



OPEN

## Mock community experiments can inform on the reliability of eDNA metabarcoding data: a case study on marine phytoplankton

Nadia Marinchel<sup>1,2</sup>✉, Alexis Marchesini<sup>3,4</sup>, Davide Nardi<sup>5</sup>, Matteo Girardi<sup>6</sup>, Silvia Casabianca<sup>2,7,8</sup>, Cristiano Vernes<sup>4,9</sup> & Antonella Penna<sup>2,7,8</sup>✉

Environmental DNA metabarcoding is increasingly implemented in biodiversity monitoring, including phytoplankton studies. Using 21 mock communities composed of seven unicellular diatom and dinoflagellate algae, assembled with different composition and abundance by controlling the number of cells, we tested the accuracy of an eDNA metabarcoding protocol in reconstructing patterns of alpha and beta diversity. This approach allowed us to directly evaluate both qualitative and quantitative metabarcoding estimates. Our results showed non-negligible rates (17–25%) of false negatives (i.e., failure to detect a taxon in a community where it was included), for three taxa. This led to a statistically significant underestimation of metabarcoding-derived alpha diversity (Wilcoxon  $p=0.02$ ), with the detected species richness being lower than expected (based on cell numbers) in 8/21 mock communities. Considering beta diversity, the correlation between metabarcoding-derived and expected community dissimilarities was significant but not strong ( $R^2=0.41$ ), indicating suboptimal accuracy of metabarcoding results. Average biovolume and rDNA gene copy number were estimated for the seven taxa, highlighting a potential, though not exhaustive, role of the latter in explaining the recorded biases. Our findings highlight the importance of mock communities for assessing the reliability of phytoplankton eDNA metabarcoding studies and identifying their limitations.

Phytoplankton is a diverse, polyphyletic group of mostly unicellular, photoautotrophic organisms, which plays a pivotal role for the marine ecosystems and the whole Earth's biosphere<sup>1,2</sup>. In marine environments, the prokaryotic group of cyanobacteria is often the most abundant numerically, but eukaryotic phytoplankton are the most significant in terms of ecosystem processes<sup>3</sup>. Among eukaryotic phytoplankton, diatoms and dinoflagellates are ones of the most abundant, diverse and ecologically significant groups; e.g., the peculiarity of diatoms is their siliceous skeleton, while most of dinoflagellates are characterized by the presence of two locomotory flagella. Diatoms and dinoflagellates are responsible for seasonal blooms in temperate and polar waters and play a key role in primary production<sup>4</sup>. Other important, often abundant microalgae groups are haptophytes and the small prasinophytes<sup>5</sup>. As a whole, phytoplankton is one of the world's major oxygen producers and is also responsible for nearly 50% of global primary production<sup>1,2</sup>. In addition of being the foundation of the aquatic food web, phytoplankton play a key role in biogeochemical cycles, being involved in the biological pump, that takes inorganic carbon (e.g., carbon dioxide) from the atmosphere, fixes it into organic matter through photosynthesis and finally deposits it in marine sediments<sup>1,2</sup>. Due to this process of carbon fixation, marine phytoplankton also has a great impact on Earth's climate<sup>6</sup>. For all these reasons, together with its fast and sensitive response to changes in the aquatic environment, phytoplankton is considered to be a good indicator of marine ecosystems<sup>5,7</sup>. Thus, monitoring the structure and composition of phytoplankton communities may be crucial for assessing water quality, evaluating shifts in the trophic state and studying the response of marine ecosystems to ongoing

<sup>1</sup>Department of Pure and Applied Sciences, University of Urbino, Urbino, Italy. <sup>2</sup>Department of Biomolecular Sciences, University of Urbino, Urbino, Italy. <sup>3</sup>Research Institute on Terrestrial Ecosystems (IRET), National Research Council (CNR), Porano, Italy. <sup>4</sup>National Biodiversity Future Center, Palermo, Italy. <sup>5</sup>DAFNAE, University of Padova, Legnaro, PD, Italy. <sup>6</sup>Conservation Genomics Research Unit, Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige, Italy. <sup>7</sup>Fano Marine Center, Inter-Institute Center for Research on Marine Biodiversity, Resources and Biotechnologies, Fano, Italy. <sup>8</sup>CoNISMa, National Inter-University Consortium for Marine Sciences, Rome, Italy. <sup>9</sup>Forest Ecology Unit, Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige, Italy. ✉email: n.marinchel@campus.uniurb.it; antonella.penna@uniurb.it

climate and environmental changes<sup>8,9</sup>. For example, in the Baltic Sea phytoplankton monitoring has been used to study the seasonal fluctuations in primary producers and their long-term role in nutrient and biogeochemical cycles, allowing the detection of environmental changes on a short to long timescale (from days to decades)<sup>10</sup>. Similarly, phytoplankton communities have been recently used to show the impact of raising temperature in the Mediterranean Sea<sup>2,11,12</sup>. Lastly, some phytoplankton species can produce dangerous biotoxins and in some situations can give rise to harmful algal blooms (HABs), that can cause problems for human health, alterations to marine ecosystems and economic losses to aquaculture and tourism activities<sup>13</sup>. Therefore, developing and implementing accurate, standardized and cost-effective methodologies for monitoring spatio-temporal trends of marine phytoplankton communities is a critical and urgent issue in many contexts<sup>14</sup>.

In the last decade, the new field of environmental DNA (eDNA) metabarcoding<sup>15</sup>, defined as the combination of DNA metabarcoding, i.e., the simultaneous identification of multiple taxa based on high-throughput sequencing (HTS) of a target DNA region, and environmental DNA, i.e., the direct extraction of genetic material from environmental samples, such as air, soil, sediment, water, etc., has emerged as a revolutionary approach to biodiversity research<sup>16</sup>. Indeed, eDNA metabarcoding represents a faster and relatively cheap alternative to conventional biodiversity assessment based on traditional sampling and morphological identification and has proven to be particularly effective in the detection of cryptic, rare and/or elusive species<sup>17</sup>. The metabarcoding of eukaryotic plankton communities by targeting 18S SSU rDNA regions is frequently used to assess diversity patterns from environmental samples<sup>18–20</sup>. In particular, the ribosomal V4 and V9 regions are popular choices for analysing communities of marine eukaryotic protists<sup>21,22</sup>, due to their highly conserved primer binding regions and the ability to detect and identify a wide spectrum of phylogenetically distant taxa<sup>17</sup>. However, the 18S rDNA subunit's gene copy number variation among various phytoplankton groups or even between individuals is a widely discussed issue, and many authors agree that it can be a cause of the discrepancy between true species abundance and metabarcoding-derived read counts introducing a bias in the latter<sup>23–25</sup>. In addition to the problems associated with the variability in rDNA gene copy numbers, a more general challenge in studies using amplicon-based metabarcoding is the choice of barcode region and primer pairs, which can limit or bias the estimated diversity, particularly when the target group of organisms is highly phylogenetically diverse<sup>26</sup>. Finding and applying appropriate markers is still a matter of debate for many metabarcoding applications, and primer bias, i.e. the preferential amplification of certain taxa over others, has become a well-known issue for several taxonomic groups. Indeed, sequence mismatches (even of a few base pairs) between primers and their target sequences, may result in the failure to detect some taxa or, together with PCR-induced biases, in skewed relative sequence abundances<sup>17,27,28</sup>. In addition, sub-optimal field (e.g., sample collection protocols), laboratory (e.g., DNA extraction) and bioinformatic procedures, including the use of incomplete or inaccurate reference sequence databases, and sequencing errors may be additional sources of error in metabarcoding studies<sup>17</sup>. All the above-mentioned shortcomings and problems of eDNA metabarcoding protocols can lead to false negatives, i.e. the failure to detect a taxon at sites (or in samples) where it is present, false positives, i.e. the detection of a taxon in a sample where it should actually be absent (i.e., not intentionally included)<sup>29,30</sup>, or inaccuracies in quantitative estimates, due to the non-correspondence of read counts with the original abundances of taxa in the community/sample<sup>31</sup>. Therefore, eDNA metabarcoding results should be interpreted with caution and several authors are beginning to claim the need for context-specific validation studies<sup>32–34</sup>.

Mock communities, i.e., simulated communities of known and pre-defined species composition, provide an opportunity to test several of the above-mentioned uncertainties and can be used to assess the reliability of metabarcoding-derived results, both in terms of qualitative and quantitative data<sup>35,36</sup>. The use of mock communities in preliminary studies can facilitate the development of more efficient and standardized protocols (from DNA extraction to primer selection and bioinformatic analysis), and their use as positive controls allows for the validation of eDNA studies, or at least can highlight their limitations.

To analyse and compare ecological communities, it is crucial to examine both the diversity of species within and among communities (alpha and beta diversity, respectively). The analysis of alpha and beta diversity levels allows to measure the spatio-temporal changes in community structure and composition, providing the information for the study of their biotic and abiotic drivers and for elucidating ecological processes operating at different scales, including biodiversity response to environmental stresses<sup>10</sup>. With the advent of eDNA metabarcoding and its application in biodiversity assessment, many researchers started to use the derived data as qualitative (number of retrieved taxa) and quantitative (number of amplicon read sequences) information for the calculation of alpha and beta diversity indices<sup>18,21,37</sup>, despite the uncertainties related to this methodological approach, as explained above. However, qualitative and quantitative biases potentially associated with metabarcoding data can lead to the non-detection of target species in biomonitoring or prevent the accurate reconstruction of the relationships between community composition and biotic and abiotic factors<sup>29</sup>.

In this study, we used mock communities to test the efficiency and accuracy of an eDNA metabarcoding protocol in reconstructing both alpha and beta phytoplankton diversity. We chose to adopt an amplicon sequencing approach, targeting the V4 region of the 18S SSU rDNA by means of a widely used broad-spectrum primer pair developed for microbial eukaryotic communities<sup>38</sup>. A total of 21 lab-generated mock communities, varying in species composition and abundance, were constructed using seven common and easily lab-cultured marine diatoms (*Chaetoceros socialis*, *Skeletonema marinoi*, *Thalassionema frauenfeldii* and *Pseudo-nitzschia* spp.) and dinoflagellates (*Alexandrium minutum*, *A. pacificum* and *Scrippsiella* spp.). As unicellular organisms, diatoms and dinoflagellates allowed us to assemble artificial communities directly based on the number of individuals (i.e., cells), instead of using DNA concentration, such as in the majority of mock community studies<sup>39</sup>, or more complex proxies for species richness<sup>35</sup>, thus allowing us to directly investigate the quantitative component of alpha and beta diversity. Using the 21 assembled mock communities and comparing metabarcoding-based and expected (computed from the number of cells included for the different taxa) composition and abundance estimates, this study aimed at: (i) evaluate the accuracy of the tested metabarcoding protocol in detecting the

selected diatom and dinoflagellate taxa; (ii) assess the potential occurrence of false negatives (i.e., the failure to detect a taxon in a mock community in which it was included); and (iii) test the reliability of metabarcoding-derived alpha and beta diversity estimates computed from both qualitative and quantitative data. Based on the results obtained, we tried to investigate the possible explanations for the patterns that emerged.

