Int. J. Lim. 2023, 59, 8
© A. Canino *et al.*, Published by EDP Sciences, 2023
https://doi.org/10.1051/limn/2023008

Available online at:
www.limnology-journal.org

RESEARCH ARTICLE OPEN ACCESS

# Which barcode to decipher freshwater microalgal assemblages? Tests on mock communities

Alexis Canino[1,3], Clarisse Lemonnier[1,3], Benjamin Alric[1,3], Agnès Bouchez[1,3], Isabelle Domaizon[1,3], Christophe Laplace-Treyture[2,3] and Frédéric Rimet[1,3*]

[1] UMR CARRTEL, INRAE, Université Savoie Mont-Blanc, Thonon, France
[2] INRAE, UR EABX, Cestas, France
[3] Pôle R&D ECLA, France

Received: 2 May 2023; Accepted: 29 June 2023

**Abstract –** DNA metabarcoding can be a promising alternative to microscopy for analysing phytoplankton, a key ecological indicator for freshwater ecosystems. The aim of this study was to evaluate the performance of different barcodes and associated primer pairs to assess microalgal diversity with DNA metabarcoding using a single barcode targeting all microalgae. We investigated barcodes in 16S and 23S rRNA genes, encoding for prokaryotic ribosomal sub-units, that are present in Cyanobacteria as well as in chloroplasts. *In silico* PCR tests were carried out on eight 16S and five 23S primer pairs using the Phytool reference library. Two and three pairs were selected for 16S and 23S, respectively, to perform an *in vitro* metabarcoding test based on a mock community made of DNA extracts of 10 microalgae strains. The 23S pairs enabled to detect all species, whereas 16S ones failed in the detection of some of them. One pair was selected for each genetic marker, based on its efficiency and specificity towards microalgae (*e.g.* not heterotrophic bacteria). Another mock community covering a larger diversity (18 microalgae strains) was used to test the efficiency of the selected pairs and their ability to estimate relative abundances. The 23S pair performed better than the 16S one for detecting target species with also more accuracy to assess their relative abundances. We conclude that the 23S primer pair ECLA23S_F1/ECLA23S_R1 appears as a good candidate to decipher freshwater phytoplankton communities. As a next step, it will be necessary to confirm these results on a large diversity of natural communities.

**Keywords:** Metabarcoding / primers / barcode / phytoplankton / mock community / eutrophication / cyanobacteria

## 1 Introduction

Freshwater ecosystems are key to support human populations (Cardinale *et al.*, 2012). Microalgae are the most abundant primary producers in these ecosystems and are used to monitor their ecological quality. For instance, phytoplankton and benthic microalgae are ecological indicators required in Europe (Water Framework Directive; European Commission, 2000) and the US (Clean Water Act; United States, 1972) to assess lake and river ecological quality.

Classically, when using microalgae to assess the ecological quality of lakes or rivers, monitoring methods require assessing the taxonomic composition of the community. For instance, phytoplankton community composition is analysed by microscopy using the Utermohl (1958) method, which is standardised at the European level (CEN, 2006). Floristic lists are established from lugol preserved samples that are set to sediment during several hours and then examined by a taxonomist expert under an inverted microscope. In addition, biomass is estimated for each species for a given sample volume. However, this method is a lengthy process based on an ever-decreasing number of taxonomic experts. Metabarcoding is an appealing approach to speed up the process, to avoid the bottleneck due to lack of experts, and to obtain more robust and accurate identifications as it was shown for some particular microalgae like diatoms (*e.g.* Vasselon *et al.*, 2019).

Deciphering microalgal communities using eDNA metabarcoding is a challenge, due to their wide taxonomic diversity, ranging from prokaryotes to eukaryotes, in diverse phyla of the tree of life (Adl *et al.*, 2019, Hug *et al.*, 2016). Several studies have already tested the metabarcoding approach for phytoplankton, many of them using a combination of two different barcodes, one targeting the prokaryotic fraction, like the 16S rRNA gene, and the other targeting the eukaryotic fraction, like the 18S rRNA gene (*e.g.* Hug *et al.*, 2016, Needham and

*Corresponding author: frederic.rimet@inrae.fr

Fuhrman, 2016, Djurhuus *et al.*, 2017, Filker *et al.*, 2019, Nowinski *et al.*, 2019, Yarimizu *et al.*, 2020, Sildever *et al.*, 2022). These two genetic markers, 18S rRNA and 16S rRNA are well represented in reference libraries like Silva (Quast *et al.*, 2013) and PR2 (Guillou *et al.*, 2013). Using two different genetic markers (*e.g.* 16S rRNA and 18S rRNA) targeting different clades (cyanobacteria and photosynthetic eukaryotes) for a study whose objective is to analyse the composition of the entire algal community can be a major drawback, especially when one seeks to have a molecular method that could replace microscopy. Indeed, assembling the floristic lists from two different markers to get a single list is something that seems difficult to achieve, especially when it comes to respect the relative proportions of each taxon in the final floristic list. For this reason, in this study we looked for primers that amplify a single barcode that targets the entire phytoplankton community, whether prokaryotic or eukaryotic. The efficiency of a barcode relies on its ability to identify a taxon at species level (Hebert *et al.*, 2003), which is linked to both its polymorphism and specificity. As microalgae represent a widely diverse, polyphyletic clade, the availability of primer pairs targeting the whole of it while being specific enough can be challenging. In this framework, the 18S rRNA gene, although it is widely used (*e.g.* Debroas *et al.*, 2015, Capo *et al.*, 2017, Keck *et al.*, 2020), was not selected as it only targets the eukaryotic compartment, overlooking the prokaryotic taxa. Moreover, it has a low-resolution power for several algal clades, *e.g.* diatoms (Kermarrec *et al.*, 2013), Dinoflagellates (Stern *et al.*, 2012), generally detecting taxa at higher taxonomic levels than species level. On the other hand, 16S rRNA (Decelle *et al.*, 2015) and 23S rRNA (Djemiel *et al.*, 2020) genes are encoding for prokaryotic ribosomal components (small and large sub-unit, respectively) that are present in Cyanobacteria, as well as in the chloroplasts of eukaryotic photosynthetic organisms, thus potentially enabling the detection of both prokaryotic and eukaryotic phytoplankton species.

In this study, we aimed to investigate (1) if one single barcode could efficiently decipher the microalgal diversity and (2) which one would be the most appropriate for microalgal monitoring in freshwater environments. To achieve these aims, the assessment of the most suited barcodes and primer pairs was achieved with a four-step approach. (1) An *in silico* step was dedicated to the evaluation of different barcodes and primer pairs selected from the literature or designed for this study. This first step enabled a 1st selection based on primers and amplicon characteristics, as well as *in silico* amplification efficiency. (2) An *in vitro* test of the 1st selection of barcodes and primer pairs was carried out using a mock community composed of strains covering part of the phytoplankton diversity and introduced in equimolar proportions. This experiment enabled to check both the efficiency and specificity of the primers in lab conditions in order to make a 2nd selection. (3) The selected barcodes/primers were then tested again *in vitro* in order to assess their ability to assess differences of abundance, using more complex mock communities with a larger set of strains in variable DNA amplicon concentrations. (4) To complement this, the assignation power of the amplicons obtained from this 2nd selection of primers/barcodes was assessed *in silico.* The results of these different analyses will enable to provide advices for future use of DNA metabarcoding for phytoplankton in environmental conditions.

## 2 Materials and methods

### 2.1 Reference libraries

Two reference libraries were built for the 16S rRNA and 23S rRNA genes respectively (later simplified as 16S and 23S), collecting phytoplankton sequences from public reference libraries (Silva_138.1–; Quast et al., 2013; PhytoRef – del Campo *et al.*, 2018; PR2–Guillou *et al.*, 2013; µgreen-db – Djemiel *et al.*, 2020). These reference libraries are available at the shiny application Phytool v2 (Canino *et al.*, 2021) at the following link: https://github-carrtel.shinyapps.io/phytool_v2/.

For each barcode, a curation procedure was carried out, based on sequences similarity and taxonomic homogeneity (see details in Phytool v2, Canino *et al.*, 2021). To enable comparisons through different barcodes (*e.g.* 16S *vs* 23S) or methods (*e.g.* molecular *vs* microscopy), Phytool v2 was also used to harmonize the taxonomy from the different public reference libraries.

### 2.2 Genes and primer pairs tested

Primer pairs from 16S and 23S genes were first selected regarding their ability to produce amplicons with a length that fits with the MiSeq (Illumina) technology (2*250 bp to 2*300 bp). Thus, *in silico* were conducted on (1) primer pairs gathered from the literature (Tab. 1), (2) new combinations of known primers and (3) newly designed primers considering phytoplankton species and clades that were sequenced since the publication of the primers given in Table 1 for 16S and 23S genes.

The methodology to design the new primers was carried out as follows for each genetic marker (16S, 23S): (1) a single sequence was kept from each species to avoid the over-representation of taxa abundant in reference libraries. (2) A random sampling without replacement was performed on all these sequences to produce multiple clusters of few sequences (*e.g.* 100 clusters of 10 sequences for 1000 initial sequences). (3) For each cluster, a multiple alignment was carried out with the msa package (Bodenhofer *et al.*, 2015) using Muscle (Edgar, 2004) with default settings. (4) Conserved regions were checked and the most redundant were selected with BioPython (Cock *et al.*, 2009). From these multiple alignments the variability of the different regions within the genes were evaluated with Shannon index based on nucleotide diversity calculated at each base position. Regions which appear to be the most redundant through the different clusters were kept as candidate regions for primers. Additionally, a special attention was given to the followings: primers length ranging from 18 to 23 base pairs; GC percentage between 40 and 60% to ensure efficient polymerisation; presence of GC in the 5 first and last nucleotides to ensure good fixation; low number of ambiguous bases; resulting amplicons size suited for MiSeq sequencing. (5) Then, candidate primers were tested for *in silico* PCR according to the methodology given in the following paragraph ("*In silico* tests").

