXB/ERW163

Young, J.N., Rickaby, R.E.M., Kapralov, M. V, Filatov, D.A., 2012. Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. https://doi.org/10.1098/rstb.2011.0145

Zhang, G.K., Chain, F.J.J., Abbott, C.L., Cristescu, M.E., 2018. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities, Evolutionary Applications. 11, 1901-1914. https://doi.org/10.1111/eva.12694

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

Zimmermann, J., Jahn, R., Gemeinholzer, B., 2011. Barcoding diatoms: Evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org. Divers. Evol. 11, 173–192. https://doi.org/10.1007/s13127-011-0050-6