## Results

### Primer testing and Sanger sequencing of the seven selected phytoplankton taxa

Sanger sequencing performed on DNA extracted from cultured cells before the assembly of mock communities, targeting the 18S V4 region using the same primer pairs as for metabarcoding, successfully amplified all the selected phytoplankton taxa and confirmed the taxonomic assignment of the cultured algal cells, ruling out potential misclassification in the preparatory steps of the experiment. The obtained sequences were deposited in GenBank (see Data Availability).

In silico PCR tests performed using FastPCR<sup>40</sup> showed comparable, relatively high (96–97%), efficiency levels for the selected 18S V4 primers, in all the seven phytoplankton taxa included in our experiment, suggesting that preferential amplification due to primer bias should not be a major issue for the considered taxa.

Sanger sequencing performed using taxon-specific primer pairs, that targeted other ribosomal regions (5.8S and ITS), undoubtedly confirmed the presence of the DNA template of the taxa which resulted undetected by metabarcoding, in all mock communities where false negatives arose (see below section).

### Biovolume and 18S rDNA copy number determination

The estimated average cell biovolume, a morphometric descriptor of phytoplankton species, varied largely among the seven taxa included in the mock communities (42–26,434  $\mu\text{m}^3$ ; Table 1), with average biovolumes for dinoflagellates largely exceeding those for diatoms (12,224  $\mu\text{m}^3$  vs 151  $\mu\text{m}^3$ , respectively). The maximum average biovolume was recorded for *A. pacificum* ( $26,434 \pm 1,675.43 \mu\text{m}^3$ ), followed by *A. minutum* ( $7,070 \pm 404.58 \mu\text{m}^3$ ) and *Scrippsiella* spp. ( $3,167 \pm 233.58 \mu\text{m}^3$ ). Among diatoms, *C. socialis* had the highest value ( $463 \mu\text{m}^3$ ), while *Pseudo-nitzschia* spp. ( $42 \mu\text{m}^3 \pm 4.21$ ) had the lowest.

Similarly, the estimated values for the average 18S rDNA copy number per cell were variable (Table 1) ranging from one to about 10. Among dinoflagellates, *A. minutum* showed the highest 18S rDNA copy number ( $10 \pm 2.04$ ) followed by *Scrippsiella* spp. ( $1.75 \pm 0.13$ ) and *A. pacificum* ( $1.1 \pm 0.0005$ ), while among diatoms the only taxon with average copy number per cell > 1 was *S. marinoi* ( $4 \pm 0.33$ ).

### Mock community metabarcoding, species detection and false negatives

The 18S rDNA amplicon sequencing of the 21 mock communities, assembled starting from lab-cultured cells of the seven diatom and dinoflagellate taxa and following the specific experimental design described in Table 2 and Fig. 1, yielded a total of 837,095 merged reads after trimming, corresponding to 111 ASVs (amplicon sequence variants, i.e. unique sequences). Of these, 830,930 reads (99.3%), corresponding to 50 ASVs, were clustered into seven OTUs, which resulted to be successfully classified into the seven target microalgal taxa. For each of the seven phytoplankton taxon included in the mock communities, the corresponding sequence IDs in the PR2 database (ver. 4.14.0) are reported in Table S1. The obtained classification was further checked by blasting the retrieved OTUs sequences in the NCBI Genbank database, to rule out the risk of misclassification due to potential bias in the reference database. The remaining 6,165 reads (0.7%), corresponding to 61 ASVs, remained unclassified and were manually assigned to non-target taxa (e.g., Insecta) with low sequence similarity or unclassified phytoplankton taxa (sequence similarity < 95%), after blasting their sequences with the NCBI Genbank database. All these sequences, most likely derived from minor contaminants or PCR artifacts, were removed from the dataset.

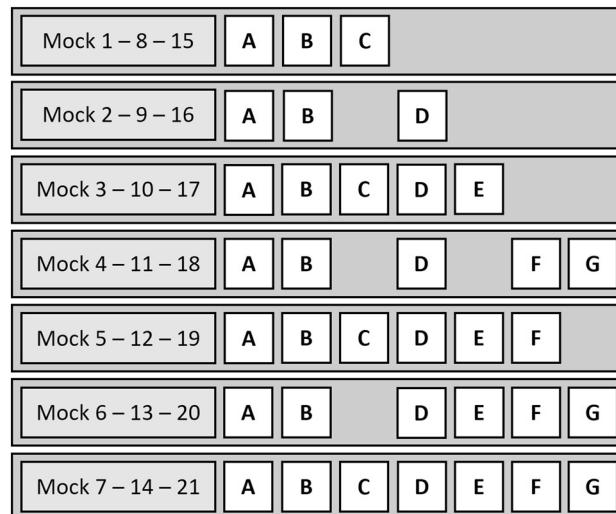
The total number of reads varied among mock communities, ranging from 25,009 to 59,335, although they were constructed starting from the same, predefined, total number of cells ( $1.0 \times 10^6$ ). Considering the single algal taxa separately, *Alexandrium minutum* showed the highest mean number of reads in the 21 mock communities ( $2.64 \times 10^4$ ), while *Skeletonema marinoi* showed the lowest average reads number of  $1.98 \times 10^1$  (Table 3). The sequenced negative control (extraction blank) resulted to be clean (0 reads detected for all the considered microalgal taxa). In general, the metabarcoding-derived community composition better reflected the expected composition (based on cell counts for each taxon) after applying the fourth-root transformation to read counts, as

Taxonomic group	Taxon	Biovolume $\pm$ SE ( $\mu\text{m}^3$ )	Mean 18S rDNA copy number/cell $\pm$ SE
Dinoflagellates	<i>Alexandrium minutum</i>	$7,070 \pm 404.58$	$10.11 \pm 2.04$
	<i>Alexandrium pacificum</i>	$26,434 \pm 1,675.43$	$1.1 \pm 0.0005$
	<i>Scrippsiella</i> spp.	$3,167 \pm 233.58$	$1.75 \pm 0.13$
Diatoms	<i>Chaetoceros socialis</i>	$463 \pm 95.91$	$1 \pm 0.0002$
	<i>Pseudo-nitzschia</i> spp.	$42 \pm 4.21$	$1 \pm 0.0003$
	<i>Skeletonema marinoi</i>	$124 \pm 14$	$4.06 \pm 0.33$
	<i>Thalassionema frauenfeldii</i>	$288 \pm 65.66$	n.a

**Table 1.** Estimated average biovolumes and 18S rDNA copy number per cell for the different phytoplankton taxa included in the mock communities (SE = standard error; n.a. = data not available).

Group	Mock	Dinoflagellates			Diatoms			
		<i>Alexandrium minutum</i> cell number	<i>Alexandrium pacificum</i> cell number	<i>Scrippsiella</i> spp. cell number	<i>Chaetoceros socialis</i> spp. cell number	<i>Pseudo-nitzschia</i> spp. cell number	<i>Skeletonema marinoi</i> cell number	<i>Thalassionema frauenfeldii</i> cell number
1	1	$3.3 \times 10^5$	0	0	$3.3 \times 10^5$	$3.3 \times 10^5$	0	0
	2	$3.3 \times 10^5$	0	$3.3 \times 10^5$	$3.3 \times 10^5$	0	0	0
	3	$2.0 \times 10^5$	$2.0 \times 10^5$	$2.0 \times 10^5$	$2.0 \times 10^5$	$2.0 \times 10^5$	0	0
	4	$2.0 \times 10^5$	0	$2.0 \times 10^5$	$2.0 \times 10^5$	0	$2.0 \times 10^5$	$2.0 \times 10^5$
	5	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	0
	6	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	$1.6 \times 10^5$	0	$1.6 \times 10^5$	$1.6 \times 10^5$
	7	$1.4 \times 10^5$	$1.4 \times 10^5$	$1.4 \times 10^5$	$1.4 \times 10^5$	$1.4 \times 10^5$	$1.4 \times 10^5$	$1.4 \times 10^5$
2	8	$2.0 \times 10^5$	0	0	$4.0 \times 10^5$	$4.0 \times 10^5$	0	0
	9	$2.0 \times 10^5$	0	$4.0 \times 10^5$	$4.0 \times 10^5$	0	0	0
	10	$1.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.5 \times 10^5$	$3.5 \times 10^5$	0	0
	11	$1.0 \times 10^5$	0	$1.0 \times 10^5$	$3.5 \times 10^5$	0	$3.5 \times 10^5$	$1.0 \times 10^5$
	12	$1.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.0 \times 10^5$	$3.0 \times 10^5$	$1.0 \times 10^5$	0
	13	$1.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.0 \times 10^5$	0	$3.0 \times 10^5$	$1.0 \times 10^5$
	14	$8.0 \times 10^4$	$8.0 \times 10^4$	$8.0 \times 10^4$	$3.0 \times 10^5$	$3.0 \times 10^5$	$8.0 \times 10^4$	$8.0 \times 10^4$
3	15	$4.0 \times 10^5$	0	0	$4.0 \times 10^5$	$2.0 \times 10^5$	0	0
	16	$4.0 \times 10^5$	0	$2.0 \times 10^5$	$4.0 \times 10^5$	0	0	0
	17	$3.5 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.5 \times 10^5$	$1.0 \times 10^5$	0	0
	18	$3.5 \times 10^5$	0	$1.0 \times 10^5$	$3.5 \times 10^5$	0	$1.0 \times 10^5$	$1.0 \times 10^5$
	19	$3.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	0
	20	$3.0 \times 10^5$	$1.0 \times 10^5$	$1.0 \times 10^5$	$3.0 \times 10^5$	0	$1.0 \times 10^5$	$1.0 \times 10^5$
	21	$3.0 \times 10^5$	$8.0 \times 10^4$	$8.0 \times 10^4$	$3.0 \times 10^5$	$8.0 \times 10^4$	$8.0 \times 10^4$	$8.0 \times 10^4$

**Table 2.** Mock community experimental design. Taxon abundance (number of cells) was reported for each taxon and mock community. Mock communities were grouped according to the scheme described in Materials and methods.