**Table 1.** Primer pairs tested *in silico*. Combinations not tested before this study are indicated with a star (*).

| Gene marker | Primer pairs | Targeted region | Mean amplicon size (pb) | References for primers definition | References for primers tests |
|---|---|---|---|---|---|
| 16S | 341F/805R | v3-v4 | 443 | Herlemann *et al.* (2011) | Eiler *et al.* (2013); Bennke *et al.* (2018) |
| | CYA359F/CYA781Rd | v3-v4 | 425 | Nübel *et al.* (1997) | Costa *et al.* (2016) ; Ivanova *et al.* (2019) |
| | CYA359F/805R | v3-v4 | 425 | Nübel *et al.* (1997) (F); Herlemann *et al.* (2011) (R) | * |
| | PLA491F/805R | v4 | 315 | Füller *et al.* (2006) (F) ; Herlemann *et al.* (2011) (R) | * |
| | 515F/926R | v4-v5 | 413 | Caporaso *et al.* (2011), modified by Parada *et al.*, 2016 | Parada *et al.*, 2016; Watanabe *et al.* (2001) |
| | CYA781Rd/E1115R | v5-v6 | 336 | Nübel *et al.* (1997) (F); Reysenbach & Pace, 1995 (R) | * |
| | ECLA16S_F1/ECLA16S_R1 | v5-v6 | 386 | *This study* | * |
| | E939F/OXY1313R | v6-v7 | 417 | Rudi *et al.* (1997); West *et al.* (2001) | * |
| 23S | p23SrV_f1/p23SrV_r1 | v5 | 408 | Sherwood & Presting, 2007 | Craine *et al.* (2018); Brown *et al.* (2022) |
| | A23SrV_F1/A23SrV_R1 | v5 | 411 | Yoon *et al.* (2016) | * |
| | A23SrV_F2/A23SrV_R2 | v5 | 405 | Yoon *et al.* (2016) | * |
| | ECLA23S_F1/ECLA23S_R1 | v5 | 402 | *This study* | * |
| | ECLA23S_F2/ECLA23S_R2 | v5 | 408 | *This study* | * |

The selection of the primer pairs is listed in Table 1, and the corresponding oligonucleotides sequences are available in Supplementary data 1.

### 2.3 *In silico* tests

The *in silico* tests of the primers rely on the different approaches are detailed hereafter.

The main characteristics of the primers′ pairs were first assessed using OligoAnalyzer (Integrated DNA Technologies OligoAnalyzer: http://scitools.idtdna.com/analyzer/Applications/OligoAnalyzer/).

The following criteria were evaluated with this tool for each pair: the mean amplicon size; the difference between forward and reverse primers mean melting temperature ($\Delta T_m = |T_{m\_FORWARD} - T_{m\_REVERSE}|$; the smaller the difference, the higher the PCR efficiency is supposed to be); the oligonucleotide structure of the primers and their predicted behaviour. Potential problematic structures could be: «Hairpin» (folding of the oligonucleotide on itself which is likely to occur around or above the melting temperature which would decrease its ability to fix to its target and thus inhibit the PCR amplification); «Homodimer» (one oligonucleotide is likely to hybridize with itself and produce a dimer, resulting in a reduction of the amplification efficiency); «Heterodimer» (the two oligonucleotides are likely to hybridize together).

The amplification efficiency was assessed *in silico* on sequences of microalgal species, allowing a maximum of 2 mismatches between the amplicon and its target sequence, following indications by Nossa *et al.* (2010) and Klindworth *et al.*, (2013). This was done using *pcr.seqs* command from Mothur (Schloss *et al.*, 2009). This enabled to assess the number of sequences that matched perfectly with the primer pair (no mismatches) and the number of sequences that are likely to be amplified by a standard PCR (maximum 2 mismatches allowed) out of the total number of reference sequences used from the corresponding reference library. In complement, an estimation of the resolution of each barcode was assessed by the proportion of the number of species with strictly different barcodes among the total number of species amplified *in silico* (maximum 2 mismatches allowed). The variability for undefined species (sp.) was not considered in this assessment.

The specificity of the primer pairs for microalgae at the expense of heterotrophic bacteria was assessed.

For 23S, an *in silico* PCR was conducted with candidate primer pairs using *pcr.seqs* (Mothur) on 183 976 sequences of 23S for heterotrophic bacteria downloaded from SILVA 138.1. For 16S, as the number of available sequences was much higher, we used a less time-consuming approach. The estimation was made using a ratio of the proportion of microalgae species amplified *in silico* using *pcr.seqs* command (Mothur, Schloss *et al.*, 2009) out of the proportion of heterotrophic bacteria amplified *in silico* using RDP Probe-Match tool (Cole *et al.*, 2014, available at http://rdp.cme.msu.edu/probematch/search.jsp), the higher the score, the higher the specificity.

In complement, the *in silico* amplicons were re-assigned using their corresponding reference library to evaluate the number of correct re-assignments. This was investigated for each taxonomic rank using two assignation methods: the DADA2 function *assignTaxonomy* (with a boostrap value of 80) and the Mothur function *classify.seqs* (with the same boostrap value and 10,000 iterations).

Based on all these tests, we performed a 1st selection of the most efficient primer pairs.

**Table 2.** Microalgal strains used to create mock communities Mock 1 and Mock 2.

| Class | Order | Species | Strain # | Mock 1 | Mock 2 |
|---|---|---|---|---|---|
| Cyanophyceae | Nostocales | *Dolichospermum flosaquae* | TCC79 | x | x |
| | Oscillatoriales | *Planktothrix rubescens* | TCC14 | | x |
| | Chroococcales | *Microcystis aeruginosa* | TCC80 | | x |
| | Synechococcales | *Anathece clathrata* | TCC300 | | x |
| Zygnematophyceae | Desmidiales | *Cosmarium regnellii* | TCC56 | x | x |
| | Zygnematales | *Mougeotia sp.* | TCC814 | x | x |
| | | *Zygnema sp.* | TCC815 | | x |
| Chlorophyceae | Chlamydomonadales | *Haematococcus lacustris* | TCC8 | | x |
| | | *Chlamydomonas reinhardtii* | TCC234-2 | x | |
| | | *Pandorina morum* | TCC9 | | x |
| | Sphaeropleales | *Tetradesmus obliquus* | TCC116 | x | x |
| Trebouxiophyceae | Trebouxiales | *Botryococcus braunii* | TCC57 | x | x |
| | Prasiolales | *Stichococcus bacillaris* | TC145-1 | x | |
| Bacillariophyceae | Bacillariales | *Nitzschia palea* | TCC139-1 | | x |
| | Fragilariales | *Staurosira venter* | TCC691 | | x |
| | Licmophorales | *Ulnaria acus* | TCC365 | | x |
| | Tabellariales | *Asterionella formosa* | TCC362 | x | x |
| Mediophyceae | Stephanodiscales | *Cyclotella meneghiniana* | TCC640 | x | |
| Cryptophyceae | Cryptomonadales | *Cryptomonas sp.* | TCC826 | | x |
| Eustigmatophyceae | Eustigmatales | *Vischeria magna* | TCC345 | | x |
| Xanthophyceae | Tribonematales | *Xanthonema montanum* | TCC165 | x | x |

**Table 3.** PCR program used to amplify the different barcodes.

| Temperature (°C) | Duration | Cycles (nb) | Steps |
|---|---|---|---|
| 95 | 3′ | 1 | Activation / initial denaturation |
| 95 | 30 | 30 | Denaturation |
| 58 | 30″ | | Hybridisation |
| 72 | 30″ | | Elongation |
| 72 | 5′ | 1 | Final elongation |

## 2.4 *In vitro* experiments

### 2.4.1 Mock communities

The primer′ pairs resulting from the *in silico* selection step, were compared in two *in vitro* experiments with mock communities composed of a mix of DNA from known pure algal cultures from the TCC culture collection (Rimet *et al.*, 2018a). Two mock communities (Mock 1 and Mock 2) were designed to cover a large taxonomic diversity of phytoplankton, including species from Eubacteria, Plantae and Chromista kingdoms (Tab. 2)

Mock 1 was made as an equimolar mix of DNA from 10 different microalgal strains (Tab. 2). DNA extractions were performed on 2 mL of each TCC strain culture following the GenElute protocol in Kermarrec *et al.* (2013). DNA concentrations were quantified using a NanoDrop spectrophotometer (Thermo Scientific) and dilutions were made to prepare aliquots with a similar DNA concentration of 5 ng $\mu L^{-1}$. Then PCRs were carried out in triplicates (Tab. 3).

Mock 2 was made from a larger taxonomic basis, including 18 microalgal strains (Tab. 2) at various DNA concentrations to mimic environmental conditions and to evaluate how these variations can be assessed by the different primer pairs. To reduce bias that could raise from the TCC non-axenic strains (*e.g.* potential presence of DNA from non-target microorganisms, as associated bacteria), Mock 2 was made as a mixture of amplicons from each of the 18 strains. Moreover, in anticipation of later routine implementation, an automated DNA extraction method was compared to GenElute DNA extraction used for Mock 1 (Kermarrec *et al.*, 2013). This automated DNA extraction was performed using the MagnetaPure 32 automated system (Dutscher) and the NucleoMag Microbiome extraction kit (Macherey-Nagel) following the manufacturer′s instructions. DNA extract from each culture was then amplified by PCR, 2 µL of the amplification product was controlled on agarose gel (1%), and the remaining volume of amplicon (23 µL) was purified using Illustra GFX PCR DNA kit (Cytiva). DNA concentrations of the purified amplicons were measured with Qubit 4 Fluorometer (Invitrogen) in order to build three types of mock communities with controlled amplicon concentrations (Supplementary data 2). Mock 2a was an «equimolar» mix (20 ng) of the DNA amplicon from each species. Mock 2b was a linear «gradient» mix with DNA quantities ranging from 5 to 50 ng of DNA

amplicon. Mock 2c was an «exponential» mix with two highly-represented species (50 ng), 8 low-concentration species (1 ng) and 8 species mimicking rare ones (0.1 ng). Finally, with two DNA extraction methods, two primer pairs and three types of mock, 12 mock samples were prepared for Mock 2.

#### 2.4.2 First *in vitro* metabarcoding experiment

This *in vitro* experiment was conducted on the primer pairs selected from the *in silico* test. Their ability to detect species along a wide taxonomic diversity and their amplification efficiency were tested performing metabarcoding on Mock 1. First, the selected primer pairs were used to amplify Mock 1, second, amplicons were sequenced, and then, the obtained community composition was compared to the expected composition of Mock 1. To that aim, Illumina libraries were prepared in a dual-step PCR approach. Briefly, forward and reverse primers were tailed with the 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′ and the 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3′ overhanging adapters to proceed to the first PCR step (PCR1) where Mock 1 was amplified in triplicates in a final volume of 25 µL using 5 µL of each forward and reverse tailed primers (1 µM), 12.5 µL of 2× KAPA HiFi HotStart Ready Mix (Roche), and 2.5 µL of DNA template of Mock 1 (5 ng µL$^{-1}$). The PCR program is given in Table 3. A common and optimal melting temperature of 58 °C was selected for all primer pairs.

Three replicates of PCR were carried out and sequenced separately. PCR products were then checked on an agarose gel and sent to the sequencing platform (PGTB, France) using sealed plates (Thermo Scientific sealed with PCR Seal 4TITUDE). Since all amplicons from the different primer pairs are close in length (see Tab. 1), not exceeding 50 bp differences, they were all gathered on the same sequencing run. The PGTB platform performed a second PCR (PCR2) amplification using the purified PCR1 amplicons as template and the Illumina-tailed primers to add dual-index specific to the samples. The final pool corresponding to an equimolar mix of all the PCR2 dual-indexed amplicons was sequenced using a NanoMiseq (Illumina) run with kit v2 (2*250 bp).

#### 2.4.3 Second *in vitro* metabarcoding experiment

This experiment was conducted on the primer pairs selected from the 1st *in vitro* metabarcoding experiment. Their ability to detect species among a wider and different taxonomic diversity, a higher richness, even when species proportions are unequal, and their amplification efficiency were tested using the different samples of Mock 2. Following the same dual-step PCR approach as in the first experiment, the primer pairs were used to amplified Mock 2a, 2b and 2c. In that experiment, the final pool was sequenced at the PGTB platform using a MiSeq (Illumina) with kit v3 (2*250 bp) in order to get a higher sequencing depth than the 1st experiment to better unravel rare species.

#### 2.4.4 Bioinformatic analyses

Primers were checked and removed from demultiplexed FASTQ reads using Cutadapt v3.5 (Martin, 2011). Only reads with no error in primer sequences were kept (max-n 0). Next steps for the read quality filtering, merging and determination of Amplicon Sequence Variants (ASVs) were done using the DADA2 package v1.20 (Callahan *et al.*, 2016) on R v4.1 (R Core Team, 2018). Briefly, reads were filtered using the function *filterAndTrim* with parameters set as default (*i.e.* maxEE = c(2,2), truncQ = 2, maxN = 0) and by only keeping the first 220 and 200 bases for the forward and reverse sequences respectively, due to their expected lower quality at the end of the read (Schirmer *et al.*, 2015). Reads were then merged with default parameters of *mergePairs* function, and chimera were detected *de novo* and removed using the function *removeBimeraDenovo* with the consensus method. Sequences were then size-filtered by keeping sequences with a size comprised between 245 and 270 bp, to avoid aspecific ASVs. Taxonomic assignment of each ASV was done with the RDP (Ribosomal Database Project) naive Bayesian classifier (Wang *et al.*, 2007) implemented in DADA2, using Phytool v2 database (Canino *et al.*, 2021) as a reference. In parallel, a blast search of each ASV sequence was done on the same database as well as against all NCBI nucleotide database using blast 2.13 (Altschul *et al.*, 2009). The combination of these three taxonomic assignations permitted an efficient recovery of almost all species that compose the Mock 2 community.

Bioinformatic pipelines are available at https://github.com/Github-Carrtel/2023_Canino_MBMG

#### 2.4.5 Taxonomic assignation test of *in silico* amplicons

To assess the efficiency of the barcodes amplified by the selected primer pairs, taxonomic assignation was processed for the amplicons obtained *in silico*. To this end, 16S sequences were downloaded from the Phytool v2 reference library and cut in order to keep only the short portions corresponding to the barcodes amplified by the primers (*in silico* amplicons). These *in silico* amplicons were then reassigned to the Phytool v2 reference library using DADA2 (command *assignTaxonomy, minBoot* = 75) and Mothur (command *classify.seqs*, same bootstrap value 75 and 10,000 iterations).

## 3 Results

### 3.1 *In silico* evaluation of the primer pairs

Primer pairs *in silico* evaluations are given in Table 4. The 16S reference library used for this step gathered 8479 sequences, representing 1 675 species, and the 23S reference library gathered 1995 sequences, representing 913 species. For both 16S and 23S, these sequences covered all microalgal phyla (Bacillariophyta, Charophyta, Chlorophyta, Chromista, Cryptophyta, Cyanobacteria, Glaucophyta, Haptophyta, Miozoa, Ochrophyta, Prasinodermatophyta, Rhodophyta).

Based on these results, a primer selection was carried out. For 16S, all tested primers presented high resolution and were able to amplify a high number of targeted taxa according to *in silico* PCR results. The primer pair CYA359F/CYA781Rd (targeting v3-v4 region) was selected based on its low ΔTm, high specificity (43.64) for phytoplankton and high proportion of amplified targets. Despite its low specificity (4.4), a second primer pair (ECLA16S_F1/ECLA16S_R1) was selected for 16S, as it targets a different variable region (v5-v6) and presented good characteristics (low $\Delta T_m$, no structure warning

**Table 4.** Results of *in silico* tests for the 16S and 23S primers. $\Delta T_m$: melting temperature difference between forward and reverse. Oligonucleotide structure warnings: warnings obtained with OligoAnalyzer. *In silico* PCR output: number of sequences that perfectly match with primers (in brackets) followed by the number of sequences likely to be amplified out of the total number of reference sequences. Estimated resolution: percentage of species with strictly different barcodes (in brackets are given the exact numbers of species with different barcodes out of the total number of species). Estimated specificity: for 16S it is the proportion of phytoplankton species out of the proportion of heterotrophic bacteria amplified *in silico*; for 23S it is given as the result of *in silico* PCR done on heterotrophic bacteria sequences (*i.e.* « *in silico* PCR output »).

| Gene marker | Primer pairs | Δ $T_m$ (°C) | Oligonucleotide structure warnings | *In silico* PCR output | Estimated resolution (%) | Estimated specificity |
|---|---|---|---|---|---|---|
| 16S | 341F/805R | 8.8 | 2 Homodimers (F)<br>1 Heterodimer | [6613]<br>7305/8479 | 83.57<br>1007/1205 | 0.97 |
| | CYA359F/CYA781Rd | 2.4 | 1 Heterodimer | [5819]<br>7119/8479 | 85.84<br>1115/1299 | 43.64 |
| | CYA359F/805R | 6.6 | 2 Homodimers (R) | [5887]<br>7260/8479 | 87.07<br>1178/1353 | 34.48 |
| | PLA491F/805R | 0.7 | 2 Homodimers (R) | [2950]<br>6356/8479 | 79.91<br>907/1135 | 42.16 |
| | 515F/926R | 14.7 | 2 Homodimers (F)<br>1 Heterodimer | [7717]<br>8076/8479 | 86.16<br>1357/1575 | 0.98 |
| | CYA781Rd/E1115R | 0.6 | 1 Homodimer (R) | [6001]<br>6637/8479 | 78.52<br>1060/1350 | 39.82 |
| | ECLA16S_F1/ECLA16S_R1 | 2 | None | [7438]<br>8075/8479 | 83.12<br>1290/1552 | 4.40 |
| | 926F/OXY1313R | 11.7 | None | [522]<br>6899/8479 | 86.12<br>1092/1268 | 18.76 |
| 23S | p23SrV_f1/p23SrV_r1 | 1.7 | None | [1608]<br>1763/1995 | 90.99<br>727/799 | [1]<br>526/183976 |
| | A23SrV_F1/A23SrV_R1 | 0.5 | None | [1608]<br>1775/1995 | 91.56<br>738/806 | [1]<br>28725/183976 |
| | A23SrV_F2/A23SrV_R2 | 2 | None | [1599]<br>1759/1995 | 91.33<br>727/796 | [1]<br>1215/183976 |
| | ECLA23S_F1/ECLA23S_R1 | 2.6 | 1 Homodimer (F) | [1727]<br>1797/1995 | 91.36<br>751/822 | [5]<br>853/183976 |
| | ECLA23S_F2/ECLA23S_R2 | 0.7 | None | [1672]<br>1809/1995 | 91.49<br>764/835 | [1]<br>870/183976 |

and a higher number of amplified targets). For 23S, all primer pairs showed satisfactory results in the *in silico* investigations among which three pairs (p23SrV_f1/p23SrV_r1, A23SrV_F1/A23SrV_R1 and A23SrV_F2/A23SrV_R2) showed comparable results in terms of amplification efficiency and resolution power. The primer pair p23SrV_f1/p23SrV_r1 was selected as it showed a higher specificity to phytoplankton organisms and had already been widely used in former environmental studies. In addition, we selected ECLA23S_F1/ECLA23S_R1 and ECLA23S_F2/ECLA23S_R2 due to their high number of perfect matches with target sequences (*in silico* PCR output of 1727 and 1672 respectively), good specificity toward microalgae organisms, combined to high resolution. At the end, two primer pairs for 16S and three for 23S were kept for further *in vitro* lab test.