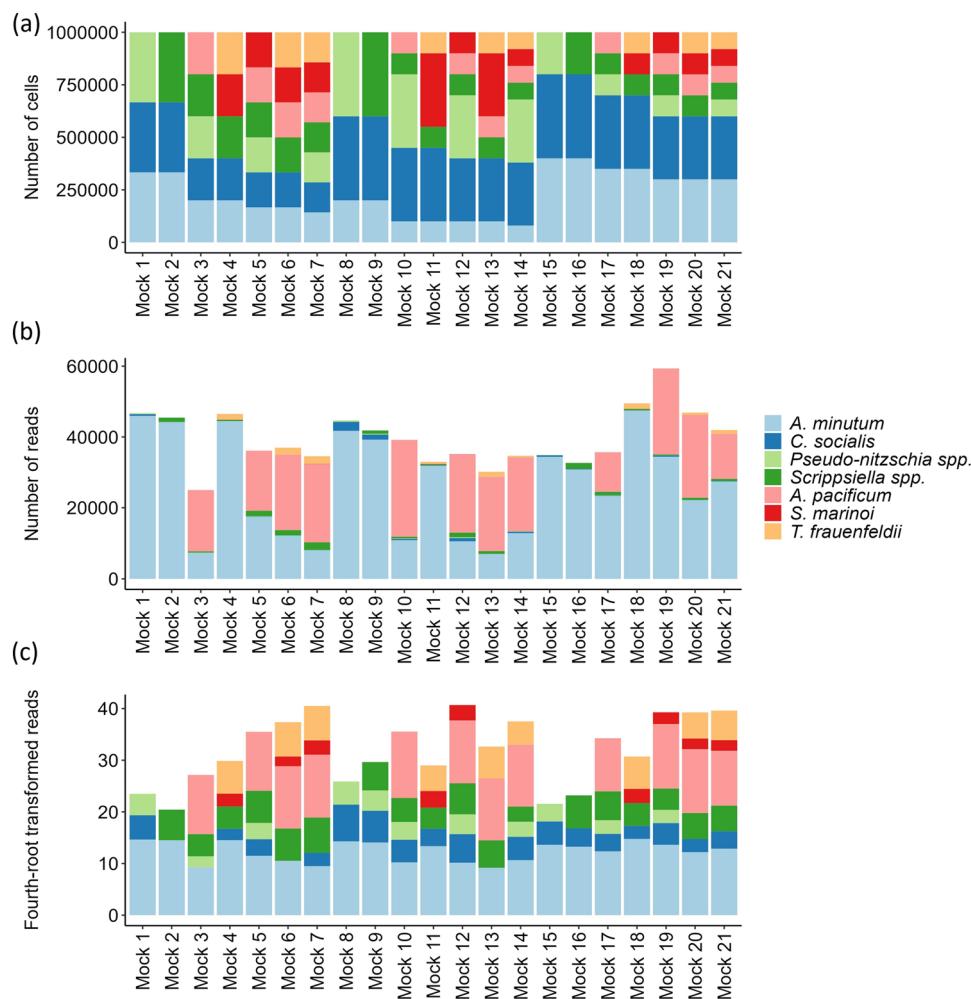
**Figure 1.** Schematic representation of the 21 phytoplankton mock communities. Each of the seven groups (see Materials and methods) is represented by three communities, which differ in taxa abundances (i.e., number of cells; see Table 2); A = *Alexandrium minutum*, B = *Chaetoceros socialis*; C = *Pseudo-nitzschia* spp., D = *Scrippsiella* spp., E = *Alexandrium pacificum*, F = *Skeletonema marinoi*, G = *Thalassionema frauenfeldii*.

suggested to account for the intrinsic unevenness of metabarcoding data (see Materials and methods; subsection Bioinformatics and statistical analysis). In the communities obtained from the untransformed metabarcoding read counts (Fig. 2b), *A. minutum* and *A. pacificum* were overrepresented compared to their expected abundances (Fig. 2a), while all the other taxa resulted to be clearly underestimated. The above-described distortion of the

Taxon	N	Mean cell number	Mean read number	Pearson R <sup>a</sup>	Ratio (R/C)	False negatives	False negatives > 100 reads
<i>Alexandrium minutum</i>	21	$2.30 \times 10^5$	$2.64 \times 10^4$	<b>0.63</b>	0.123	0	0
<i>Alexandrium pacificum</i>	12	$6.84 \times 10^4$	$1.14 \times 10^4$	<b>0.89</b>	0.181	0	0
<i>Scrippsiella</i> spp.	18	$1.32 \times 10^5$	$7.39 \times 10^2$	<b>0.62</b>	0.006	0	1
<i>Chaetoceros socialis</i>	21	$3.02 \times 10^5$	$3.68 \times 10^2$	<b>0.54</b>	0.001	4	8
<i>Pseudo-nitzschia</i> spp.	12	$1.27 \times 10^5$	$8.14 \times 10^1$	<b>0.78</b>	0.0004	2	7
<i>Skeletonema marinoi</i>	12	$8.98 \times 10^4$	$1.98 \times 10^1$	<b>0.55</b>	0.0002	3	11
<i>Thalassionema frauenfeldii</i>	8	$5.09 \times 10^4$	$5.44 \times 10^2$	<b>0.94</b>	0.0106	0	0

**Table 3.** Species detection levels in the analysed mock communities. For each taxon, the overall number of expected occurrences (N, based on the pre-defined community compositions) was reported together with the mean cell number, the mean number of metabarcoding sequence reads, the Pearson's correlation coefficients between number of cells and reads, the R/C ratio (total number of reads divided by total number of cells; see Materials and methods), and the number of false negatives (with and without the 100 reads cut-off).

<sup>a</sup>Statistically significant correlations were reported in bold.



**Figure 2.** Stacked bar plots showing the composition of the 21 phytoplankton mock communities analysed; (a) expected composition (based on cell counts): for each community, the number of cells of each taxon included is shown, according to the colours reported in the legend; (b) metabarcoding-based estimated composition: for each community, the number of sequence reads recorded for each taxon is shown; (c) metabarcoding-based estimated composition, after fourth-root transformation of the read counts (see Materials and methods for details).

estimated taxa abundances appeared to be mitigated after fourth-root transformation of read counts (Fig. 2c). Consequently, we decided to rely on the fourth-root transformed read counts data for further analysis.

Four out of seven microalgal taxa were always amplified in all the mock communities where cells of the considered taxa were included: *Alexandrium minutum*, *Alexandrium pacificum*, *Scrippsiella* spp. and *Thalassionema frauenfeldii*, while three taxa, i.e., *Chaetoceros socialis*, *Pseudo-nitzschia* spp. and *Skeletonema marinoi* occasionally gave rise to false negatives (i.e., the failure to detect a taxon in a mock community where it was expected based on the included algal cells).

Specifically, false negatives were detected in mock 2, 3, 6 and 13 for *C. socialis* (4/21 communities where it was included; 19%); in mock 5, 13 and 14 for *S. marinoi* (3/12 communities where it was included; 25%); and in mock 7 and 21 for *Pseudo-nitzschia* spp. (2/12 communities where it was included; 17%; see Table 3). Moreover, one false positive case was found for *Pseudo-nitzschia* spp., although with a very low number of reads (mock 9; 239 reads).

For all the taxa that occasionally gave rise to false negatives, the R/C ratio (defined as the ratio between total number of reads and total number of included cells; see Bioinformatics and statistical analysis subsection in Materials and methods), resulted to be  $\leq 0.001$ .

When OTUs were filtered applying the 100 reads per sample cut-off threshold, to control for potential minor contaminations, sequencing errors, etc. (see Materials and methods for details), and therefore when adopting a less conservative definition for determining the presence of a taxon, also the dinoflagellate *Scrippsiella* spp. gave rise to a case of false negative, in addition to *C. socialis*, *Pseudo-nitzschia* spp. and *S. marinoi*. The R/C ratio for *Scrippsiella* spp. resulted to be 0.006, therefore close to the previously stated value of 0.001, characterizing the other taxa that occasionally reported false negatives. Moreover, the frequency of false negatives for *C. socialis*, *Pseudo-nitzschia* spp. and *S. marinoi* in the 21 communities increased by 100–300% (*C. socialis*: 8 false negatives; *S. marinoi*: 11; *Pseudo-nitzschia* spp.: 7). Therefore, adopting a less conservative definition for determining the presence of a taxon in each mock community (i.e., using the 100 reads cut-off) resulted in a higher number of false negatives.

Notably, the diatom *T. frauenfeldii* and the two dinoflagellates *A. pacificum* and *A. minutum*, which never gave rise to false negatives with or without the 100 reads cut-off threshold, showed a R/C ratio  $> 0.01$  indicating an overall high sequencing yield.

For each diatom and dinoflagellate taxon, Pearson's correlation test (Table 3) showed a significant and positive correlation between the number of cells included in the mock communities and the number of metabarcoding-derived reads (after fourth-root transformation; see Materials and methods). *T. frauenfeldii* and *A. pacificum* showed the strongest correlations ( $R^2 = 0.94$ ,  $R^2 = 0.89$ , respectively;  $p$ -value  $< 0.001$ ), whereas *S. marinoi* and *C. socialis* showed the weakest correlations ( $R^2 = 0.55$ ,  $R^2 = 0.54$ , respectively;  $p$ -value  $< 0.001$ ). Similar patterns were highlighted when repeating the correlation tests after normalization of read counts (i.e. total sum equal to one), with *T. frauenfeldii* and *A. pacificum* again showing the highest correlation coefficients and with the only difference consisting in the non-detection of a significant correlation for *Scrippsiella* spp. (although  $R^2 = 0.41$ ; Table S2).

### Alpha diversity

In the analysed mock communities, the estimated levels of species richness (i.e., computed from metabarcoding data after fourth-root transformation of read counts; see Materials and methods; average value: 4.62) were lower than the expected (i.e., computed based on the number of cells included for each taxon in the community; average value: 5) and this difference was statistically significant (average difference =  $- 0.38$ ; Wilcoxon signed rank test,  $z = - 2.24$ ,  $p = 0.02$ ). Specifically, based on metabarcoding data, one of the included species was not detected in mock community 2, 3, 5, 6, 7, 14 and 21; while two of the included species were not detected in mock community 13 (Table 4), as a result of the false negatives cases discussed above. Only in one mock community, species richness was higher in the metabarcoding-based estimate than expected from the included cells, due to a false positive (*Pseudo-nitzschia* spp.; detected with a low number of read counts in mock 9, where no cells for the considered taxon were included). Similar patterns were recorded for Pielou's evenness and Simpson's index. An average, statistically significant difference of  $- 0.05$  was observed between the estimated and the expected values for Pielou's evenness (Wilcoxon signed rank test:  $z = - 3.07$ ,  $p < 0.01$ ). The metabarcoding estimates indicated that the taxa were less evenly distributed than the expected (based on cell counts) in 14/21 mock communities (maximum recorded difference:  $- 0.16$ , in mock 1 and 4). An average difference of  $- 0.06$  between estimated and expected values was recorded for Simpson's index (Wilcoxon signed rank test:  $z = - 3.29$ ,  $p < 0.01$ ). Based on metabarcoding data, 17/21 mock communities showed less diversity than expected (maximum observed difference:  $- 0.16$ , in mock 2).

### Beta diversity

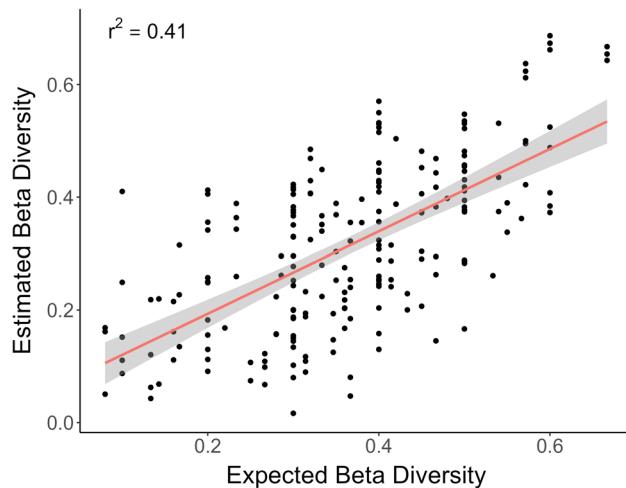
Considering total beta diversity ( $\beta_{tot}$ ), multiple regression on distance matrices (MRM) highlighted a linear and moderate, positive relationship between the estimated (i.e., metabarcoding-derived) and expected (i.e., based on cell-counts) pairwise dissimilarity matrices ( $R^2 = 0.41$ ,  $p < 0.01$ ) (Fig. 3). This result indicated that the dissimilarities between the 21 mock communities, assembled with different taxa composition and abundances (by means of controlling the number of cells per taxon), were captured by the metabarcoding-derived estimates, but the correspondence between the estimated and expected (cell-count based) beta diversity patterns was not perfect.

### Discussion

Marine phytoplankton includes a wide range of eukaryotic unicellular organisms, some of which can be grown in laboratory cultures, and are therefore representing an appropriate target group for the creation of mock communities with controlled species composition to test the efficacy of metabarcoding protocols.

	Expected richness	Estimated richness	Expected Pielou's evenness	Estimated Pielou's evenness	Expected Simpson's diversity index	Estimated Simpson's diversity index
Mock 1	3	3	1	0.84	0.67	0.54
Mock 2	3	2	1	0.87	0.67	0.41
Mock 3	5	4	1	0.88	0.80	0.67
Mock 4	5	5	1	0.84	0.80	0.69
Mock 5	6	5	1	0.91	0.83	0.75
Mock 6	6	5	1	0.92	0.83	0.75
Mock 7	7	6	1	0.93	0.86	0.79
Mock 8	3	3	0.96	0.90	0.64	0.59
Mock 9	3	4	0.96	0.91	0.64	0.68
Mock 10	5	5	0.89	0.92	0.73	0.74
Mock 11	5	5	0.89	0.89	0.73	0.71
Mock 12	6	6	0.92	0.94	0.78	0.79
Mock 13	6	4	0.92	0.96	0.78	0.72
Mock 14	7	6	0.89	0.91	0.79	0.78
Mock 15	3	3	0.96	0.83	0.64	0.53
Mock 16	3	3	0.96	0.88	0.64	0.57
Mock 17	5	5	0.89	0.90	0.73	0.74
Mock 18	5	5	0.89	0.86	0.73	0.69
Mock 19	6	6	0.92	0.87	0.78	0.75
Mock 20	6	6	0.92	0.89	0.78	0.76
Mock 21	7	6	0.89	0.91	0.79	0.78

**Table 4.** Alpha diversity indeces for each phytoplankton mock community. Expected values correspond to those calculated based on the number of cells included for each taxon in the community; estimated values are those derived from the metabarcoding data, after fourth-root transformation of the read counts (see Bioinformatics and statistical analysis subsection, in Materials and methods).



**Figure 3.** Linear regression of estimated beta diversity (based on metabarcoding read counts, after fourth root transformed read counts; see Materials and methods) vs expected beta diversity (based on cell counts).

In this study, we selected seven unicellular diatom and dinoflagellate taxa, which can be commonly found in marine coastal samples collected in the north-western Adriatic Sea, and constructed 21 mock communities of varying species composition and abundances (for more details on the selected taxa and the experimental design, see Materials and methods) by controlling the number of cells for each taxon according to a pre-defined experimental design. This allowed us to evaluate the species detection accuracy of an eDNA metabarcoding protocol, targeting the V4 region of 18S SSU rDNA (a common choice in eukaryotic phytoplankton metabarcoding<sup>38</sup>), and to directly compare estimated (i.e., derived from metabarcoding) with expected (i.e., computed from the cell numbers of the different taxa included in the communities) alpha and beta diversity patterns, both from qualitative (taxa composition) and quantitative (cell abundances) data.

Our main findings were: (a) low species detection rates for some of the seven phytoplankton taxa and, as a consequence, (b) biased metabarcoding estimates of alpha and beta diversity. Specifically, three of the seven microalgal taxa (*Chaetoceros socialis*, *Pseudo-nitzschia* spp. and *Skeletonema marinoi*) gave rise to non-negligible rates of false negatives in the metabarcoding results (i.e., no sequence reads were detected in mock communities where they were expected based on the algal cells included). The other four taxa, on the other hand, never gave rise to false negatives. Therefore, some taxa seemed to be more prone than others to generate false negatives with the metabarcoding protocol employed, and recovery of these taxa by means of Sanger sequencing, using taxon-specific primers, excluded laboratory errors in mock community preparation and DNA extraction as a possible explanation. Notably, one of the taxa exhibiting false negatives, i.e., *Pseudo-nitzschia* spp., actually includes also potentially toxic algae<sup>41</sup>.

Overall, false negatives in metabarcoding results were not strictly related to taxa abundances (i.e., number of cells) in the mock communities considered. Indeed, false negatives also occurred in mock communities where the considered taxon was included with high abundance of cells, such as in mock 13, where *C. socialis* and *S. marinoi* were both represented by  $3 \times 10^5$  cells, three times more abundant than all the other included taxa.

The discussed bias in species detection for some taxa, as well as biased abundance estimates, led to a significant, though not severe, underestimation of metabarcoding-derived alpha and beta diversity levels. The latter seemed to be more robust to bias in our case study, as estimated and expected (i.e., based on cell counts) community dissimilarities were moderately correlated; however, it should be noted that in more species-rich communities (such as in marine environmental samples) the effect of imperfect species detection could be more severe.

It should be noted that several precautions were taken to account for PCR stochasticity and the intrinsic unevenness of metabarcoding data: three independent PCR replicates were performed for each community and pooled together for sequencing, and fourth-root transformation was applied to read counts<sup>19, 42</sup>. The latter clearly resulted in a better correspondence between metabarcoding-derived and expected (based on cell counts) compositional data (see Fig. 2) but didn't eliminate the discussed biases.

Primer bias, a frequent cause of false negatives in metabarcoding studies (see Introduction), seemed not to be the most likely explanation for the occasional detection failures in our experiment, since the chosen primer pair resulted in the same efficiency for all the seven phytoplankton sequences, after in silico PCR tests, and successfully amplified all the seven taxa when performing Sanger sequencing for taxonomic confirmation of the cultured cells. However, it should be noted that results of in silico and in vitro PCR can be somehow different, as the tuning of the PCR conditions can affect the performance of primers; moreover, amplification conditions can be different in single-taxon PCR compared to mock community metabarcoding PCR.

On the other hand, the calculated R/C ratio (i.e., for each taxon, the ratio between the total sum of read counts and the total sum of cells, in all communities) resulted to be related to the arising of false negatives: the taxa having the lowest R/C ratio values (*C. socialis*, *S. marinoi* and *Pseudo-nitzschia* spp.) corresponded with the taxa exhibiting the highest frequency of false negatives. This result may indicate that the overall frequency of false negatives for the different taxa in our experiments was related to the overall, taxon-specific, sequencing yield. Concordant evidence came from Pearson's correlation test computed between the number of cells included in the mock communities and the number of metabarcoding-derived reads: although significant correlations were found for all taxa, they varied widely in magnitude and those with the strongest correlations, i.e., *Thalassionema frauenfeldii* and *Alexandrium pacificum* ( $R^2 = 0.94$  and  $R^2 = 0.89$ , respectively; Table 3), never gave rise to false negatives, while those with the weakest correlations, i.e., *S. marinoi* and *C. socialis* ( $R^2 = 0.55$  and  $R^2 = 0.54$ , respectively), corresponded to those with the highest frequency of false negatives in the metabarcoding results (three and four cases, respectively). In other words, when there was a strong correspondence between cell abundances and metabarcoding reads, no false negatives occurred; on the contrary, when the correspondence was weak (i.e., the same number of cells resulted in fewer sequence reads), false negatives occurred in some of the mock communities. These results were not surprising, but they clearly indicated that the arising of false negatives was not a stochastic process in our experiment.

In marine phytoplankton, several studies have raised the issue of rDNA gene copy number variation among different groups of organisms and species, highlighting it as a potential source of bias in metabarcoding results<sup>18, 23, 25</sup>, and several dinoflagellates are known to have large and repetitive genomes<sup>43, 44</sup>. Our qPCR-based estimates of the average 18S rDNA gene copy number per cell, highlighted *A. minutum* and *S. marinoi* as the taxa with highest value (approximately 10 and four, respectively), followed by *Scrippsiella* spp. and *A. pacificum*, with an average copy number per cell between one and two. The two dinoflagellates *A. minutum* and *A. pacificum* always dominated the communities in terms of read counts (Fig. 2c; Table S3 and Table S4); however, the mean number of read counts recorded for *S. marinoi* was relatively low. None of the four species with the highest average gene copy number gave rise to false negatives, except for *S. marinoi* (three false negatives out of 12 communities where it was included). On the other hand, the lowest values for 18S rDNA gene copy number were recorded for *C. socialis* and *Pseudo-nitzschia* spp. and both taxa gave rise to false negatives, in four and two communities, respectively. Therefore, our data indicated a relationship between low average 18S rDNA gene copy number, low detection rates and abundance bias for some taxa, but with the notable exception of *S. marinoi*. Moreover, it should be noted that the systematic overamplification of *A. pacificum*, clearly evident in the mean number of read counts, can be hardly attributed only to the 18S rDNA gene copy number, since the mean value assessed for this species was close to the taxa giving rise to false negatives. However, the nature of rDNA gene copy number variation can be multifaceted, complicating its estimation and introducing several sources of uncertainty. Indeed, numerous studies indicated both intra-specific and intra-individual variability in rDNA sequences, and both genetic and environmental factors have been reported as potential drivers influencing it<sup>45</sup>. Moreover, the cause of variation may be not only biological, but also due to experimental errors, such as PCR-induced mutations<sup>46</sup>.