In order to avoid amplification issues in PCR1 due to the addition of adapters to the primers, as required by the sequencing facility (for them to add Illumina adapters in a PCR2 step), complementary *in silico* investigations were performed prior to the *in vitro* evaluation steps. Checking the behaviour of the five selected pairs to which the adapters were added, no major issue was observed (Tab. 5), except some heterodimer structures which were more likely to occur for some pairs. However, since homodimers which may reduce amplification efficiency were not suspected to occur, all five pairs were kept for following *in vitro* tests.

## 3.2 In vitro tests results

### 3.2.1 *In vitro* experiment with Mock 1

Table 6 shows the detection results of the five primer pairs for each of the 10 microalgal strains in Mock 1. In some cases, the taxonomic assignation of ASV with DADA2 matched correctly with the target species, while in others it did not. In such cases, the taxonomic assignation was completed manually by recognizing the highest taxonomical rank at which the target species belongs based on DADA2 and Blast results (Tab. 6). Finally, all control-species could be detected with the three 23S primer pairs, while 3 species remained undetected with the 16S primer pairs.

**Table 5.** Results of *in silico* tests for the 16S and 23S primers including adapters from the sequencing platform: $\Delta T_m$ is the melting temperature difference between forward and reverse; Oligonucleotide structure warnings.

| Gene marker | Primer pairs | Δ Tm (°C) | Oligonucleotide structure warnings |
|---|---|---|---|
| 16S | CYA359F/CYA781R | 1.1 | 1 Heterodimer |
| | ECLA16S_F1/ECLA16S_R1 | 2 | 1 Heterodimer |
| 23S | p23SrV_f1/p23SrV_r1 | 0.6 | 1 Heterodimer |
| | ECLA23S_F1/ECLA23S_R1 | 0.5 | none |
| | ECLA23S_F2/ECLA23S_R2 | 1 | none |

**Table 6.** Detection of the 10 different control-species in Mock 1 with 16S and 23S primers. **x:** ASV assigned to the target species with DADA2; ***:** ASV manually assigned to the target species using Blast, *nd*: not detected.

| Mock 1 | 16S | 23S | | | |
|---|---|---|---|---|---|
| | CYA359F/ CYA781Rd | ECLA16S_F1/ ECLA16S_R1 | p23SrV_f1/ p23SrV_r1 | ECLA23S_F1/ ECLA23S_R1 | ECLA23S_F2/ ECLA23S_R2 |
| *Asterionella formosa* | **x** | **x** | **x** | **x** | **x** |
| *Botryococcus braunii* | **x** | **x** | **x** | **x** | **x** |
| *Chlamydomonas reinhardtii* | *nd* | *nd* | * | * | * |
| *Cosmarium regnellii* | *nd* | * | * | * | * |
| *Cyclotella meneghiniana* | * | * | * | * | * |
| *Dolichospermum flosaquae* | * | * | * | * | * |
| *Mougeotia sp.* | *nd* | *nd* | * | * | * |
| *Stichococcus bacillaris* | **x** | *nd* | * | * | * |
| *Tetradesmus obliquus* | **x** | **x** | * | * | * |
| *Xanthonema montanum* | * | * | * | * | * |

Complementary, an overview of the percentage of ASV and reads assigned to control-species in Mock 1 for each primer pair is presented in Figures 1a and 1b. These results assess the primers specificity for microalgae. The CYA359F/ CYA781Rd pair has a very good specificity to microalgae since no ASV were matching other organism but microalgae. This was not the case for ECLA16S_F1/ ECLA16S_R1 which had a low specificity to microalgae since only 27% of the reads, corresponding to 23% of the ASV, matched to microalgae. Good efficiency was observed for 23S, with 75% of the ASV assigned to target species for ECLA23S_F1/ ECLA23S_R1, and slightly less for ECLA23S_F2/ ECLA23S_R2 and p23SrV_f1/p23SrV_r1 (52%). The specificities were high when considering the number of reads: 96% for ECLA23S_F1/ECLA23S_R1 and ECLA23S_F2/ ECLA23S_R2, and 93% for p23SrV_f1/p23SrV_r1. The non-target organisms which were detected with 16S and 23S primer pairs were heterotrophic bacteria (*e.g.* alphaproteobacteria, betaproteobacteria, uncultured bacteria, Pseudomonadales).

Based on these results, one primer pair was selected for each barcode for further evaluation. For 16S, CYA359F/ CYA781Rd outperformed the specificity and efficiency of the other tested pair. For 23S, all three primer pairs showed quite similar results in this first *in vitro* test. However, ECLA23S_F1/ECLA23S_R1 was selected based on its slightly better *in silico* performance, with 1727 sequences that perfectly match with primers (Tab. 3).

### 3.2.2 *in vitro* experiment with Mock 2

The metabarcoding results for the two selected primer pairs on Mock 2 communities (2a: equimolar; 2b: gradient; and 2c: exponential) are presented in Figure 2.

Among the 18 control-species of the Mock 2a sample (Fig. 2), all species were detected with 23S while the Chlorophyceae *Haematococcus lacustris* could not be detected with 16S. Although some deviation from the expected proportion can be observed for 23S (*Cosmarium regnellii* and *Planktothrix rubescens* with GenElute DNA extraction; *Pandorina morum* with the Automate DNA extraction), the equimolar proportion is well represented by read numbers. For 16S, proportions were more variable, with an often-higher number of reads when DNA was extracted with the automate.

When considering Mock 2b (gradient concentration, Fig. 2), with DNA Automate extraction, the correlation between DNA concentrations and the number of reads obtained for control-species was higher for the 23S primer pair ($R^2 = 0.88$) than for the 16S one ($R^2 = 0.73$). With the GenElute extraction, the tendency was reversed ($R^2 = 0.61$ for 23S; $R^2 = 0.77$ for 16S). However, all correlations were high and significant.

When considering Mock 2c (Fig. 2) all species could be detected with 23S primer pairs, whatever their initial DNA concentration, and the exponential proportions were well represented by read numbers. For 16S primer pairs, in addition to *H. lacustris*, two species present at the lowest DNA

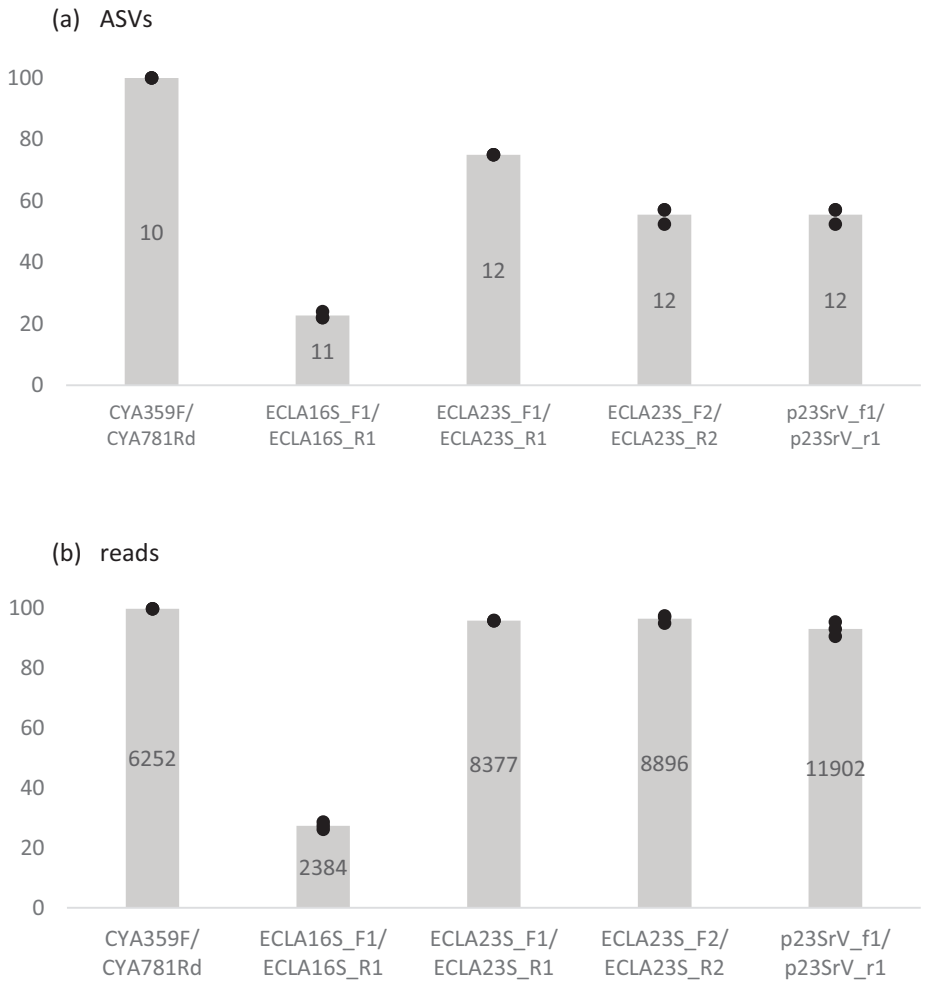


**Fig. 1.** Percentage of ASV (a) and reads (b) assigned to control-species in Mock 1. Grey bars give the average percentage of ASV (or reads) with the number of ASV (or reads) written inside. Black dots give the percentage of ASV (or reads) assigned to control-species for the three replicates taken separately.

concentration (0.1 $\mu g\,L^{-1}$) could not be detected (*Mougeotia sp*. and *Zygnema sp*.). The two species at the highest concentration (50 $\mu g\,L^{-1}$) were represented by 84.3% of the reads for 23S and 93.6% for 16S, while their DNA was expected to represent 91.2% of the DNA mix. The nine species at the medium concentration (1 $\mu g\,L^{-1}$) were represented by 13.0% of the reads for 23S and 5.6% for 16S, while their DNA was expected to represent 8.2% of the DNA mix.

### 3.2.3 Resolutive power of *in silico* amplicons

Results of the taxonomic assignation of *in silico* amplicons are given in Table 7 for the CYA359F/CYA781Rd and ECLA23S_F1/ECLA23S_R1 primer pairs on 16S and 23S reference libraries, respectively. Whatever the command used for taxonomic assignation (*i.e.* assignTaxonomy in DADA2 and classify.seqs in Mothur), results were rather similar. The percentages of amplicons correctly assigned from Kingdom to Family levels were comparable for both primer pairs. However, at lower taxonomic levels (Genus and Species levels), assigned percentages were higher for the 23S primer pair than for the 16S one.

## 4 Discussion

This study followed a four-step approach with successive steps of primer pair evaluation done either *in silico* or *in vitro* using, respectively curated reference libraries and mock communities. This approach allowed us to converge to a small subset of primer pairs with good efficiency. The first *in silico* evaluation allowed to successfully select a subset of five primer pairs that well amplified mock samples in following *in vitro* steps. The first *in vitro* step, enabled to retain only two primer pairs being efficient to amplify most of the species in the simplified community of Mock 1, before focusing on the Mock 2 samples. Those more complex Mock 2 samples allowed to test the 2 primer pairs on a larger taxonomical diversity and to evaluate their ability to reveal relative abundances. Finally, the last *in silico* step aimed at evaluating de taxonomic resolution of the two primer pairs on their respective reference libraries, thus exploring a larger microalgal diversity than in the mock communities. This four-step approach appears to be time- and cost-efficient by minimizing wetlab and sequencing efforts on five and two pairs (out of 13), at steps 2 (Mock 1) and 3 (Mock 2) respectively.

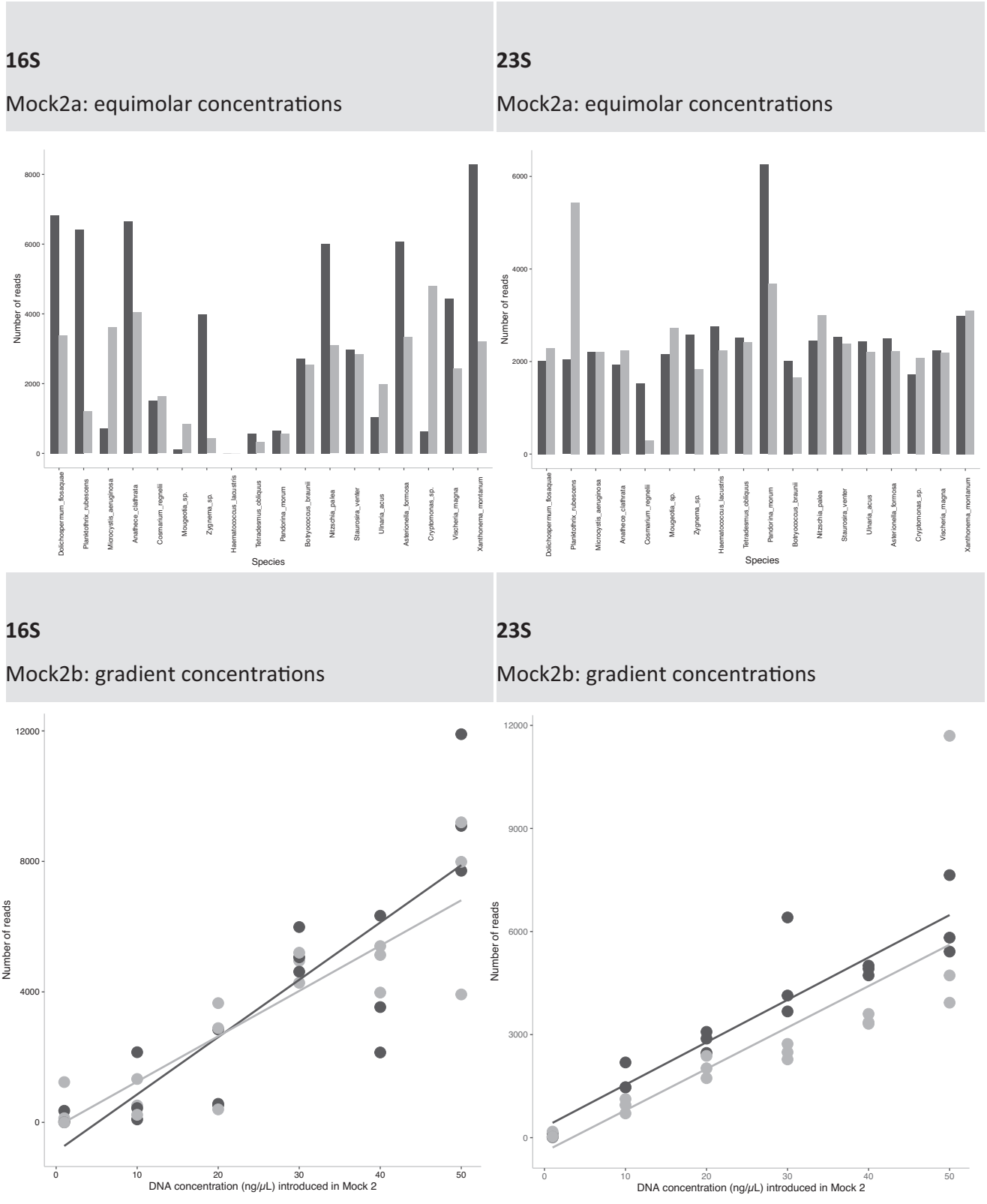


**Fig. 2.** Metabarcoding results for the 16S primer pair CYA359F/CYA781Rd (left) and for the 23S primer pair ECLA23S_F1/ECLA23S_R1 (right) on Mock 2a, 2b, and 2c for DNA extracted with automate (dark grey) and with GenElute (light grey). Mock 2a: number of reads obtained per control-species. Mock 2b: number of reads obtained per amplicon DNA initial concentration.

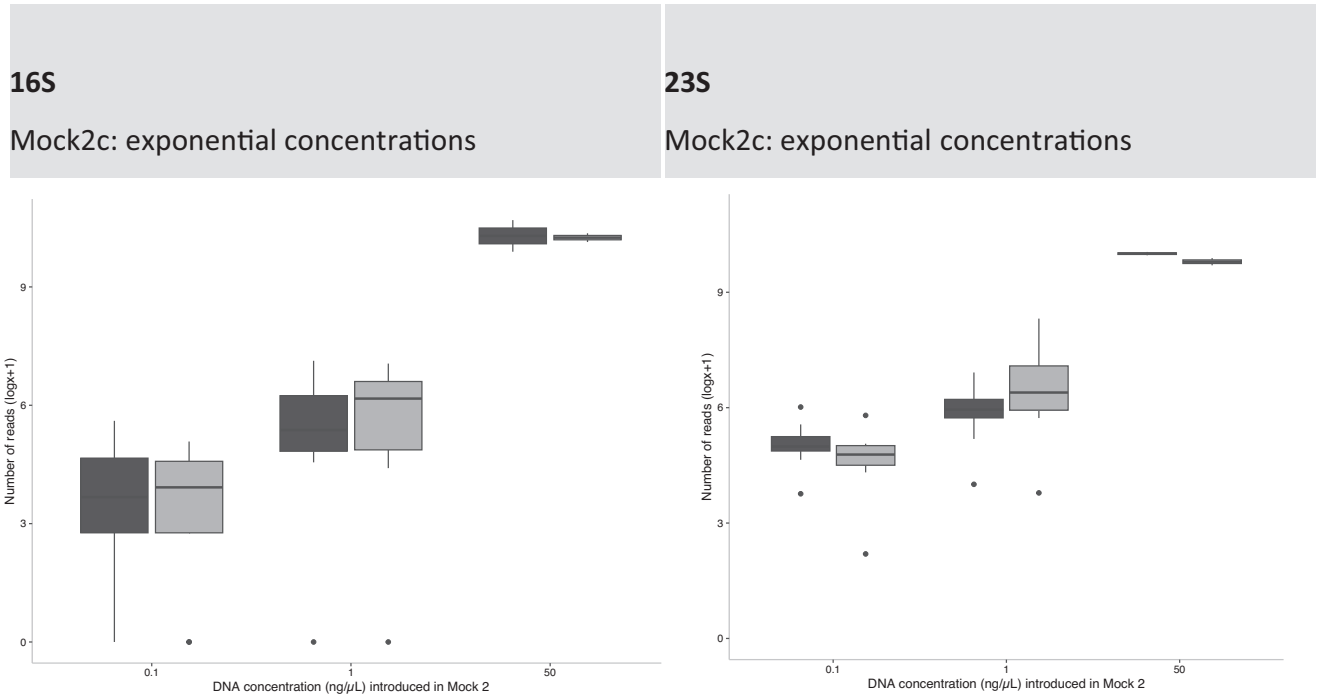


**Fig. 2.** (Continued). Mock 2c: box-plot of the number of reads obtained per amplicon DNA initial concentration.

**Table 7.** Percentages of *in silico* amplicons correctly assigned with Mothur and DADA2 from Kingdom to species levels: for CYA359F/CYA781Rd primer pair with the 16S reference library; for ECLA23S_F1/ECLA23S_R1 primer pair with the 23S reference library.