Biovolume, a morphometric descriptor of phytoplankton cells, has been found to correlate with the relative abundance of rRNA genes in several, although not all, microalgae species<sup>24, 47</sup>. We therefore tried to investigate the

effects of different biovolumes on the metabarcoding patterns recorded. Our estimates showed that *A. pacificum* and *A. minutum* had average biovolume values of 30 and 10 times bigger than other taxa, respectively (Table 1): these dinoflagellates had the highest yields in terms of overall sequence reads and they were always detected by the tested metabarcoding protocol when they were included in the community (i.e. no false negatives) (Table 3). On the other hand, the smallest biovolumes were measured for *Pseudo-nitzschia* spp. and *S. marinoi* and false negatives were recorded for both taxa (two and three cases, respectively). However, the taxon with the highest frequency of false negatives, i.e., the diatom *C. socialis*, showed an intermediate value for the average biovolume, higher than that of *T. frauenfeldii*, the diatom that never gave false negatives. Therefore, biovolume, which is sometimes used as a proxy for rDNA copy number, seemed to partially but not entirely explain the patterns observed.

Taken together, our results suggested a role for 18S rDNA gene copy number variation in biasing metabarcoding results, but not exhaustively and not for all taxa. In addition to the aforementioned uncertainties in 18S rDNA gene copy number estimation, which may act as a confounding factor, other potential explanations may be involved, such as imperfect DNA extraction for some taxa, the presence of species-specific PCR inhibitors and primer bias<sup>48, 49</sup>, although the latter was at least partially excluded in our experiment (see above). In some cases, another potential source of bias in 18S metabarcoding is the possible presence of secondary structures in the amplified region due to insertions/deletions or introns, with shorter and structurally simpler fragments being preferentially amplified over longer ones<sup>29</sup>. In order to disentangle the influence of gene copy number variation from all the other potential drivers, further experiments should be carried out, with specifically assembled mock communities that control for the above-mentioned factors.

It must be noted that with eDNA metabarcoding, several other factors not considered here, can give rise to false negatives and introduce errors into the results. In this experiment, we “sampled” whole, pre-defined communities and we were confident to sample all the taxa included (as a further check, this was confirmed by Sanger sequencing): therefore, sampling error was absent. However, for environmental samples, inappropriate sampling methods, environmental heterogeneity, low DNA persistence and other factors related to field sampling can play a major role in introducing uncertainties into the data. Moreover, suboptimal DNA extraction protocols and bioinformatics parameters, e.g. the adopted sequence identity threshold for OTUs clustering as well as inaccurate and incomplete reference sequence databases (or unequal representation of taxa) can be additional source of errors<sup>30</sup>. Again, due to the simplified nature of our experiment with low species richness communities and controlled conditions, we didn't test different extraction protocols, nor different bioinformatic pipelines. Regarding the latter aspect, however, we manually inspected all the unclassified sequences (although being an almost negligible fraction of the total: 0.7%), focusing on the taxa that give rise to false negatives in our experiment. By comparing the unclassified sequences with the OTU sequences of the considered phytoplankton species, we excluded the possibility that their non detection was due to the clustering and classification parameters. It should be noted that some of the seven target taxa were underrepresented in term of corresponding sequences included in the PR2 reference database (see Table S1); however, the visual inspection of unclassified sequences ruled out also this possibility as potential explanation for the recorded false negatives and this was further confirmed by the fact that *Scrippsiella* spp. and *Thalassionema frauenfeldii*, the only two taxa represented by a single sequence in the PR2 database, did not give rise to false negatives in our experiment (except for one case when applying the 100 reads cut-off threshold).

Nevertheless, in DNA metabarcoding from environmental samples, great attention should be paid to all the above-mentioned aspects related to sampling strategies and laboratory and bioinformatic protocols<sup>30</sup>.

Lastly, it is common practice in metabarcoding to apply post-processing filtering to ASV/OTU tables to account for the effects of false positives, potentially arising from weak cross-sample or exogenous contamination, PCR/sequencing errors and tag-jumps. As the sequenced negative control for our experiment was clean (0 reads for all of the target taxa), we decided to apply a fixed cut-off threshold and we chose the commonly used value of 100 reads (see subsection Bioinformatics and statistical analysis in Materials and methods, for a detailed explanation). The results from both unfiltered and filtered data were compared: when the 100 reads cut-off was applied to each community, false negatives increased in the three taxa already affected by this issue and an additional taxon gave rise to a false negative in one mock community. This is not surprising, given that (a) applying a cut-off based on the read counts implies a less conservative definition of presence/absence, and (b) the taxa affected by false negative issues correspond to those with the lowest yield in terms of number of reads (see above). Interestingly, we also found a false positive (i.e., detection of a taxon in a community where no cells for that taxon were included) in our metabarcoding results, although with a limited number of reads (possibly due to low cross-contamination of the sample). The potential occurrence of false positives is a well-known issues in eDNA metabarcoding and can affect results even when following strict sampling and laboratory procedures<sup>50</sup>; non-negligible rates of false positives had already been reported also for 18S metabarcoding of plankton mock communities<sup>51</sup>.

Given the simplified nature of our experiment, based on predefined communities with known composition, we were easily able to identify both false negative and false positive cases and our primary aim here was not to test the effect of different post-processing filtering strategies. However, in DNA metabarcoding from environmental samples, great attention should be paid to strategies for mitigating detection errors<sup>29</sup>. During the different steps of collection, extraction, amplification, and sequencing process, negative controls (i.e., blanks) should be included and sequenced to quantify the levels of contamination present; positive controls (i.e., samples which were already sequenced or with known composition) should be also included when possible. Moreover, rigorous post-processing error mitigation strategies should be carried out. Different approaches can be applied: in addition to the application of an arbitrary, fixed cut-off threshold based on read counts for ASVs/OTUs exclusion (which can vary according to the barcode region, target organisms and aims of the study; see e.g.<sup>52, 53</sup>, a proportional threshold based on relative frequencies (e.g., 0.01%, 0.1% or 1%<sup>52, 54</sup>) can be used, or more

complex post-sequencing filtering pipelines, which rely on information from blanks to calculate and remove the contaminant reads for each ASV/OTU in each sample<sup>55</sup>. Lastly, when multiple replicates per sample/site are performed, site occupancy-detection models can be applied to assess the reliability of species detections in a probabilistic manner<sup>42</sup>.

## Conclusions

This study highlighted the weaknesses of a standard amplicon-based eDNA metabarcoding approach based on a single primer pair, when applied to a phylogenetically diverse group of organisms such as phytoplankton. The non-negligible rates of false negatives, recorded for some of the target taxa, warn us of the need to test the accuracy of species detection in phytoplankton eDNA metabarcoding protocols, prior to analyse environmental samples where species composition is unknown. Although this is obviously not possible for all the potentially present phytoplankton species, it is highly recommended particularly when one of the objectives is the monitoring of potentially toxic microalgae, which can cause harmful algal blooms<sup>41</sup> with negative consequences for the marine environment, human health and the economy<sup>56–59</sup>. The frequency of these events is increasing due to climate change<sup>60, 61</sup>, and developing cost-effective and efficient strategies for monitoring potentially harmful algae is therefore becoming increasingly important. Notably, one of the taxa exhibiting false negatives in our experiment was *Pseudo-nitzschia* spp., which actually also includes potentially toxic species<sup>41</sup>.

Using a simplified experiment with controlled conditions, we showed how a metabarcoding protocol with suboptimal detection power, particularly when biased against some species, can lead to an underestimation of alpha and beta diversity. Biased estimates of alpha and beta diversity can limit our power to detect local extinctions, patterns of species turnover in space and time, and community nestedness<sup>62</sup>, making it difficult to accurately characterize spatio-temporal trends in phytoplankton communities and investigating their underlying ecological processes.

To partially mitigate some of the concerns raised, the use of multiple markers and/or primer sets<sup>27, 36</sup> may be advisable in phytoplankton metabarcoding. Moreover, while Illumina MiSeq is still the most popular platform for amplicon sequencing, some studies showed that more modern platforms, such as the NovaSeq, may be able to detect more diversity (even at the same sequencing depth), therefore showing a higher detection power particularly for rare and/or low-abundance taxa<sup>63</sup>. Finally, it is worth mentioning that more recently developed PCR-free metabarcoding approaches, such as hybridisation capture techniques<sup>64, 65</sup>, could overcome the intrinsic limitations of amplicon-based metabarcoding approaches and potentially reduce the associated qualitative and quantitative biases.

As a general indication of good practice, we strongly recommend the use of mock community experiments in phytoplankton metabarcoding, particularly when large and expensive projects are planned or when monitoring potentially harmful algae. Mock communities should be used to set up and validate protocols, from DNA extraction to bioinformatics analysis, to control for potential biases and guide the interpretation of the resulting data, and can also be included as "positive controls" in routine analyses to assess the reproducibility of results. The assembly of mock communities based on cell counts or biomass (for multicellular organisms), rather than using DNA concentration, allows their use as standards to better control for abundance-related biases.

## Materials and methods

### Microalgal cultures, cell counts and biovolume determination

A total of seven marine phytoplankton taxa were used in this study, representing dinoflagellates and diatoms which can be commonly found in the Mediterranean Sea and can be easily grown in lab-cultures. Specifically, the selected taxa were the dinoflagellates *Alexandrium minutum* (strain CNR AMIV1), *Alexandrium pacificum* (strain CNR ACAT15P), *Scrippsiella* sp. (strain VGO1140) and the diatoms *Chaetoceros socialis* (strain CBA22), *Skeletonema marinai* (strain CBA4), *Pseudo-nitzschia* spp. (strain CBAA2), *Thalassionema frauenfeldii* (strain CBA98), which were grown in F/2-Si and F/2 media<sup>66</sup>, respectively. Culture conditions were:  $18 \pm 1^\circ\text{C}$  and  $20 \pm 1^\circ\text{C}$ , for diatom and dinoflagellate taxa, respectively, under a standard 12:12 h light-dark cycle. The light was provided by cool-white, fluorescent bulbs (photon flux of  $100 \mu\text{E m}^{-2} \text{s}^{-1}$ ). Cells were collected during the exponential growth phase and counted using an inverted microscope (Zeiss Axiovert 40 CFL, Germany) equipped with phase contrast according to the Utermöhl method<sup>67</sup>. Counting was carried out at  $\times 200$  or  $\times 400$  magnification on the entire Utermöhl chamber. Size measure of each taxon was done with an inverted microscope Nikon model Ti2-U (Tokyo, Japan). The average cell biovolume, a phytoplankton morphometric descriptor based on different cell features such as shape, length, diameter and height, was estimated for each of the seven diatom and dinoflagellate taxa by measuring morphometric parameters in 10 cells per taxon; calculations were made according to the principle of the most similar geometric shape<sup>68</sup>.

### Primer testing and Sanger sequencing of the seven selected phytoplankton taxa

For our metabarcoding experiment, we chose to amplify the V4 region of the 18S SSU rDNA using the TAREuk454FWD1 and TAREukREV3 primer pair (Table S5), specifically developed for broad-spectrum microbial eukaryotic communities<sup>38</sup>.

Before the assembly of mock communities, for each of the seven selected diatom and dinoflagellate taxa, cultured cells were collected and processed via Sanger sequencing, using the above mentioned primer pair, to confirm taxonomic assignment. For each taxon, genomic DNA was extracted from cultured cells using DNeasy Plant Kit (Qiagen, Hilden, Germany) following manufacturers' instructions and the extracted DNA was stored at  $-80^\circ\text{C}$  until molecular analyses. PCR amplification was carried out in a  $50 \mu\text{l}$  reaction volume containing  $1 \times$  Reaction Buffer (Hot-Start Taq DNA Polymerase, 5U/ $\mu\text{l}$ , Biotechrabbit GmbH, Germany),  $2 \text{ mM MgCl}_2$ ,  $1 \times$  PCR Enhancer,  $200 \mu\text{m}$  dNTP mix,  $200 \text{ nm}$  of each primer,  $1 \text{ U}$  Taq DNA Polymerase and  $10 \mu\text{l}$  of DNA

volume. Thermocycling conditions were as follow: 95 °C for 2 min, followed by 12 cycles at 94 °C for 30 s; 57 °C for 45 s followed by 72 °C for 1 min; 28 cycles at 94 °C for 30 s, 45 °C for 45 s, 72 °C for 1 min, with a final extension at 72 °C for 1 min. PCR products were purified using the ExoSAP-IT kit (USB Corporation, Cleveland, OH). DNA sequencing reactions was performed following the ABI Prism Big-Dye Terminator Kit (Applied Biosystems) standard protocol and the sequencing reaction products were run on an ABI 3130xl Genetic Analyzer (Applied Biosystems). Raw sequences, after the removal of primer sequences, were edited and assembled in contigs using Sequencer 4.7 (Gene Codes Corporation, USA) and taxonomically assigned blasting their sequences with the NCBI Genbank database (National Center for Biotechnology Information; [www.ncbi.nlm.nih.gov/genbank](http://www.ncbi.nlm.nih.gov/genbank)).

Prior to mock communities metabarcoding, in silico PCR were performed using FastPCR<sup>40</sup>, testing the selected 18S V4 primers on the sequences of the seven selected phytoplankton taxa (obtained as described above), with the aim of assessing primer efficiency for all the selected taxa.

For the taxa giving rise to false negatives in the metabarcoding results, the presence of template DNA in the mock communities was checked by Sanger sequencing using taxon-specific primers (Table S5): CsocF and CsocR for *C. socialis*, SkelsspF and SkelsspR for *S. marinoi*, Pseudo 5' and Pseudo 3' for *Pseudo-nitzschia* spp.<sup>69</sup>, with the following PCR thermal cycle conditions: initial denaturation 95 °C for 2 min; 35 cycles at 95 °C for 30 s, 28 °C for 30 s, 72 °C for 30 s; final extension at 72 °C for 5 min. For this specific aim, we decided not to use the same universal 18S primers chosen for mock community metabarcoding, but instead a set of already tested taxon-specific primers, because our primary aim here was to rule out completely the possibility that the detected false negatives could be due to potential errors in laboratory procedure during mock community assembly. Therefore, we wanted to avoid any potential risk of primer bias which, at least in theory, could derive from the use of less specific primers (though this possibility seems unlikely based on the results of our in silico test; see above).

### Quantification of average 18S rDNA copy number

In order to quantify the average 18S rDNA copy number per cell for each of the selected diatom and dinoflagellate taxon, real time PCR was performed, using class-specific primers targeting the 18S rDNA gene (Table S5): 1209f and Diatom18SR1 for diatoms; EUK528f and Dino18SR1 for dinoflagellates<sup>47</sup>. PCR reactions were carried out in a total volume of 25 µl, containing 1 × Hot Rescue Real Time master mix (Hot-Rescue Real Time PCR Kit-SG, Diatheva, Italy), 2 mM MgCl<sub>2</sub>, 300 nM each primer for both assays, 0.625 U of Hot-Rescue Plus DNA Polymerase, and 1 µl of undiluted and 1:10 diluted sub-samples of DNA. The qPCRs were carried out using the StepOne Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) with different thermal cycle programs for the two phytoplankton assays<sup>69</sup>. The determination of rDNA copy number per cell was achieved by means of standard curve method described in Casabianca et al.<sup>69</sup>.

### Mock community experimental design

A total of 21 mock communities were prepared using the seven selected unicellular microalgae, following the experimental design described in Table 2 and Fig. 1.

The algal communities were assembled based on pre-defined different composition and taxa abundances (i.e., number of cells), for testing both qualitative (based on presence/absence) and quantitative (based on taxa abundances). The total number of cells per community was kept constant at a value of  $1.0 \times 10^6$  cells. This value roughly corresponds to the average phytoplankton cell concentration in surface seawater samples collected in the NW Adriatic Sea by means of Niskin bottles (personal observation; see also Neri et al.<sup>2</sup>, Totti et al.<sup>70</sup>) and it was chosen to simulate an example of environmental conditions (although they can vary largely). The 21 assembled mock communities can be ideally divided into three groups, each composed of seven communities (Table 2). In Group 1 (mock 1–7), the different taxa were included in the community with the same abundance (number of cells), but communities differed in species richness (i.e. number of taxa). In Group 2 (mock 8–14), two taxa were included in the communities with higher abundances than the others, one of these being the dinoflagellate *A. minutum*, a species with hypothesized amplification bias; communities also differed in species richness. In Group 3 (mock 15–21), communities were assembled following the scheme of Group 2, but *A. minutum* was included with lower abundance compared to all other taxa. The dinoflagellate *A. minutum* was identified as a candidate species to test for the effects of amplification bias, since this taxon showed the highest average 18S rDNA gene copy number based on qPCR quantification (see Table 1): for this reason, it was included in all mock communities with varying abundance combinations. *C. socialis*, one of the species exhibiting the lowest value for average 18S rDNA gene copy number, was also included in all mock communities with varying abundance combinations, with the purpose of testing how the two extremes (high and low 18S rDNA gene copy numbers) behave in different combinations.

Each mock community was prepared by mixing the required cell culture volumes in a 50 ml sterile falcon tube, and centrifuging them at 4000 rpm for 10 min. After removing supernatant, the pellets were stored at –80 °C until DNA extraction.

### Mock community metabarcoding: DNA extraction, amplification and sequencing

Genomic DNA was extracted from the 21 assembled phytoplankton mock communities using DNeasy Plant Kit (Qiagen, Hilden, Germany) following manufacturers' instructions and the extracted DNA was stored at –80 °C until molecular analyses. One negative control (extraction blank) was included in the extraction batch, to test for DNA carry-over during the extraction process.

For amplicon sequencing, the V4 region of the 18S SSU rDNA was amplified using the TAREuk454FWD1 and TAREukREV3 primer pair (Table S5), and PCR amplification was performed under the same conditions as for Sanger sequencing (see previous section). A PCR negative control (PCR blank) was added to each PCR reaction batch, to check for potential contaminations. For each mock community, three independent PCR replicates

were performed and then pooled together in a single sample for sequencing, to correct for the effect of PCR stochasticity. Each PCR product (including blanks) was checked on a 1.5% (w/v) agarose gel and visualized using a Gel Doc System (Bio-Rad). Extraction and PCR blanks showed no bands on gel, but we decided to proceed with the sequencing of the first, as a further check. Pooled PCR products were purified using the MinElute PCR Gel Extraction Kit (Qiagen, Hilden, Germany), following manufacturer's protocol. Library preparation was performed by means of two amplification steps: an initial PCR amplification using specific primers and a subsequent amplification that integrates relevant flow-cell binding domains and unique indices (using the NexteraXT Index Kit; FC-131-1001/FC-131-1002). The obtained libraries were sequenced at Fondazione Edmund Mach (S. Michele a./A., Trento, Italy), using a MiSeq sequencing platform (Illumina, San Diego, CA) with a  $2 \times 300$  bp paired-end mode.

### Bioinformatics and statistical analysis

Demultiplexed fastq sequences were checked for sequencing quality using FastQC v0.12<sup>71</sup> and trimmed with Cutadapt version 3.5<sup>72</sup>. Primer sequences and their reverse complements were trimmed in both directions, from 5' to 3' using the "-g" and "-G" parameters and from 3' to 5' using the "-A" and "-a" parameters, with the default maximum error rate of 0.1; sequences with a length shorter than 80 base pair were filtered out using the -m flag. Trimmed reads were then imported in QIIME2<sup>73</sup>, which was used for quality filtering (minimum per base quality threshold set to 30), denoising, merging and classification. Denoising was conducted using DADA2<sup>74</sup> through the q2-dada2 plugin, with the following parameters: -p-trunc-len-f 295 and -p-trunc-len-r 220. The obtained Amplicon Sequence Variants (ASVs) were clustered into OTUs with a 97% similarity threshold and classified using the VSEARCH algorithm<sup>75</sup> and the PR2 Protist Ribosomal Reference database, ver. 4.14.0<sup>76</sup>. OTUs with a percentage of identity < 97% were labelled as unassigned and, after a further check blasting their sequences with the NCBI Genbank database (National Center for Biotechnology Information; [www.ncbi.nlm.nih.gov/genbank](http://www.ncbi.nlm.nih.gov/genbank)), removed from the dataset.

To account for the effects of false positives, potentially arising from weak cross-sample or exogenous contamination, PCR/sequencing errors and tag-jump, we applied a cut-off based on a fixed read counts threshold: when an OTU was present in a sample with number of reads < 100, it was removed from that sample. Although more complex filtering methods are available, considering the peculiar features of our study system (based on a low number of taxa and communities with pre-defined composition, which can be viewed as "intrinsic positive controls"), we decided to adopt the simplest strategy. The applied cut-off value of 100 reads per sample corresponds to a common choice in eDNA metabarcoding studies targeting different barcode regions (see e.g.,<sup>52</sup>, including 18S<sup>77,78</sup>, to exclude false positives due to contaminations in reagents or cross-contamination of samples, or to exclude eDNA molecules that can still be present in the water environment but not deriving from living organisms present at the place and time of collection (see e.g., Kelly et al.<sup>79</sup>). It is worth mentioning that a less stringent threshold (e.g., 50 or 10 reads per sample–taxon; e.g., Esenkulova et al.<sup>80</sup>) can be used, but we opted for a relatively high one since this approach is considered more appropriate for reliable ecological inferences<sup>54</sup>. It should be noted that more accurate post-sequencing strategies (see Discussion) should be applied to mitigate detection errors in metabarcoding results from environmental samples, which are characterized, contrary to mock communities, by unknown species composition. Results from both filtered (cut-off: 100) and unfiltered data were then compared. For each taxon, we calculated the ratio between the obtained number of reads and the number of cells which were included in the mock communities (R/C ratio), as a raw estimate of the taxon-specific sequencing yield.

We computed both estimated (from metabarcoding data) and expected (based on cell counts of the different taxa) alpha diversity using the R package VEGAN, version 2.6–4<sup>81</sup>, by means of the following indices: species richness (function *specnumber*), Simpson's Diversity Index and Shannon's Entropy Index (function *diversity*). In addition, Pielou's evenness was calculated as  $H/\log(S)$ , where  $H$  is the Shannon's entropy, and  $S$  is the total number of species (i.e. species richness). In order to compare the estimated (metabarcoding-based) and expected (cell count-based) alpha diversity levels, we used Wilcoxon signed rank test, to account for the non-normal distribution of data, using the R function *wilcox.test*. We then computed the total beta diversity ( $\beta_{tot}$ ) among mock communities, obtained by the sum of the species replacement and species richness components, using BAT package, ver. 2.9.2<sup>82</sup>. Matrices of pairwise dissimilarity for total beta diversity among sites were computed using the function *beta*, for both estimated (metabarcoding-based) and expected (cell count-based) data. On the obtained pairwise dissimilarity matrices, multiple regression on distance matrix (MRM) was then applied in order to test the relationship between estimated and expected beta diversity, using the *MRM* function of R package ECODIST, version 2.0.9<sup>83</sup>. Statistical analyses were conducted in R, ver. 4.1.0. 2021.05.18<sup>84</sup> and Rstudio ver. 2022.12.03<sup>85</sup>; graphs were created using GGPlot2 package ver. 3.4.0<sup>86</sup>.

For the graphical representation of compositional data and for the computation of correlation tests and alpha and beta diversity estimates, read counts were transformed using the fourth-root method with *tran* function of R package ANALOGUE, ver. 0.17–6<sup>87</sup>. This transformation was suggested when comparing metabarcoding read counts with other quantities, to account for the exponential nature of the PCR<sup>19</sup>.

### Data availability

The sequences obtained via Sanger sequencing for each of the seven selected phytoplankton taxon are deposited in NCBI GenBank under accession numbers OR752866-OR752872. The raw sequence data for the DNA metabarcoding experiment are deposited in the NCBI Sequence Read Archive (Bioproject accession PRJNA956448). An R markdown file with the R code used for the statistical analysis, together with the associated input files (OTUs table, cell counts, etc.) are available in the Mendeley Data repository at <https://data.mendeley.com/datasets/hkh7nmdbrv/2> (doi: <https://doi.org/10.17632/hkh7nmdbrv.2>).

Received: 29 April 2023; Accepted: 14 November 2023  
 Published online: 17 November 2023

## References

- Mattei, F. & Scardi, M. Collection and analysis of a global marine phytoplankton primary production dataset. *Earth Syst. Sci. Data* **13**, 4967–4985 (2021).
- Neri, F. *et al.* Phytoplankton communities in a coastal and offshore stations of the northern Adriatic Sea approached by network analysis and different statistical descriptors. *Estuar. Coast. Shelf Sci.* **282**, 108224 (2023).
- Falkowski, P. G. *et al.* The evolution of modern eukaryotic phytoplankton. *Science* **305**, 354–360 (2004).
- Simon, N., Cras, A. L., Foulon, E. & Lemée, R. Diversity and evolution of marine phytoplankton. *Comptes Rendus Biol.* **332**, 159–170 (2009).
- Castellani, C. & Edwards, M. (Eds.) *Marine Plankton: a practical guide to ecology, methodology, and taxonomy* (1st edition) (Oxford University Press., 2017).
- Wilson, J. D., *et al.* The biological carbon pump in CMIP6 models: 21st century trends and uncertainties. *Proc. Natl. Acad. Sci. U.S.A* **119**, e2204369119. <https://doi.org/10.1073/pnas.2204369119> (2022).
- Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive). 19–40 <http://data.europa.eu/eli/dir/2008/56/oj> (2008).
- Desrosiers, C., Leflaive, J., Eulin, A. & Ten-Hage, L. Bioindicators in marine waters: Benthic diatoms as a tool to assess water quality from eutrophic to oligotrophic coastal ecosystems. *Ecol. Indic.* **32**, 25–34 (2013).
- Machado, K. B., de Andrade, A. T., de Almeida, M. F. & Nabout, J. C. Systematic mapping of phytoplankton literature about global climate change: Revealing temporal trends in research. *Hydrobiologia* **850**, 167–182 (2023).
- Olli, K., Nyman, E. & Tamminen, T. Half-century trends in alpha and beta diversity of phytoplankton summer communities in the Helsinki Archipelago, the Baltic Sea. *J. Plankton Res.* **45**, 146–162 (2022).
- El Hourany, R., Mejia, C., Faour, G., Crépon, M. & Thiria, S. Evidencing the impact of climate change on the phytoplankton community of the Mediterranean Sea through a bioregionalization approach. *J. Geophys. Res. Oceans* **126**, e2020JC016808. <https://doi.org/10.1029/2020JC016808> (2021).
- Casabianca, S., Capellacci, S., Ricci, F., Scardi, M. & Penna, A. A phytoplankton time series in the Northwestern Adriatic Sea: Structure and dynamics of the assemblages in a coastal ecosystem. *Estuar. Coast. Shelf Sci.* **278**, 108109. <https://doi.org/10.1016/j.ecss.2022.108109> (2022).
- Hallegraeff, G.M., Anderson, D.M., Cembella, A.D. & Enevoldsen, H.O. (eds) *Manual on Harmful Marine Microalgae in Monographs on Oceanographic Methodology*, 11 (ed. 2nd revised edition) 25–50 (Paris, France, UNESCO, 2004).
- Campbell, L. & Thompson, P. A. Trends in phytoplankton over space and time. *J. Plankton Res.* **45**, 144–145 (2023).
- Didaskalou, E. A., Trimbos, K. B. & Stewart, K. A. Environmental DNA. *Curr. Biol.* **32**, R1250–R1252 (2022).
- Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. *Environmental DNA: For Biodiversity Research and Monitoring* (Oxford University Press, 2018).
- Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* <https://doi.org/10.1016/j.gecco.2019.e00547> (2019).
- Martin, J. L., Santi, I., Pitta, P., John, U. & Gypens, N. Towards quantitative metabarcoding of eukaryotic plankton: An approach to improve 18S rRNA gene copy number bias. *Metabarcoding Metagenom.* **6**, 245–259 (2022).
- Ershova, E. A., Wangensteen, O. S., Descoteaux, R., Barth-Jensen, C. & Præbel, K. Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. *ICES J. Mar. Sci.* **78**, 3342–3355 (2021).
- Penna, A., Casabianca, S., Guerra, A. F., Vernesi, C. & Scardi, M. Analysis of phytoplankton assemblage structure in the Mediterranean Sea based on high-throughput sequencing of partial 18S rRNA sequences. *Mar. Genom.* **36**, 49–55 (2017).
- Armeli Minicante, S., *et al.* Habitat heterogeneity and connectivity: effects on the planktonic protist community structure at two adjacent coastal sites (the Lagoon and the Gulf of Venice, Northern Adriatic Sea, Italy) revealed by metabarcoding. *Front. Microbiol.* **10**, 2736. <https://doi.org/10.3389/fmicb.2019.02736> (2019).
- Siano, R., *et al.* Sediment archives reveal irreversible shifts in plankton communities after World War II and agricultural pollution. *Curr. Biol.* **31**, 2682–2689 (2021).
- Gong, W. & Marchetti, A. Estimation of 18S gene copy number in marine eukaryotic plankton using a next-generation sequencing approach. *Front. Mar. Sci.* **6**, 1–5 (2019).
- Vasselon, V., *et al.* Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods Ecol. Evol.* **9**, 1060–1069 (2018).
- Saad, O. S., *et al.* Genome size, rDNA copy, and qPCR assays for Symbiodiniaceae. *Front. Microbiol.* **11**, 847. <https://doi.org/10.3389/fmicb.2020.00847> (2020).
- Hajibabaei, M., Porter, T. M., Wright, M., & Rudar, J. COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS ONE* **14**, e0220953. <https://doi.org/10.1371/journal.pone.0220953> (2019).
- Ficetola, G. F. & Taberlet, P. Towards exhaustive community ecology via DNA metabarcoding. *Mol. Ecol.* <https://doi.org/10.1111/mec.16881> (2023).
- Liu, M., Burridge, C. P., Clarke, L. J., Baker, S. C. & Jordan, G. J. Does phylogeny explain bias in quantitative DNA metabarcoding?. *Metabarcoding Metagenomics* **7**, 73–88 (2023).
- Santoferrara, L. F. Current practice in plankton metabarcoding: Optimization and error management. *J. Plankton Res.* **41**, 571–582 (2019).
- Burian, A., *et al.* Improving the reliability of eDNA data interpretation. *Mol. Ecol. Resour.* **21**, 1422–1433 (2021).
- Lamb, P. D., *et al.* How quantitative is metabarcoding: A meta-analytical approach. *Mol. Ecol.* **28**, 420–430 (2019).
- McElroy, M. E., *et al.* Calibrating environmental DNA metabarcoding to conventional surveys for measuring fish species richness. *Front. Ecol. Evol.* **8**, 276. <https://doi.org/10.3389/fevo.2020.00276> (2020).
- Morey, K. C., Bartley, T. J. & Hanner, R. H. Validating environmental DNA metabarcoding for marine fishes in diverse ecosystems using a public aquarium. *Environ. DNA* **2**, 330–342 (2020).
- Hajibabaei, M. Demystifying eDNA validation. *Trends Ecol. Evol.* **37**, 826–828 (2022).
- Iwaszkiewicz-Eggebrecht, E., *et al.* Optimizing insect metabarcoding using replicated mock communities. *Methods Ecol. Evol.* <https://doi.org/10.1111/2041-210X.14073> (2023).
- Smith, K. F., Kohli, G. S., Murray, S. A. & Rhodes, L. L. Assessment of the metabarcoding approach for community analysis of benthic-epiphytic dinoflagellates using mock communities. *N. Z. J. Mar. Freshw. Res.* **51**, 555–576 (2017).
- Gran-Stadniczeñko, S., *et al.* Protist diversity and seasonal dynamics in Skagerrak plankton communities as revealed by metabarcoding and microscopy. *J. Eukaryot. Microbiol.* **66**, 494–513 (2019).
- Stoeck, T., *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19**, 21–31 (2010).

39. Hilário, H. O., Mendes, I. S., Guimaraes Sales, N. & Carvalho, D. C. DNA metabarcoding of mock communities highlights potential biases when assessing Neotropical fish diversity. *Environ. DNA* **00**, 1–11 (2023).
40. Kalendar, R., Khassenov, B., Ramankulov, Y., Samuilova, O. & Ivanov, K. I. FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics* **109**, 312–319 (2017).
41. Trainer, V. L. *et al.* Pseudo-nitzschia blooms in the northeastern Pacific Ocean. *PICES Sci. Rep.* **53**, 37–48 (2017).
42. Chen, W. & Ficetola, G. F. Numerical methods for sedimentary-ancient-DNA-based study on past biodiversity and ecosystem functioning. *Environ. DNA* **2**, 115–129 (2020).
43. Casabianca, S., Perini, F., Battocchi, C., Casabianca, A. & Penna, A. Monitoring toxic *Ostreopsis cf. ovata* in recreational waters using a qPCR based assay. *Mar. Pollut. Bull.* **88**, 102–109 (2014).
44. Perini, F. *et al.* New approach using the real-time PCR method for estimation of the toxic marine dinoflagellate *Ostreopsis cf. ovata* in marine environment. *PLoS ONE* **6**, e17699. <https://doi.org/10.1371/journal.pone.0017699> (2011).
45. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
46. Wang, C., *et al.* Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. *Proc. Royal Soc. B P Roy. Soc. B-Biol. Sci.* **284**, 20170425. <https://doi.org/10.1098/rspb.2017.0425> (2017).
47. Godhe, A. *et al.* Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008).
48. Tapolczai, K., *et al.* Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Front. Ecol. Evol.* **7**, 409. <https://doi.org/10.3389/fevo.2019.00409> (2019).
49. Mónika, D., *et al.* Applicability of diatom metabarcoding in the ecological status assessment of Hungarian lotic and soda pan habitats. *Ecol Indic.* **130**, 108105. <https://doi.org/10.1016/j.ecolind.2021.108105> (2021).
50. Ficetola, G. F., Taberlet, P. & Coissac, E. How to limit false positives in environmental DNA and metabarcoding?. *Mol. Ecol. Resour.* **16**, 604–607 (2016).
51. Zhang, G. K., Chain, F. J., Abbott, C. L. & Cristescu, M. E. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evol. Appl.* **11**, 1901–1914 (2018).
52. Tommasi, N., Ferrari, A., Labra, M., Galimberti, A. & Biella, P. Harnessing the power of metabarcoding in the ecological interpretation of plant-pollinator DNA data: strategies and consequences of filtering approaches. *Diversity (Basel)* **13**, 437. <https://doi.org/10.3390/d13090437> (2021).
53. Drake, L. E. *et al.* An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. *Methods Ecol. Evol.* **13**, 694–710 (2022).
54. Alberdi, A., Aizpurua, O., Gilbert, M. T. P. & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* **9**, 134–147 (2018).
55. McKnight, D. T. *et al.* microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environ. DNA* **1**, 14–25 (2019).
56. Grattan, L. M., Holobaugh, S. & Morris, J. G. Jr. Harmful algal blooms and public health. *Harmful Algae* **57**, 2–8 (2016).
57. Griffith, A. W. & Gobler, C. J. Harmful algal blooms: a climate change co-stressor in marine and freshwater ecosystems. *Harmful Algae* **91**, 101590. <https://doi.org/10.1016/j.hal.2019.03.008> (2020).
58. Trainer, V. L. & Yoshida, T. Proceedings of the workshop on economic impacts of harmful algal blooms on fisheries and aquaculture. *PICES Sci. Rep.* **47**, 2–79 (2014).
59. Diaz, P. A. *et al.* Impacts of harmful algal blooms on the aquaculture industry: Chile as a case study. *Perspect. Phycol.* **6**, 39–50 (2019).
60. Brosnahan, M. L., Fischer, A. D., Lopez, C. B., Moore, S. K. & Anderson, D. M. Cyst-forming dinoflagellates in a warming climate. *Harmful Algae* **91**, 101728. <https://doi.org/10.1016/j.hal.2019.101728> (2020).
61. Gobler, C. J. Climate change and harmful algal blooms: Insights and perspective. *Harmful Algae* **91**, 101731. <https://doi.org/10.1016/j.hal.2019.101731> (2020).
62. Bevilacqua, S. & Terlizzi, A. Nestedness and turnover unveil inverse spatial patterns of compositional and functional β-diversity at varying depth in marine benthos. *Divers. Distrib.* **26**, 743–757 (2020).
63. Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A. & Hajibabaei, M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.* **9**, 5991. <https://doi.org/10.1038/s41598-019-42455-9> (2019).
64. Günther, B. *et al.* Capture by hybridization for full-length barcode-based eukaryotic and prokaryotic biodiversity inventories of deep sea ecosystems. *Mol. Ecol. Resour.* **22**, 623–637 (2022).
65. Chakoori, O., Comtet-Marre, S. & Peyret, P. RiboTaxa: Combined approaches for rRNA genes taxonomic resolution down to the species level from metagenomics data revealing novelties. *NAR genom. bioinform.* **4**, lqac070. <https://doi.org/10.1093/nargab/lqac070> (2022).
66. Guillard, R. R. L. Culture of phytoplankton for feeding marine invertebrates in Culture of Marine Invertebrate Animals (ed. W. L. Smith & M. H. Chanley) 26–60 (Plenum Press, 1975).
67. Hasle, G. R. The Inverted-Microscope Methods. In *Phytoplankton Manual, Monographs on Oceanographic Methodology* (Ed. Sournia, A.) 88–96 (UNESCO, Paris, 1978).
68. Hillebrand, H., Dürselen, C. D., Kirschtel, D., Pollingher, U. & Zohary, T. Biovolume calculation for pelagic and benthic microalgae. *J. Phycol.* **35**, 403–424 (1999).
69. Casabianca, S. *et al.* Structure and environmental drivers of phytoplanktonic resting stage assemblages in the central Mediterranean Sea. *Mar. Ecol. Prog. Ser.* **639**, 73–89 (2020).
70. Totti, C. *et al.* Phytoplankton communities in the northwestern Adriatic Sea: Interdecadal variability over a 30-years period (1988–2016) and relationships with meteo-climatic drivers. *J. Mar. Syst.* **193**, 137–153 (2019).
71. Andrews, S. FastQC: a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
72. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
73. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
74. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
75. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584. <https://doi.org/10.7717/peerj.2584> (2016).
76. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, 597–604 (2013).
77. Meyer, A. *et al.* Morphological vs. DNA metabarcoding approaches for the evaluation of stream ecological status with benthic invertebrates: testing different combinations of markers and strategies of data filtering. *Mol. Ecol.* **30**, 3203–3220 (2021).
78. DeMone, C., *et al.* Application of next generation sequencing for detection of protozoan pathogens in shellfish. *Food Waterborne Parasitol.* **21**, e00096. <https://doi.org/10.1016/j.fawpar.2020.e00096> (2020).

79. Kelly, M., Boonham, N., Juggins, S., Mann, D., & Glover, R. Further development of a DNA based metabarcoding approach to assess diatom communities in rivers. *Chief Scientist's Group report* [https://assets.publishing.service.gov.uk/media/61a0def28fa8f5037778c1f4/Assessing\\_river\\_nutriennu\\_using\\_diatom\\_DNA\\_-\\_further\\_development\\_of\\_an\\_operational\\_method\\_-\\_report.pdf](https://assets.publishing.service.gov.uk/media/61a0def28fa8f5037778c1f4/Assessing_river_nutriennu_using_diatom_DNA_-_further_development_of_an_operational_method_-_report.pdf) (2020).
80. Esenkulova, S. *et al.* Comparing metabarcoding and morphological approaches to identify phytoplankton taxa associated with harmful algal blooms. *Facets* **5**, 784–811 (2020).
81. Oksanen, F.J., *et al.* Vegan: Community Ecology Package. <https://CRAN.R-project.org/package=vegan> (2017).
82. Cardoso, P., Rigal, F. & Carvalho, J. C. BAT—Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity. *Methods Ecol. Evol.* **6**, 232–236 (2015).
83. Goslee, S. C. & Urban, D. L. The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* **22**, 1–19 (2007).
84. R Core Team. R: a language for statistical computing. <https://www.R-project.org> (2022).
85. RStudio Team. RStudio: integrated development for R. <https://www.rstudio.com> (2020).
86. Wickham, H. ggplot2: elegant graphics for data analysis. <https://ggplot2.tidyverse.org> (2016).
87. Simpson, G. L. & Oksanen, J. analogue: Analogue and weighted averaging methods for palaeoecology. <https://cran.r-project.org/package=analogue> (2021).

## Acknowledgements

This work was supported by a grant of REMEST (REsearch MEthods in Science and Technology) Ph.D School of University of Urbino, Italy. We thank Capellacci Samuela for her technical support in caring, counting and computing cell biovolumes of microalgal cultures, and Fracasso Ilaria for her help in approaching bioinformatic analysis. We also thank Fondazione E. Mach (FEM) Sequencing and Genotyping Platform for library preparation, amplicon sequencing and demultiplexing. Lastly, we would like to express special thanks to two anonymous reviewers for their valuable and constructive comments on an earlier version of the manuscript.

## Author contributions

N.M., A.M., D.N., S.C., C.V. and A.P. designed the study. N.M. and M.G. performed laboratory analysis; N.M., D.N. and A.M. performed data analyses. N.M. wrote the first draft of the manuscript. All authors contributed to the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-47462-5>.

**Correspondence** and requests for materials should be addressed to N.M. or A.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023