| Taxonomic ranks | **16S:** CYA359F/CYA781Rd | | **23S:** ECLA23S_F1/ECLA23S_R1 | |
|---|---|---|---|---|
| | DADA2 | Mothur | DADA2 | Mothur |
| Kingdom | 99.62% | 99.63% | 99.83% | 99.72% |
| Phylum | 99.30% | 99.30% | 99.61% | 99.50% |
| Class | 99.13% | 99.13% | 99.39% | 99.33% |
| Order | 94.79% | 94.82% | 98.83% | 98.83% |
| Family | 88.43% | 88.59% | 95.66% | 95.83% |
| Genus | 77.20% | 77.31% | 90.26% | 90.43% |
| Species | 54.91% | 54.62% | 79.24% | 79.91% |

## 4.1 Advantages and disadvantages of the tested barcodes and primer pairs

### 4.1.1 Amplification efficiency and specificity to phytoplankton

The five primer pairs selected from the *in silico* tests all showed good amplification potential and high specificities to eukaryotic and prokaryotic microalgae (cyanobacteria). However, one of them, the ECLA16S_F1/ECLA16S_R1 pair, showed a lower specificity in Mock 1. This mock community was made prior PCR amplification, as a pool of DNA extracted from strains that were non-axenic cultures and may thus include DNA from both target and non-target taxa. The lower specificity of this pair to eubacteria revealed the presence of heterotrophic bacteria (*e.g.* Proteobacteria, Pseudomonadales) associated to the algal strains and which DNA was thus co-extracted. In parallel, this revealed the good specificity of the four other pairs tested on Mock 1. For the two finally selected primer pairs (CYA359F/CYA781Rd, ECLA23S_F1/ECLA23S_R1 respectively for 16S and 23S), the proportion of sequences that could be amplified *in silico* was very high (84% and 90% for 16S and 23S, respectively). The good amplification efficiency evaluated *in silico* was confirmed *in vitro* with 70% and 100% of the 10 species of Mock 1 detected by 16S and 23S pairs, respectively. This trend was similar for Mock 2, with 83% and 100% of the 18 species, for 16S and 23S pairs respectively. Although both selected pairs (CYA359F/CYA781R, ECLA23S_F1/ECLA23S_R1) showed good efficiency to amplify phytoplankton taxa, it is notable that the 23S barcode on the v5 region performed systematically better than the 16S one, detecting more species from the mocks. Finally, we warn the future users of these primers who would like to work specifically on microalgae, that the specificity of these primers was tested against heterotrophic bacteria, a group that can be dominant in plankton or aquatic biofilm samples. If samples would contain significant biomass of higher plant tissues (*Pinus, Solanum* ...) alongside algae, then these higher plant tissues would be co-amplified with algae (further tests on Primer-Blast tool in NCBI

for ECLA23S_F1/ECLA23S_R1 showed good affinity to higher plants alongside microalgae).

#### 4.1.2 Taxonomic resolution

From the last *in silico* step, the 23S pair clearly appears to be the most resolutive with ∼90% of the *in silico* amplicons obtained from the reference library that could be assigned to genus level, and ∼80% to species level. The 16S pair performed less well (77% and 55% respectively for genus and species level, respectively).

#### 4.1.3 Capacity to assess relative abundances

For Mock 1, it was not possible to assess the ability of tested pairs to assess the relative abundance of the target taxa. Indeed, although it was made of equimolar DNA quantities, the DNA quantification was not a good proxy since DNA came from both the target strains and non-target associated microorganisms (*e.g.* heterotrophic bacteria) present in the cultures. From the *in vitro* test on Mock 2a (equimolar) and Mock 2b (gradient) it was possible to evaluate the potential of the barcodes and pairs for assessing taxa relative abundances. When DNA from all species were present in equimolar proportions (Mock 2a), the number of reads obtained was similar for all species with the 23S pair, while it was more variable from one species to the other with 16S. When the proportion between taxa was unequal (Mock 2b), the DNA quantities introduced in the mock and the number of reads were significantly correlated. Relative abundances were better assessed with the 23S pair than with the 16S one. Finally, Mock 2c (exponential) enabled to assess the ability for detecting low-abundant taxa (*e.g.* taxa at a DNA concentration 500 times lower than others). For 16S, three out of the 7 less abundant taxa could not be detected. The 23S pair performed better as all taxa were detected. As in environmental communities, taxa are often present in unequal proportions, the 23S barcode appears to be more promising to decipher microalgae (eukaryotic algae and cyanobacteria) assemblages. Moreover, when the ultimate goal of metabarcoding is ecological quality assessment, presence-absence data are often not sufficient (*e.g.* IPLAC index – Laplace-Treyture et Feret, 2016). Thus, the 23S pair shows good potential to both identify phytoplankton taxa and assess their relative abundances.

#### 4.1.4 Completeness of reference libraries

The reference libraries used in this study have contrasted completeness depending on the barcode with 8 479 sequences for 16S and about four-time less for 23S (1 995 sequences). Although the 16S reference library had more sequences, most of them (nearly 3/4) are Cyanobacteria, while eukaryotic microalgae are poorly referenced. For 23S, more eukaryotic species (nearly 1/3) were represented proportionally. Reference libraries are clearly a pitfall of these two gene makers (16S, 23S), compared to others like 18S. The operationalisation of the metabarcoding approach for biomonitoring phytoplankton communities will not be possible without completing reference libraries as shown for other aquatic organisms (Weigand *et al.*, 2019).

### 4.2 A single barcode to decipher phytoplankton communities?

This study aimed at finding a single barcode to decipher microalgal assemblages in their entirety, especially to get simultaneously eubacteria and eukaryotic microalgae. Although the eukaryotic primers of 18S are largely used and advised for phytoplankton surveys (Jerney *et al.*, 2022), they do not allow cyanobacteria detection (*e.g.* Wang *et al.*, 2022), which requires the use of a complementary barcode (*e.g.* 16S). A single barcode ideally fitted to phytoplankton biomonitoring must meet several criteria: (1) not exclude any microalgal phyla or families, (2) be specific enough in excluding as much as possible non-microalgal diversity, (3) offer a high taxonomic resolution, (4) provide accurate taxa proportions in final inventories.

In our study, we wanted to ensure that taxa from most of the phyla could be amplified. All tested pairs for 16S and 23S did amplify both prokaryotic and eukaryotic algae. However, some taxa could not be detected by 16S although they were present in the tested mock communities. For 16S, all cyanobacteria could be well amplified, while two Chlorophyceae and one Eustigmatophyceae could not. 23S appears as the most efficient to amplify the whole microalgal diversity, without excluding any of the tested families. Moreover, 23S primers were specific enough to microalgae taxa, as only a limited number of reads of non-target taxa were obtained (4% of reads matched to heterotrophic bacteria). Obtaining a high taxonomic resolution is also important in order to get a good ecological quality assessment. Indeed, bioassessment metrics generally require identification of taxa at species (or genus) level as it is the case for several phytoplankton biomonitoring indices (*e.g.* IPLAC in Laplace-Treyture & Feret 2016, Brettum Index in Dokulil *et al.*, 2005). Although the 23S reference library is far from complete, it seems to provide a high enough taxonomic resolution from phyla to species level. The *in silico* evaluation toward the reference library is encouraging and bodes well for good resolution in environmental samples. Finally, mock communities with gradient and exponential proportions have shown the promising ability of 16S and, even better, of 23S to assess relative abundances and to detect rare taxa. However, as Mock 2 samples were made from DNA of purified amplicons, this still needs to be confirmed on DNA from environmental samples.

All these results point to 23S v5 having the potential to be a good gene marker to decipher microalgal and cyanobacterial diversity. Several primer pairs were tested *in silico* and *in vitro* for amplifying this v5 region. After the first selection step, three pairs amplifying this region were explored, two of them being designed in this study and the third one proposed by Sherwood and Presting (2007). With Mock 1 community, all three presented good potential which was not surprising as forward and reverse primers were all designed from the same gene region, with only little nucleotide changes from one to the other. Indeed, this region was also the one explored by Yoon *et al.* (2016) and Kang *et al.* (2018). Only slight advantages led us to select the ECLA23S_F1/ECLA23S_R1 pair, which performed also well with the more complex and uneven Mock 2 community samples.

## 5 Conclusion and perspectives

DNA metabarcoding is a powerful tool that can enhance freshwater ecosystem monitoring programs (Pawlowski *et al.*, 2018). Here, we focused on the first of the five important steps managers and researchers should consider when developing eDNA monitoring program (as suggested by Gold *et al.*, 2022): "select genes and primers to target taxa". After the *in silico* and *in vitro* tests carried out in this study, evaluating the performance of the selected 23S primer pair on environmental samples is now mandatory. Even if several environmental samples were studied with success (Craine *et al.*, 2018; Brown *et al.*, 2022 ) using Sherwood & Presting (2007) primers, which are close to the 23S selected pair (ECLA23S_F1/ ECLA23S_R1), it will be necessary to confirm our *in silico* and *in vitro* results on environmental samples from lake ecosystems.

Actually, if this 23S primer pair is to be part of the biomonitoring toolbox for lake ecosystems, it has to prove its potential to decipher a large phytoplankton biodiversity at a precise taxonomic level and to provide acceptable relative abundances of taxa, in order to accurately feed the biomonitoring metrics. This potential could be evaluated with samples originating from lakes in a large range of climatic, geological and trophic conditions. Challenging the 23S metabarcoding approach will still require to get in parallel microscopy data and 16S metabarcoding data. This will ensure that no species are missed and that no phyla or families are under or over-represented in the taxonomic inventories as shown in the study of Brown *et al.* (2022) where diatoms were over-represented in 23S metabarcoding relative to microscopy. The comparison of all three approaches (microscopy, 16S, 23S) in environmental samples for diversity assessment, taxonomic identification and quantification, will strengthen conclusions.

Producing reliable taxonomic inventories with metabarcoding is also closely linked to the quality and completion of the reference library (Rimet *et al.*, 2021). Being far from complete at the moment (Canino *et al.*, 2021), the 23S reference library will first require a gap analysis (*e.g.* Weigand *et al.*, 2019) to identify completion priorities. Then, 23S sequencing should be processed first on strains that may already be available in culture collections; second, with a prior strain isolation step that is unavoidable for microorganisms. In complement, Rimet *et al.* (2018b) suggested a more operational way using barcode sequences from metabarcoding data of natural samples as a source of primary taxonomic information for reference libraries. Thus, sampling lakes where gap-taxa are known to be present in high relative abundance would be a way to both ascertain the 23S efficiency, and complete its reference library. Finally, a fully operational and standardised reference database should follow the recommendations by Rimet *et al.* (2021) when building barcode reference libraries for aquatic life.

Providing tools for lake biomonitoring requires them to be well-fitted to environmental stakeholder needs in terms of production rate and ease of use (Blancher *et al.*, 2022). Here we tested an automated DNA extraction protocol to start assessing the potential for higher throughput sample processing. Although promising, these first results show a need to improve the protocol. To be operational for professionals, the approach tested here will also require going toward standardisation. Recently, handbooks for DNA-based methods in aquatic monitoring have been released (*e.g.* Bruce *et al.*, 2021; Jerney *et al.*, 2022); inter-laboratory comparisons have been conducted (Baričević *et al.*, 2022; Vasselon *et al.*, 2021). Going to international standards will require next steps as done for example at the European standardisation level (CEN) for diatoms metabarcoding (*e.g.* phytobenthos sampling step: CEN 2018a, b).

*Acknowledgements.* This study was carried out in the framework of the project PhytoDOM, funded by OFB and INRAE, and developed at the Pôle R&D ECLA (ECosystèmes LAcustres).

## Supplementary Material

**Supplementary data 1:** Sequences of the primers tested for 16S.

**Supplementary data 2:** DNA amplicon concentrations introduced in the three type of communities of Mock 2 (in ng/L of DNA). EQ: equimolar, GR: linear gradient, EX: exponential gradient.

The Supplementary Material is available at https://www.limnology-journal.org/10.1051/limn/2023008/olm.

## References

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, Cárdenas P, Čepička I, Chistyakova L, Campo J del, Dunthorn M, Edvardsen B, Eglit Y, Guillou L, Hampl V, Heiss AA, Hoppenrath M, James TY, Karnkowska A, Karpov S, Kim E, Kolisko M, Kudryavtsev A, Lahr DJG, Lara E, Gall LL, Lynn DH, Mann DG, Massana R, Mitchell EAD, Morrow C, Park JS, Pawlowski JW, Powell MJ, Richter DJ, Rueckert S, Shadwick L, Shimano S, Spiegel FW,Torruella G, Youssef N, Zlatogursky V, Zhang Q. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukary Microbiol* 66: 4–119.

Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu YK. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* 37: 815–824.

Baričević A, Chardon C, Kahlert M, Karjalainen SM, Maric Pfannkuchen D, Pfannkuchen M, Rimet F, Smodlaka Tankovic M,Trobajo R, Vasselon V, Zimmermann J, Bouchez A. 2022. Best practice recommendations for sample preservation in metabarcoding studies: a case study on diatom environmental samples. *Metabarcod Metagenom* 6: 349–365.

Bennke CM, Pollehne F, Müller A, Hansen R, Kreikemeyer B, Labrenz M. 2018. The distribution of phytoplankton in the Baltic Sea assessed by a prokaryotic 16S rRNA gene primer system. *J Plankton Res* 40: 244–254.

Blancher P, Lefrançois E, Rimet F, Vasselon V, Argillier C, Arle J, Beja P, Boets P, Boughaba J, Chauvin C, Deacon M, Duncan W, Ejdung G, Erba S, Ferrari B, Fischer H, Hänfling B, Haldin M, Hering D, Hette-Tronquart N, Hiley A, Järvinen M, Jeannot B, Kahlert M, Kelly M, Kleinteich J, Koyuncuoğlu S, Krenek S, Langhein-Winther S, Leese F, Mann D, Marcel R, Marcheggiani S, Meissner K, Mergen P, Monnier O, Narendja F, Neu D, Pinto VO, Pawlowska A, Pawlowski J, Petersen M, Poikane S, Pont D,

Renevier MS, Sandoy S, Svensson J, Trobajo R, Zagyva AT, Tziortzis I, van der Hoorn F B, Vasquez MI, Walsh K, Weigand A, Bouchez A. 2022. A strategy for successful integration of DNA-based methods in aquatic monitoring. *Metabarcod Metagenom* 6: 215–226.

Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. 2015. MSA: an R package for multiple sequence alignment. *Bioinformatics* 31: 3997–3999.

Brown PD, Craine JM, Richards D, Chapman A, Marden B. 2022. DNA metabarcoding of the phytoplankton of Great Salt Lake's Gilbert Bay: spatiotemporal assemblage changes and comparisons to microscopy. *J Great Lakes Res* 48: 110–124.

Bruce K, Blackman R, Bourlat SJ, Hellström AM, Bakker J, Bista I, Bohmann K, Bouchez A, Brys R, Clark K, Elbrecht V, Fazi S, Fonseca V, Hänfling B, Leese F, Mächler E, Mahon AR, Meissner K, Panksep K, Pawlowski J, Schmidt Yáñez P, Seymour M, Thalinger B,Valentini A, Woodcock P, Traugott M, Vasselon V, Deiner K. 2021. A practical guide to DNA-based methods for biodiversity assessment. *Adv Books*. https://doi.org/10.3897/ab.e68634

Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A, Mace GM, Tilman D, Wardle DA, Kinzig AP, Daily GC, et al., 2012. Biodiversity loss and its impact on humanity. Nature 486:59–67.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DAD A2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13: 581–583.

Canino A, Bouchez A, Laplace-Treyture C, Domaizon I, Rimet F. 2021. Phytool, a ShinyApp to homogenise taxonomy of freshwater microalgae from DNA barcodes and microscopic observations. *Metabarcod Metagenom* 5: 199.

Capo E, Domaizon I, Maier D, Debroas D, Bigler C. 2017. To what extent is the DNA of microbial eukaryotes modified during burying into lake sediments? *A repeat-coring approach on annually laminated sediments. J Paleolimnol* 58: 479–495.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci* 108: 4516–4522.

CEN. 2006. *Water quality – E N15204: 2006–Guidance standard on the enumeration of phytoplankton using inverted microscopy (Utermöhl technique)*, 1–42.

CEN. 2018a. *Water quality – CEN/TR 17244-Technical report for the management of diatom barcodes*, 1–11.

CEN. 2018b. Water quality – CEN/TR 17245-Technical report for the routine sampling of benthic diatoms from rivers and lakes adapted for metabarcoding analyses. *CEN/TC 230/WG23, Aquatic Macrophytes and Algae*, 1–8.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–3.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucl Acids Res* 42: 633–642.

Costa NB, Kolman MA, Giani A. 2016. Cyanobacteria diversity in alkaline saline lakes in the Brazilian Pantanal wetland: a polyphasic approach. *J Plankton Res* 38: 1389–1403.

Craine JM, Henson MW, Cameron Thrash J, Hanssen J, Spooner G, Fleming P, Pukonen M, Stahr F, Spaulding S, Fierer N. 2018. Environmental DNA reveals the structure of phytoplankton assemblages along a 2900-km transect in the Mississippi River. *bioRxiv* 261727.

Debroas D, Hugoni M, Domaizon I. 2015. Evidence for an active rare biosphere within freshwater protists community. *Mol Ecol* 24: 1236–1247.

Decelle J, Romac S, Stern RF, Bendif EM, Zingone A, Audic S, Guiry MD, Guillou L, Tessier D, Le Gall F, Gourvil P, Dos Santos AL, Probert I, Vaulot D, de Vargas C, Christen R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol Ecol Resour* 15: 1435–1445.

Del Campo J, Kolisko M, Boscaro V, Santoferrara LF, Nenarokov S, Massana R, Guillou L, Simpson A, Berney C, de Vargas C, Brown MW, Keeling PJ, Wegener Parfrey L. 2018. EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol* 16: e2005849.

Djemiel C, Plassard D, Terrat S, Crouzet O, Sauze J, Mondy S, Nowak V, Wingate L, Ogée J, Maron PA. 2020. µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. *Sci Rep* 10: 5915.

Djurhuus A, Mikalsen SO, Giebel HA, Rogers AD. 2017 Cutting through the smoke: the diversity of microorganisms in deep-sea hydrothermal plumes. *Royal Soc Open Sci* 4: 160829.

Dokulil M, Teubner K, Greisberger S. 2005. Typenspezifische Referenzbedingungen für die integrierende Bewertung des okologischen Zustandes stehender Gewasser Österreichs gemss der EU-Wasserrahmenrichtlinie. Modul 1: Die Bewertung der Phytoplankton struktur nach dem Brettum-Index. Projektstudie Phase 3, *Abschlussbericht. Im Auftrag des Bundesministeriums für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Wien*.xx

Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* 32: 1792–1797.

Eiler A, Drakare S, Bertilsson S, Pernthaler J, Peura S, Rofner C, Simek K, Yang Y, Znachor P, Lindström ES. 2013. Unveiling distribution patterns of freshwater phytoplankton by a next generation sequencing based approach. *PLoS ONE* 8: e53516.

European Commission. 2000 The European parliament and of the council. *Water Framework Directive. Off. J. L* 327: 1–73.

Filker S, Kühner S, Heckwolf M, Dierking J, Stoeck T. 2019. A fundamental difference between macrobiota and microbial eukaryotes: Protistan plankton has a species maximum in the freshwater-marine transition zone of the Baltic Sea. *Environ Microbiol* 21: 603–617.

Füller NJ, Campbell C, Allen DJ, Pitt FD, Zwirglmaier K, Le Gall F, Vaulot D, Scanlan DJ. 2006. Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids. *Aquat Microb Ecol* 43: 79–93.

Gold Z, Wall AR, Schweizer TM, Pentcheff ND, Curd EE, Barber PH, Meyer RS, Wayne R, Stolzenbach K, Prickett K, Luedy J, Wetzer R. 2022. A manager's guide to using eDNA metabarcoding in marine ecosystems. *PeerJ* 10: e14071.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet AL, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucl Acids Res* 41: D597– D604.

Hebert PD, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proc Royal Soc B* 270: 313–321.

Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. 2011. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* 5: 1571–1579.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nat Microbiol* 1: 16048.

Ivanova NV, Watson LC, Comte J, Bessonov K, Abrahamyan A, Davis TW, Bullerjahn GS, Watson SB. 2019. Rapid assessment of phytoplankton assemblages using Next Generation Sequencing-Barcode of Life database: a widely applicable toolkit to monitor biodiversity and harmful algal blooms (HABs). *bioRxiv* 873034.

Jerney J, Hällfors H, Oja J, Reunamo A, Suikkanen S, Lehtinen S. 2022. Guidelines for using environmental DNA in Finnish marine phytoplankton. *Reports of the Finnish Environment Institute 40*. http://hdl.handle.net/10138/351131

Keck F, Millet L, Debroas D, Etienne D, Galop D, Rius D, Domaizon I. 2020. Assessing the response of micro-eukaryotic diversity to the Great Acceleration using lake sedimentary DNA. *Nat Commun* 11: 3831.

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 Resources* 13: 607–619.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucl Acids Res* 41: e1.

Laplace-Treyture C, Feret T. 2016. Performance of the Phytoplankton Index for Lakes (IPLAC): a multimetric phytoplankton index to assess the ecological status of water bodies in France. *Ecol Indic* 69: 686–698.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17: 10–12.

Needham D, Fuhrman J. 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 1: 16005.

Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. 2010. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol* 16: 4135–4444.

Nowinski B, Smith CB, Thomas CM, Esson R, Marin K, Preston CM *et al*. 2019. Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci Data* 6: 129.

Nübel U, Garcia-Pichel F, Muyzer G. 1997. PCR primers to amplify 16S rRNA genes from cyanobacteria. *Appl Environ Microbiol* 63: 3327–3332.

Parada AE, Needham DM, Fuhrman JA, 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology* 18:1403–1414.

Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L, Beja P, Boggero A, Borja A, Bouchez A, Cordier T, Domaizon I, Feio MJ, Filipe AF, et al., 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Science of The Total Environment 637–638:1295–1310.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl Acids Res* 41: D590–596.

R Core Team. 2018. R: a language and environment for statistical computing. Vienna, *Austria: R Foundation for Statistical Computing*.

Reysenbach AL, Pace NR. 1995. Archaea: a laboratory manual—thermophiles. Robb FT, Place AR (Eds.), New York: Cold Spring Harbour Laboratory Press, pp. 101–107.

Rimet F, Chardon, C, Lainé L, Bouchez, A, Jacquet S, Domaizon, I, Guillard J. 2018a. *Thonon Culture Collection −TCC- a freshwater microalgae collection*. https://doi.org/10.15454/UQEMVW

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

Rimet F, Aylagas E, Borja A, Bouchez A, Canino A, Chauvin C, Chonova T, Čiampor FJr, Costa FO, Ferrari BJD, Gastineau R, Goulon C, Gugger M, Holzman M, Jahn R, Kahlert M, Kusber WH, Laplace-Treyture C, Leese F, Leliaert F, Mann DG, Marchand F, Méléder V, Pawlowski J, Rasconi S, Rougerie R, Schweizer M, Trobajo R, Vivien R, Weigand A, Witkowski A, Zimmermann J, Ekrem T. 2021. Metadata standards and practical guidelines for specimen and DNA curation when building DNA barcode reference libraries for aquatic life. *Metabarcod Metagenom* 5: e58056.

Rudi K, Skulberg OM, Larsen F, Jakobsen KS. 1997. Strain characterization and classification of oxyphotobacteria in clone cultures on the basis of 16S rRNA sequences from the variable regions V6, V7, and V8. *Appl Environ Microbiol* 63: 2593–2599.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucl Acids Res* 43: e37.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537–7541.

Sherwood AR, Presting GG. 2007. Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *J Phycol* 43: 605–608.

Sildever S, Nishi N, Inaba N, Asakura T, Kikuchi J, Asano Y, Kobayashi T, Gojobori T, Nagai S. 2022. Monitoring harmful microalgal species and their appearance in Tokyo Bay, *Japan, using metabarcoding. Metabarcod Metagenom 6: e* 79471.

Stern RF, Andersen RA, Jameson I, Küpper FC, Coffroth MA, Vaulot D, Gall FL, Véron B, Brand JJ, Skelton H, Kasai F, Lilly EL, Keeling PJ. 2012. Evaluating the Ribosomal Internal Transcribed Spacer (ITS) as a Candidate Dinoflagellate Barcode Marker. *PLOS ONE* 7:e42780.

United States. 1972. Federal Water Pollution Control Act Amendments of 1972. *Pub.L*. 92–500.

Utermohl H. 1958 Zur Vervollkommung der quantitativen phytoplankton-methodik. *Mitt Int Ver Limnol* 9: 38.

Vasselon V, Rimet F, Domaizon I, Monnier O, Reyjol Y, Bouchez A. 2019. Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: experience and developments from France water framework directive networks. *Metabarcod Metagenom* 3: 101–115.

Vasselon V, Ács É, Almeida S, Andree K, Apothéloz-Perret-Gentil L, Bailet B, Baricevic A, Beentjes K, Bettig J, Bouchez A, Capelli C, Chardon C, Duleba M, Elersek T, Genthon C, Hurtz M, Jacas L, Kahlert M, Kelly M, Lewis M, Macher JN, Mauri F, Moletta-Denat

M, Mortágua A, Pawlowski J, Pérez Burillo J, Pfannkuchen M, Pilgrim E,Pissaridou P, Porter J, Rimet F, Stanic K, Tapolczai K, Theroux S,Trobajo R, van der Hoorn B, Vasquez Hadjilyra MI, Walsh K, Wanless D, Warren J,Zimmermann J, Zupančič M. 2021. The Fellowship of the Ring Test: DNAqua-Net WG2 initiative to compare diatom metabarcoding protocols used in routine freshwater biomonitoring for standardisation. *ARPHA Conference Abstracts* 4: e65142.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261–5267.

Wang Z, Liu L, Tang Y, Li A, Liu C, Xie C, Xiao L, Lu S. 2022. Phytoplankton community and HAB species in the South China Sea detected by morphological and metabarcoding approaches. *Harmful Algae* 118: 102297.

Watanabe K, Kodama Y, Harayama S. 2001. Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *J Microbiolog Methods* 44: 253–262.

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zatovičová Z, Ferreira S, Dijkstra KD, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, B van der Hoorn BB, Japoshvili B,Keresztes L, Keskin E, Leese F, Macher J, Mamos T, Paz G, Pešić V,Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, 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.

West NJ, Schönhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF, Scanlan DJ. 2001. Closely related Prochlorococcus genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* 147: 1731–1744.

Yarimizu K, Fujiyoshi S, Kawai M, Norambuena-Subiabre L, Cascales EK, Rilling JI, Vilugrón J, Cameron H, Vergara K, Morón-López J, Acuña JJ, Gajardo G, Espinoza-González O, Guzmán L, Jorquera MA, Nagai S, Pizarro G, Riquelme C, Ueki S, Maruyama F. 2020. Protocols for monitoring harmful algal blooms for sustainable aquaculture and coastal fisheries in Chile. *Int J Environ Res Public Health* 17: 7642.

Yoon TH, Kang HE, Kang CK, Lee SH, Ahn DH, Park H, Kim HW. 2016. Development of a cost-effective metabarcoding strategy for analysis of the marine phytoplankton community. *PeerJ* 4: e2115.

**Cite this article as**: Canino A, Lemonnier C, Alric B, Bouchez A, Domaizon I, Laplace-Treyture C, Rimet F 2023. Which barcode to decipher freshwater microalgal assemblages? Tests on mock communities. *Int. J. Lim.* 59: 8: