Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes

Caterina R. Giner, Irene Forn, Sarah Romac, Ramiro Logares, Colomban de Vargas, Ramon Massana

A Research Paper submitted to *Applied and Environmental Microbiology*

Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes

Caterina R. Giner[a#], Irene Forn[a], Sarah Romac[b,c], Ramiro Logares[a], Colomban de Vargas[b,c], and Ramon Massana[a#]

[a] Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM-CSIC), Barcelona, Spain.

[b] CNRS, UMR 7144, Station Biologique de Roscoff, Roscoff, France.

[c] Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, France.

# Address correspondence to: Caterina R. Giner, caterina@icm.csic.es, and Ramon Massana, ramonm@icm.csic.es

Institut de Ciències del Mar (CSIC), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain. Phone: 34-93-2309500; Fax: 34-93-2309555;

Running title: Correspondence between HTS and cell abundance

Abstract

High-throughput sequencing (HTS) is revolutionizing environmental surveys of microbial diversity in the three domains of life by providing detailed information on which taxa are present in microbial assemblages. However, it is still unclear how the relative abundance of specific taxa gathered by HTS correlates with cell abundances. Here, we quantified the relative cell abundance of 6 picoeukaryotic taxa in 13 planktonic samples from six European coastal sites using epifluorescence microscopy on TSA-FISH preparations. These relative abundance values were then compared with HTS data obtained in three separate molecular surveys: 454 sequencing the V4 region of the 18S rDNA using DNA and RNA extracts (DNA-V4 and cDNA-V4), and Illumina sequencing the V9 region (cDNA-V9). The microscopic and molecular signals were generally correlated, indicating that a relative increase in specific 18S rDNA was the result of a large proportion of cells in the given taxa. Despite these positive correlations, the slopes often deviated from 1, precluding a direct translation of sequences to cells. Our data highlighted clear differences depending on nucleic-acid template or the 18S rDNA region targeted. Thus, the molecular signal obtained using cDNA templates was always closer to relative cell abundances, while the V4 and V9 regions gave better results depending on the taxa. Our data supports the quantitative use of HTS data but warn about considering it as direct proxy of cell abundances.

Key words: 18S rDNA, 454 pyrosequencing, FISH, Illumina sequencing, picoeukaryotes, specific abundances

Importance

Direct studies on marine picoeukaryotes by epifluorescence microscopy are problematic due to the lack of morphological features, in addition to the limited number and poor resolution of specific phylogenetic probes used in FISH routines. As a consequence, there is an increasing use of molecular methods, including high-throughput sequencing (HTS), to study marine microbial diversity. HTS can provide a detailed picture of the taxa present in a community, and can reveal diversity not evident using other methods, but it is still unclear the meaning of the sequence abundance in a given taxa. Our aim is to investigate the correspondence between the relative HTS signal and relative cell abundances in selected picoeukaryotic taxa. Environmental sequencing provides reasonable estimates of the relative abundance of specific taxa. Better results are obtained when using RNA extracts as templates, while the region of the 18S rDNA influenced differently depending on the taxa assayed.

Introduction

Protists are key components of marine ecosystems, being major players in the global respiration and production budgets (1, 2) and playing central roles in marine food webs (3). Despite their importance and ubiquity, it was only during the past decade that environmental studies, based on molecular (i.e. culture-independent) techniques, revealed an unsuspected protist diversity in a large variety of marine ecosystems (4-13). These studies were based on the analysis of 18S ribosomal RNA (rRNA) genes retrieved directly from natural assemblages by PCR amplification, cloning and sequencing. Nowadays, the development and use of high-throughput sequencing tools (HTS), e.g. 454 or Illumina, which produce more than thousands of sequences from a single sample, has revolutionized the field, allowing deeper assessments of diversity (14), as well as better estimates of specific relative abundances. One of the main challenges of this approach, however, is to understand the correspondence between the relative abundances of sequences and cells. That is, how close is the specific diversity detected in molecular surveys to the true species composition of natural assemblages.

Few studies have analyzed the relationship between direct microscopic inspections and sequencing data in protists. One of the first studies compared cloning and sequencing results with an accurate list of protists species (5-100 μm size range) identified by microscopy (15). In that case, as the sequencing effort was very limited (less than 100 clones), few of the protists identified by morphology were detected in the sequencing set. In addition, the few sequences obtained did not represent the dominant observed species, a clear sign of the biases in this molecular approach. More recent comparative studies used HTS, therefore were not limited by the sequencing effort, but focused on specific taxa, in particular marine and freshwater ciliates (2, 16-18). Ciliate species have the advantage of having conspicuous morphological traits that allow proper identification by inverted microscopy. In most cases, the same species were found in microscopic and molecular datasets, but the relative abundance of sequences and morphotypes were not in agreement, so each approach revealed a different community structure. Other studies prepared mock communities and the results obtained were similar: all individual taxa were detected, but the relative proportion of sequence types was different from cell mixes (19, 20). Overall, the popularization of HTS now allows a high-resolution exploration of protist richness present in natural samples, yet when it comes to evenness, the picture obtained is still limited.

Among protists, picoeukaryotes (protists up to 3 μm in size) are known to be very diverse, widely distributed, and ecologically important in the marine plankton realm (21). Picoeukaryotes are counted as a group by epifluorescence microscopy using a general DNA stain (22) or by flow cytometry (23), but due to their small size and lack of morphological traits (24) they cannot be taxonomically identified by these tools. This can be achieved with FISH (Fluorescence *in situ* Hybridization), which enables the visualization and quantification of specific cells in natural assemblages by using oligonucleotide probes as phylogenetic stains (25). FISH has served to identify the cells from novel environmental clades (11, 26, 27), and has been applied in a few marine surveys (28-31). But this approach is relatively time consuming and targets only one taxon at a time.

In this study, we assess the feasibility of using HTS data as a quantitative metric in picoeukaryote diversity studies, by comparing relative HTS read abundances with relative FISH cell counts in selected picoeukaryotic taxa. Differently to the previous studies in which a single taxa (ciliates) or artificial communities were analyzed, here we focus in a set of highly divergent lineages found in geographically separated and unrelated microbial assemblages. Any pattern emerging from this heterogeneous and noisy dataset is expected to be rather robust. We also investigate if there is a difference in community composition assessed by using environmental DNA or RNA extracts as templates (DNA and cDNA reads, respectively), sequencing different regions of the 18S rDNA (V4 versus V9), or using different HTS platforms (454 versus Illumina). To address these questions we used published sequencing datasets from several European coastal samples (Massana *et al*., 2015 for DNA/cDNA-V4 (32) and Logares *et al*., 2014 for cDNA-V9 (33)) and chose 6 picoeukaryote taxa (<3 µm) for which we had specific FISH probes for quantification.

Materials and Methods

Sampling

Samples were taken during the BioMarKs project (http://www.biomarks.org) in six European coastal sites: Blanes (Spain, 41° 40' N, 2° 48' E), Gijon (Spain, 43° 40' N; 5° 35' W), Naples (Italy, 40° 48' N, 14° 15' E), Oslo (Norway, 59° 16' N, 10° 43' E), Roscoff (France, 48° 46' N, 3° 57' W) and Varna (Bulgaria, 43°10' N, 28° 50' E) (Table 1). Seawater was collected with Niskin bottles attached to a CTD (conductivity-temperature-depth) rosette at surface and deep chlorophyll maximum (DCM) depths. For molecular surveys, ~20 L of seawater was pre-filtered through a 20 µm metallic mesh and then sequentially filtered through 3 µm and 0.8 µm polycarbonate filters (142 mm diameter). The later filter contained the picoplankton (0.8-3 µm size fraction) and was flash-frozen and stored at -80ºC. The filtration time was less than 30 minutes to avoid RNA degradation.

Unfiltered seawater was taken for direct cell counts. For total microscopic counts, seawater samples were fixed with glutaraldehyde (1% final concentration) and left for 1-24 h at 4ºC. Then, aliquots of 20 ml were filtered through 0.6 µm polycarbonate black filters and stained with DAPI (4',6-diamidino-2-phenylindole, 5 µg ml^{-1}). Filters were mounted on a slide and stored at -20ºC until processed. For TSA-FISH (Tyramide Signal Amplification-Fluorescent *in situ* Hybridization) specific counts, aliquots of 100 ml were fixed with filtered formaldehyde (3.7% final concentration), incubated for 1-24 h in the dark at 4ºC and filtered through 0.6 µm polycarbonate filters (25 mm diameter). Filters were kept at -80ºC until processed. For flow cytometry counting of photosynthetic picoeukaryotes, aliquotes of 1.5 ml were fixed with a mix of paraformaldehyde/glutaraldehyde (1%/0.25% final concentrations), frozen in liquid nitrogen and stored at -80ºC until processed.

Picoeukaryote cell abundance by DAPI staining and flow cytometry

Total cell abundance of picoeukaryotes was estimated in DAPI-stained filters. Cells were counted with an epifluorescence microscope (Olympus BX61) at 1000X under UV excitation, changing to blue light excitation to verify the presence or absence of chlorophyll autofluorescence (phototrophic and heterotrophic cells, respectively). A transect of about 13 mm was inspected and cells were classified in size classes: 2 µm, 3 µm, 4 µm, 5 µm and >5 µm. All data reported in the study refers to cells within the two smaller size classes (2-3 µm), which account on average for 82% of the cells.

Cell abundance of photosynthetic picoeukaryotes was determined in a FACSort flow cytometer by using the red fluorescence signal (chlorophyll) after exciting in a 488 nm laser and the SSC (side-scattered light) of each particle. Fluorescent microspheres (0.95 µm beads) were added as an internal standard (at 10^5 beads ml^{-1}). Data was acquired for 2-4 minutes with a flow rate of 50 to 100 µl min^{-1} using the settings previously described (34).

Cell abundance of specific picoeukaryote taxa by TSA-FISH

The specific oligonucleotide probes used targeted several picoeukaryote taxa: NS4 and NS7 targeted the uncultured clades MAST-4 and MAST-7; CRN02 and MICRO01 the species *Minorisa minuta* and *Micromonas* spp.; PELA01 the class Pelagophyceae; and ALV01 the environmental clade MALV-II (Table 2). These probes have been published in other studies (see Table 2 for references) except NS7. Probe NS7 was designed here with ARB (35) and targeted 91% of the 192 sequences from MAST-7 available in GenBank, had 1 mismatch with the remaining MAST-7 sequences, and at least 2 central mismatches with non-target sequences. Probe NS7 gave better signal when combined with oligonucleotide helpers contiguous to the probe region (NS7-HelperA: AACCAACAAAATAGCAC; NS7-HelperB: CCCAACTATCCCTATTAA) that were added in the hybridization buffer at same concentration than the probe. We tested a range of formamide concentration to find the best hybridization condition, and checked that the probe gave negative signal with a variety of non-target cultures. Finally, a probe targeting all eukaryotes (EUK502, 36) was also used. All probes were labeled with horseradish peroxidase (HRP).

Hybridizations were performed as previously described (37). Filter pieces (about 1/10) of the 0.6 µm polycarbonate filters were covered with 20 µl of hybridization buffer (40% deionized formamide [except 30% for probe CNR01], 0.9 M NaCl, 20 mM Tris-HCl [pH 8], 0.01% sodium dodecyl sulfate [SDS]) and 2 µl of HRP-labeled probes (stock at 50 ng $µl^{-1}$), and incubated overnight at 35ºC. After the hybridization, filter pieces were washed twice for 10 min at 37ºC with a washing buffer (37 mM NaCl [74mM NaCl when hybridizing with 20% formamide], 5 mM EDTA, 0.01% SDS, 20mM Tris-HCl [pH 8]). Tyramide signal amplification (TSA) was carried out in a solution (1x PBS, 2 M NaCl, 1 mg ml^{-1} blocking reagent, 100 mg ml^{-1} dextran sulfate, 0.0015% H_2O_2) containing Alexa 488-labeled tyramide (4 µg ml^{-1}), by incubating in the dark at room temperature for 30-60 min. Filter pieces were transferred twice to a phosphate buffer (PBS) bath in order to stop the enzymatic reaction and air dried at room temperature. Cells were countersained with DAPI (5 µg ml^{-1}) and filter pieces were mounted on a slide. Targeted FISH cells were counted by epifluorescence under blue light excitation and checked with UV radiation (DAPI staining) for the presence of the nucleus. Cells labeled with the probe EUK502 were counted using the same size classes as for

DAPI counts. Data reported refers to cells of 2-3 µm, which account on average for 84% of the cells.

High-throughput sequencing by 454 and Illumina

HTS data derives from previously published papers taken during the BioMarKs project (http://www.biomarks.org/). Total DNA and RNA from 13 picoplankton samples were extracted simultaneously from the same filter. For RNA extracts, contaminating DNA was removed and RNA was immediately reverse transcribed to cDNA. Data for the 454 sequencing derives from Massana *et al*. (32) and used the eukaryotic universal primers TAReuk454FWD1 and TAReukREV3 (38), which amplified the V4 region of the 18S rDNA (~ 380 bp). Amplicon sequencing from DNA and cDNA templates was carried out on a 454 GS FLX Titanium system (454 Life Sciences, USA) in Genoscope (http://www.genoscope.cns.fr, France). The complete sequencing dataset is available at the European Nucleotide Archive (ENA) under the accession number PRJEB9133 (http://www.ebi.ac.uk/ena/data/view/PRJEB9133). Data for the Illumina sequencing derives from Logares *et al*. (33) and used the eukaryotic universal primers 1398f and 1510r (39), which amplified the V9 region of the 18S rDNA (~130 bp). Paired-end 100 bp sequencing was performed using a Genome Analyzer IIx (GAIIx) system located at Genoscope. Only RNA (cDNA) samples were sequenced with Illumina. Sequences are publicly available at MG-RAST (http://metagenomics.anl.gov) under accession numbers 4549958.3, 4549965.3, 4549959.3, 4549945.3, 4549943.3, 4549927.3, 4549941.3, 4549954.3, 4549922.3.

Sequence analysis of HTS reads

HTS reads by 454 and Illumina were quality checked following similar criteria as detailed in the original papers (32, 33). After the quality control, chimera detection was run with UCHIME (40) and ChimeraSlayer (41) using SILVA108 and PR^2 (42) as reference databases. The final curated reads were clustered into OTUs (Operational Taxonomic Units) by using UCLUST 1.2.22 (43) with similarity thresholds of 97% for V4-reads and 95% for V9-reads. Representative reads of each OTU were taxonomically classified by using BLAST against SILVA108, PR^2 and a marine microeukaryote database (44). After the taxonomic assignment, metazoan OTUs were removed. From the complete OTU tables for 454 (32) and Illumina datasets (33), the samples targeting the picoplankton were extracted: 13 samples for DNA-V4, 13 samples for cDNA-V4 and 9 samples for cDNA-V9. Then, OTUs corresponding to taxa typically larger than 3 µm (Dinophyceae, Ciliophora, Acantharia, Diatomea, Polycystinea, Raphydophyceae, Ulvophyceae, Rodophyta and Xanthophyceae; in this order of relative abundance) were removed. These groups accounted for 8.0% to 87.7% (average of 36.9%) of the 454 dataset and 11.5% to 73.5% (average of 33.9%) of the Illumina dataset. The read number in the final OTU tables of picoeukaryotes was 110,258 for DNA-V4, 77,554 for cDNA-V4 and 1,753,600 for cDNA-V9.

The relative abundance of the picoeukaryotic groups of interest was retrieved from these taxonomically classified OTU tables, by dividing the number of reads of the specific OTUs corresponding to the groups of interest by the total number of reads in the sample. Altogether, the six taxa of interest accounted for 36.4% of the DNA-V4 reads, 23.5% of the cDNA-V4 reads, and 32.4% of the cDNA-V9 reads. Besides the taxonomic classification of OTUs in the OTU table, we did an additional classification of the unclustered 454 and Illumina reads, to obtain the raw reads for probe checking (see results) and to double-check the taxonomic classification. For this second classification we downloaded GenBank sequences representative of each picoeukaryotic group of interest and used this specific taxa-database to retrieve HTS reads by local BLAST (sequence similarity >97%).

Results

An overview of total picoeukaryote counts in marine coastal waters

We estimated the total cell abundance of picoeukaryotes by epifluorescence microscopy and flow cytometry in 13 planktonic samples taken in six geographically separated European coastal sites and different depths (Table 1). Total picoeukaryote counts (cells <3 μm) by epifluorescence microscopy of DAPI-stained samples revealed a wide range of cell abundances, from 3,139 cells ml^{-1} in Naples-2010 DCM to 24,346 cells ml^{-1} in Oslo-2010 DCM (average of 10,500 cells ml^{-1} in all samples). Phototrophic and heterotrophic cells were differentiated while counting the DAPI samples. The total abundance of phototrophic cells was generally higher than heterotrophic cells (average of of 8,200 and 2,400 cells ml^{-1}, respectively), with the exception of Naples-2010 Surface, where both assemblages have similar abundances. In some cases (Blanes, Oslo-2010 DCM, Roscoff and Varna DCM) phototrophic cells were >6 times more abundant than heterotrophic cells. Counts of phototrophic picoeukaryotes obtained by flow cytometry correlated well with the microscopic counts in the 10 samples analyzed (linear slope = 0.74, Pearson's r = 0.9, $P < 0.001$). When forcing the regression line to intercept at 0, the slope was 0.90.

The general eukaryotic probe EUK502 was also used to estimate total picoeukaryotic abundance. Cell counts by TSA-FISH were always lower than DAPI counts (60% on average) (Fig. 1). In fact, the sample with the highest total cell abundance was different if estimated by DAPI (Oslo-2010 DCM) or by TSA-FISH (Oslo-2009 Surface). The regression between both datasets was significant, but with a slope very distant from 1 (linear slope = 0.26, Pearsons' r = 0.74, $P < 0.05$). When forcing the line to intercept at 0, the slope was still very low, 0.43. There was some tendency to this discrepancy, as TSA-FISH seemed to underestimate more severely the total cell counts in samples dominated by very small cells. Clearly, DAPI counts provided a better estimate of total picoeukaryotic abundance than TSA-FISH counts, and therefore DAPI counts were used to calculate the relative cell abundances of each of the 6 specific picoeukaryotic groups: TSA-FISH counts of each group were at the numerator and total DAPI counts at the denominator.

Abundance of specific picoeukaryotic taxa

We used TSA-FISH to estimate the total abundance of six groups of picoeukaryotes, chosen because they were well represented in the sequencing datasets of the picoplankton from the studied samples (and poorly represented in the nanoplankton, Table S1). They belonged to different eukaryotic supergroups: the Stramenopiles (MAST clades and Pelagophyceae), Alveolates (the parasite clade MALV-II), Archaeplastida (*Micromonas* spp.) and Rhizaria (*Minorisa minuta*). The taxonomic coverage of the used probes varied from being very narrow targeting a species (*Minorisa minuta*) or a constrained phylogenetic clade (*Micromonas* spp. and the MAST lineages), to being very wide targeting an algal class (Pelagophyceae) or the diverse MALV-II group (formed by 44 phylogenetic clades). The sum of heterotrophic cells (MASTs, *M. minuta* and MALV-II) represented on average 36% of heterotrophic picoeukaryotes counted by DAPI, whereas the phototrophic cells targeted (*Micromonas* and Pelagophyceae) represented on average only 22% of phototrophic picoeukaryotes (Table 1).

The cell abundance of the six targeted groups varied strongly among the different samples (Table S2). We found *Micromonas*, MAST-4, MAST-7 and MALV-II as the most abundant taxa (averaged cell abundances of 1492, 279, 160, and 127 cells ml^{-1}, respectively), detected in all samples. *Minorisa minuta* was very abundant in some sites, but absent in others. By contrast, Pelagophyceae was the least abundant taxa (averaged cell abundances of 59 cells ml^{-1}). These cell counts pointed out that each sample contained a different community. *Micromonas* was the most abundant taxa in 7 samples, MAST-4 in 4 samples and *Minorisa* and MALV-II in the other two samples (Table S2).

In-silico validation of the FISH probes against raw V4-reads

Before applying TSA-FISH, we evaluated the effectiveness of the probes against the V4-reads obtained from the same samples. This analysis was done with raw reads (extracted from the initial dataset by using GenBank sequences of each group as search templates) to take into account all sequence variants. The number of raw reads per group obtained from this way was very similar to the number derived from the OTU table (Table 2). About 1000 to 3000 reads were extracted per group (except for MALV-II, about 30,000 reads). Then, we calculated the percentage of raw reads having a 100% match with the probes (Table 2). The five specific probes validated this way retrieved a very high percentage of reads, more than 95% in all cases except in MALV-II (83%). Therefore, the vast majority of reads from these five groups in our samples had the target region of the probes.

The probe targeting *Micromonas* was not designed at the V4 region of the 18S rDNA, so it could not be directly evaluated with V4-reads from this study. Therefore, we took the OTUs affiliating to *Micromonas* (7 OTUs and 11,166 reads), retrieved the closest GenBank complete sequence from these OTUs (nearly identical at the V4 region), and verified the efectiveness of the probe against these 7 GenBank sequences. Only 3 sequences (accounting for 30% of the reads) exhibited a perfect match, whereas the remaining 4 sequences had a mismatch in the first position of the probe. Thus, probe MICRO01 could be improved perhaps removing the first base, but since this mismatch is located in the first position it likely does not affect the FISH counts.

Comparison of group specific read abundance and TSA-FISH counts

The relative abundance of 454 V4-reads (from DNA and cDNA templates) and illumina V9-reads (from cDNA templates) of each group of interest was compared with the relative cell abundance assessed by epifluorescence microscopy (specific TSA-FISH counts relative to total DAPI counts) in 13 samples for the V4-reads, and 9 samples for the V9-reads (DCM samples from Naples and Oslo were excluded) (Fig. 2). The statistics of these plots are shown in Table 3. For the DNA-V4, the correlation of the relative abundance of cells and DNA reads was significant for all groups ($p < 0.05$) except for MAST-4 and Pelagophyceae, and the goodness of these correlations varied among groups, being strongest for *Minorisa minuta* ($R^2 = 0.97$) and weakest for MALV-II ($R^2 = 0.29$). Despite these good correlations, linear slopes of the plots were always different from 1 except in MAST-7. In most cases slopes were below 0.5, indicating an underestimation of cell abundance by 454 reads, while in MALV-II the slope was very high (4.46), indicating a severe overestimation of the molecular signal in this group.

By contrast, the correlations between relative cell and read abundances in the cDNA-V4 survey were generally better for all groups, being also significant for Pelagophyceae and MAST-4 (Table 3). Similar to the DNA-V4 survey, each group had a different slope, but in this case there were three taxa (MAST-7, *M. minuta* and *Micromonas*) with slopes statistically not different from 1, indicating that their relative abundances obtained by cell counts and 454 reads were comparable. In the six groups analysed, the slopes obtained in the cDNA survey were closer to 1 than the slopes derived from the DNA survey, showing a better performance of the cDNA approach.

For the Illumina cDNA-V9 survey, the correlations were slightly worse than for the cDNA-V4 survey (Fig. 2, Table 3), as they were non-significant ($p > 0.05$) for MAST-4 and

MAST-7. Regarding the linear slopes, the three groups with a good performence at the cDNA-V4 survey –*M. minuta*, Pelagophyceae and *Micromonas*– had slopes statistically different from 1, indicating that in these groups the V4 region (and not the V9) could be used as a proxy of cell counts. On the contrary, MALV-II had a better correlation with the V9-cDNA reads than with the V4-reads, and its slope was not statistically different from one. This highlights that there is not a "best region" that applies to all taxa.

Differences when targeting V4 and V9 regions of the 18S rDNA

To discard that the differences observed between the V4 and the V9 regions were due to the use of different sequencing platforms (454 for V4 and Illumina for V9), we sequenced with Illumina (MiSeq platform) the V4 region of one sample of the dataset (Oslo-2009 DCM) using both templates (DNA and cDNA). The relative abundance of ~60 taxonomic groups inferred from the same targeted region (V4) in the two platforms displayed a very good agreement, with R^2 of 0.97 and 0.91 (for DNA and cDNA, respectively), and linear slopes of 0.92 to 1.02. Both slopes were not significantly different from 1. Furthemore, this analysis was also done in an additional set of 14 samples (from other planktonic size fractions and sediments; data not shown) and both platforms performed similarly, with R^2 ranging from 0.57 to 1.00 (average of 0.91) and slopes ranging from 0.73 to 1.21 (average of 0.99). Therefore, sequencing the same 18S rDNA region with 454 or Illumina (MiSeq) gave highly consistent results.

Therefore, the differences outlined above between V4-454 and V9-Illumina sequencing (Table 3) were due to targeting different 18S rDNA regions and not due to the sequencing platform. In order to observe these differences in more detail, we compared the relative abundance of cDNA-V4 reads and cDNA-V9 reads for the six picoeukaryotic taxa studied here (Fig. 3). Clear and consistent differences were identified in each case. As before, the correlations were good and significant, with R^2 ranging from 0.68 to 0.98 (being MALV-II lower, 0.45), but the slopes deviated significantly from 1 ($p < 0.05$). The V9 analysis increased significantly the relative abundance of the stramenopile groups (the two MAST clades and Pelagophyceae), with slopes ranging from 2.3 to 3.4 while it was the opposite for *Micromonas* and MALV-II (slope of 0.2 and 0.3, respectively) and the same for *Minorisa minuta* (slope of 1.1).

Discussion

Identifying marine picoeukaryotes by direct microscopy is problematic because of their small sizes, and as a consequence there is an increasing interest in using high-throughput sequencing (HTS) technologies to explore their diversity. HTS surveys provide a detailed picture of the taxa present in the community, including rare species in the assemblage (18, 33), and reveal diversity not evident using other methods. However, the interpretation of the HTS signal in terms of total cell abundances is not straightforward. Interestingly, TSA-FISH is able to bridge microscopic and sequencing approaches by using specific phylogenetic probes to estimate true cell abundances (28, 45). FISH, besides being very laborious, is also limited by the number of taxa-specific probes available as well as by their phylogenetic resolution (46). Moreover, TSA-FISH could be inaccurate due to putative mismatches of the probes with the target group, which would result in cell counts underestimates. We addressed this issue by evaluating the six probes against sequences obtained from the same samples, and found an acceptable performance (very good in four cases, 83% of reads for MALV-II and only one terminal mismatch for *Micromonas*). This validated that the TSA-FISH cell counts performed here were accurate and enabled the main objective of this study, which was to evaluate how well the HTS signal estimates community structure in terms of specific abundance.

More sequences imply more cells

Since the HTS signal is always relative (number of reads of a given taxa respect to the total read number), we needed the total picoeukaryote abundance to calculate relative cell abundances. In principle, using TSA-FISH with a universal eukaryotic probe would be consistent with the study and would also provide an extra layer of certainty, since it allows an easier differentiation of eukaryotic cells from fluorescent particles and large bacteria. However, TSA-FISH counts systematically resulted in fewer cells than direct DAPI counts, and we noticed protists that were not labeled with the EUK502 probe. Moreover, this discrepancy was particularly critical in samples dominated by very small cells. The wide size spectra of protist cells in natural samples implied a large variation in the fluorescent signal, so small cells with dim fluorescence may remain unnoticed when close to large fluorescent cells, and easily faded away while counting a field having many cells with diverse sizes and morphologies. This problem did not happen when using specific probes, since then we focused in counting a defined cell type (even with dim fluorescence). Therefore, we used the direct DAPI counts to calculate relative cell abundances.

When comparing the relative abundance of HTS reads against the relative cell abundance obtained by TSA-FISH for the different taxa, we generally found a good correlation between both methods. The R^2 coefficients of each picoeukaryotic taxa were similar in the three comparisons conducted (DNA-V4, cDNA-V4 and cDNA-V9 vs. TSA-FISH), except a very poor correlation for Pelagophyceae in the DNA-V4 survey. Nevertheless, the statistical significance was always better for the cDNA survey than for the DNA. These correlations imply that relative read abundance was proportional to relative cell abundance, i.e. an increase in the HTS signal from a particular taxon is the result of an increase of the proportion of targeted cells in the sample. However, the correlation coefficients were far from 1 in most cases, and this noisy signal was probably related to molecular biases plus the large differences in the picoeukaryotic composition of each sample.

Molecular surveys based on a single gene are affected by the widely discussed PCR biases (47). During PCR, some phylotypes can be amplified preferentially, some groups can remain undetected due to primer mismatches (48) or there could be biases due to the number of PCR cycles (49). So, it has been suggested that the relative read abundance can no longer reflect the real composition of the original community, biasing diversity estimates and producing over or underestimations of specific groups (2). Furthermore, sequencing errors may create false or chimeric taxa (16, 50, 51). Our results indicate that PCR biases and putative sequencing artifacts are not affecting proportionality between relative read and cell abundance: more reads imply high proportion of cells. The significant correlations detected here using this sample dataset, where each sample has large differences in the picoeukaryotic composition because they were taken in distant sites and times of the year, justifies the use of relative read abundance as a proxy of community composition for comparative purposes.

Relative abundances of sequences and cells often disagree

Despite the significant correlations discussed above, HTS and TSA-FISH surveys did not give the same quantitative information, as often the regression line was statistically different from 1. Moreover, these slopes varied strongly among the three HTS surveys. In order to compare these surveys, we analyzed the relative abundances of the six picoeukaryotic groups (among themselves) in the different samples (Fig. 4). This showed a general agreement between TSA-FISH and the two cDNA surveys, but depending on the composition of the sample, the agreement was better using the V4 region or the V9 region. In samples dominated by *Micromonas* (e.g Blanes, Oslo-2010, Roscoff, Varna DCM), the picture obtained with the V4 region matched better the cell abundance, while in samples dominated by stramenopiles (MAST-4, MAST-7, Pelagophyceae), the V9 region performed better. In our samples, the cDNA-V4 survey gives a better representation of the true species composition for 5 of the samples while cDNA-V9 performed better in 4 of the samples.

In all cases, the DNA survey gave a more biased perspective of the relative abundance of the 6 picoeukaryotic taxa, being influenced by a very high abundance of MALV-II reads in all samples. This is probably due to a particularly high number of rDNA-operon copies in MALV groups (2, 30, 32). The SSU rDNA copy number can vary orders of magnitude among protist taxa, from few copies per cell in some green algae (52) to about 30 copies in MAST-4 (53) or several thousand copies in some dinoflagellates (52), depending on the cell size and genome size (54). Large differences in the copy number of the targeted gene will affect the abundance estimates in DNA surveys (2). Moreover, reads retrieved in DNA surveys could derive from dead organisms or dissolved extracellular DNA. It is known that dissolved DNA is preserved in marine waters (55), escaping from degradation and persisting for different periods of time, from hours to days (56). On the contrary, reads from cDNA surveys derive from ribosomes and represent metabolically active taxa in the community, as ribosomes are needed to perform the RNA translation in metabolically active cells (57, 58). This, in addition to the SSU rDNA copy number, could explain the differences observed between DNA and cDNA surveys. Moreover, our data also highlighted the impact of targeting different regions of the 18S rDNA gene for estimating relative abundances. For example, the cDNA-V9 survey showed a higher signal (more reads) for MAST taxa and a lower signal for *Micromonas* when compared with cDNA-V4. It is known that the range of taxonomic groups detected by V4 and V9 is different (38, 59, 60) and that some groups can be over- or underrepresented. In particular, in our samples the V4 region gives good estimates of cell counts for MAST-7 and

Micromonas spp., the V9 for MALV-II, and both regions for *Minorisa minuta*. So, the region targeted (and the primers used) is fundamental to interpret any existing molecular data.

Concluding remarks

To our knowledge, this is the first study investigating the correspondence between HTS and cell counts for selected and relevant taxa of marine picoeukaryotes. Indeed, true cell abundances of picoeukaryotic taxa require the TSA-FISH approach, but as this approach has inherent limitations (is time consuming, few probes are available, fine resolution can not be provided), we see the need to pursue with HTS studies. Our results indicate a good correlation between both methods, implying that more cells results in more sequences, although they give different quantitative information, i.e. the relative read abundance cannot be directly related to relative cell abundance. The cDNA-V4 survey showed the best agreement with TSA-FISH abundance, providing 1:1 relationships in half of the assayed taxa, but the cDNA-V9 was best for other taxa. So, the region of the 18S rDNA gene targeted clearly affected the relative abundance of specific taxa. Finally, based in the data mentioned here, we suggest that the sequencing platform used (454 or Illumina) does not produce major biases in diversity. In conclusion, the most quantitative option is to use cDNA templates rather than DNA, while the choice of the targeted region will result in different relative abundances in each particular taxa.

Acknowledgments

We thank Marta Vila for her help on the TSA-FISH counts. We thank the editor and four anonymous reviewers for their critical reading and constructive comments.

Funding information

This work was supported by the European project BioMarKs (2008-6530, ERA-net Biodiversa, EU), DEVOTES (FP7-ENV-2012, 308392, EU), MICRO-3B (FP7-OCEAN-2011 287589, EU) and MEFISTO (CTM2013-43767-P, MINECO). The funding agencies had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We declare no conflicts of interest.

References

1. **Boenigk J, Arndt H**. 2002. Bacterivory by heterotrophic flagellates. Antonie Van Leeuwenhoek **81**:465–480.

2. **Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J**. 2010. Diversity in a hidden world: Potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol Ecol **19**:32–40.

3. **Sherr EB, Sherr BF**. 2002. Significance of predation by protists in aquatic microbial food webs. Antonie Van Leeuwenhoek **81**:293–308.

4. **López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D**. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature **409**:603–607.

5. **Moon-van der Staay SY, De Wachter R, Vaulot D**. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409**:607–10.

6. **Amaral-Zettler LA, Gómez F, Zettler E, Keenan BG, Amils R, Sogin ML**. 2002. Microbiology: eukaryotic diversity in Spain's River of Fire. Nature **417**:137.

7. **Dawson SC, Pace NR**. 2002. Novel kingdom-level eukaryotic diversity in anoxic environments. Proc Natl Acad Sci USA **99**:8324–8329.

8. **Stoeck T, Epstein S.** 2003. Novel eukaryotic lineages inferred from small-subunit rRNA analyses of oxygen-depleted marine environments. Appl Environ Microbiol **69**:2657–2663.

9. **Berney C, Fahrni J, Pawlowski J**. 2004. How many novel eukaryotic "kingdoms"? Pitfalls and limitations of environmental DNA surveys. BMC Biol. **2**:13.

10. **Lovejoy C, Massana R, Pedrós-Alió C**. 2006. Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. Appl Environ Microbiol **72**:3085–3095.

11. **Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ**. 2007. Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. Environ Microbiol **9**:1233–52.

12. **Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, Scanlan DJ, Worden AZ**. 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environ Microbiol **10**:3349–3365.

13. **Massana R, Pedrós-Alió C**. 2008. Unveiling new microbial eukaryotes in the surface ocean. Curr Opin Microbiol **11**:213–8.

14. **Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK**. 2012. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Mol Ecol **21**:1048–1059.

15. **Savin MC, Martin JL, LeGresley M, Giewat M, Rooney-Varga J**. 2004. Plankton diversity in the bay of fundy as measured by morphological and molecular methods. Microb Ecol **48**:51–65.

16. **Bachy C, Dolan JR, López-García P, Deschamps P, Moreira D**. 2012. Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. ISME J. **7**:244–255.

17. **Santoferrara LF, Grattepanche JD, Katz LA, McManus GB**. 2014. Pyrosequencing for assessing diversity of eukaryotic microbes: Analysis of data on marine planktonic ciliates and comparison with traditional methods. Environ Microbiol **16**: 2752-2763

18. **Stoeck T, Breiner HW, Filker S, Ostermaier V, Kammerlander B, Sonntag B**. 2014. A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. Environ Microbiol **16**:430–444.

19. **Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B**. 2013. 454 Pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: A test for marine Haptophytes. PLoS One **8**: e74371.

20. **Weber AAT, Pawlowski J**. 2013. Can abundance of Protists be inferred from sequence data: A case study of Foraminifera. PLoS One **8**:1–8.

21. **Massana R**. 2011. Eukaryotic picoplankton in surface oceans. Annu Rev Microbiol **65**:91–110.

22. **Porter KG, Feig YS**. 1980. The use of DAPI for identifying aquatic microfloral. Limnol Oceanogr **25**:943–948.

23. **Marie D, Partensky F, Simon N, Guillou L, Vaulot D.** 2000. Flow cytometry analysis of marine picoplankton. In Living Colors: Protocols in Flow Cytometry and Cell Sorting. Diamond, R.A. and DeMaggio, S. (eds). Springer Verlag, Berlin. pp. 421–454.

24. **Potter D, Lajeunesse T**. 1997. Convergent evolution masks extensive biodiversity among marine coccoid picoplankton. Biodiv Conserv **107**:99–108.

25. **Delong EF, Wickham GS, Pace NR**. 1989. Phylogenetic stains: Ribosomal RNA-based probes for the idenfication of single cells. Science **243**:1360–1363.

26. **Massana R, Guillou L, Díez B, Pedrós-Alió C**. 2002. Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. Appl Environ Microbiol **68**:4554–4558.

27. **Chambouvet A, Morin P, Marie D, Guillou L**. 2008. Control of toxic marine dinoflagellate blooms by serial parasitic killers. Science **322**:1254–1257.

28. **Not F, Latasa M, Marie D, Cariou T, Vaulot D, Simon N**. 2004. A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. Appl Environ Microbiol **70**:4064–4072.

29. **Massana R, Terrado R, Forn I, Lovejoy C, Pedrós-Alió C**. 2006. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. Environ Microbiol **8**:1515–1522.

30. **Siano R, Alves-de-Souza C, Foulon E, Bendif EM, Simon N, Guillou L, Not F.**. 2010. Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. Biogeosciences **8**:267–278.

31. **Lin YC, Campbell T, Chung CC, Gong GC, Chiang KP, Worden AZ**. 2012. Distribution patterns and phylogeny of marine stramenopiles in the North Pacific Ocean. Appl Environ Microbiol **78**:3387–3399.

32. **Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, Chambouvet A, Christen R, Claverie JM, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Forn I, Forster D, Guillou L, Jaillon O, Kooistra W, Logares R, Mahé F, Not F, Ogata H, Pawlowski J, Ernice MC, Probert I, Romac S, Richards T, Santini S, Shalchian-Tabrizi K, Siano R, Simon N, Stoeck T, Valuot D, Zingone A, de Vargas C.** 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ Microbiol **17:**4035–4049.

33. **Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie JM, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Gobet A, Kooistra WHCF, Mahé F, Not F, Ogata H, Pawlowski J, Pernice MC, Romac S, Shalchian-Tabrizi K, Simon N, Stoeck T, Santini S, Siano R, Wincker P, Zingone A, Richards TA, de Vargas C, Massana R**. 2014. Patterns of rare and abundant marine microbial eukaryotes. Curr Biol **24**:813–821.

34. **Marie D, Partensky F, Vaulot D, Brussaard C.** 1999. Enumeration of phytoplankton, bacteria, and viruses in marine samples. Current Protocols in Cytometry. John Wiley & Sons, Inc.

35. **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. Nucleic Acids Res **41**:590–596.

36. **Lim EL, Dennett MR, Caron DA.** 1999. The ecology of *Paraphysomonas imperforata* based on studies employing oligonucleotide probe identification in coastal water samples and enrichment cultures. Limnol Oceanogr **44**:37–51.

37. **Pernice MC, Forn I, Gomes A, Lara E, Alonso-Sáez L, Arrieta JM, del Carmen Garcia F, Hernando-Morales V, MacKenzie R, Mestre M, Sintes E, Teira E, Valencia J, Varela MM, Vaqué D, Duarte CM, Gasol JM, Massana R**. 2015. Global abundance of planktonic heterotrophic protists in the deep ocean. ISME J. **9**:782–792.

38. **Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, Richards TA.** 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol **19**: 21–31.

39. **Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM**. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One **4**:1–9.

40. **Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R**. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**:2194–2200.

41. **Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ, Petrosino JF, Knight R, Birren BW**. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res **21**:494–504.

42. **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. Nucleic Acids Res **41**:597–604.

43. **Edgar RC**. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**:2460–2461.

44. **Pernice MC, Logares R, Guillou L, Massana R**. 2013. General patterns of diversity in major marine microeukaryote lineages. PLoS One **8**: e57170.

45. **Not F, Simon N, Biegala IC, Vaulot D.**. 2002. Application of fluorescent in situ hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. Aquat Microb Ecol **28**:157–166.

46. **Alonso-Sáez L, Balagué V, Sà EL, Sánchez O, González JM, Pinhassi J, Massana R, Pernthaler J, Pedrós-Alió C, Gasol JM**. 2007. Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. FEMS Microbiol Ecol **60**:98–112.

47. **Wintzingerode FV, Göbel UB, Stackebrandt E**. 1997. Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev **21**:213–229.

48. **Hong S, Bunge J, Leslin C, Jeon S, Epstein S.** 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J. **3**: 1365-1373.

49. **Suzuki M, Rappé MS, Giovannoni SJ**. 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. Appl Environ Microbiol **64**:4522–4529.

50. **Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT**. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods **6**:639–641.

51. **Kunin V, Engelbrektson A, Ochman H, Hugenholtz P**. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol **12**:118–123.

52. **Zhu F, Massana R, Not F, Marie D, Vaulot D**. 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol Ecol **52**:79–92.

53. **Rodríguez-Martínez R, Labrenz M, del Campo J, Forn I, Jürgens K, Massana R.** 2009. Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. Environ Microbiol **11**:397-408.

54. **Prokopowich CD, Gregory TR, Crease TJ**. 2003. The correlation between rDNA copy number and genome size in eukaryotes. Genome **46**:48–50.

55. **Danovaro R, Corinaldesi C, Dell'Anno A, Fabiano M, Corselli C.** 2005. Viruses, prokaryotes and DNA in the sediments of a deep-hypersaline anoxic basin (DHAB) of the Mediterranean Sea. Environ Microbiol **7(4)**:586–592.

56. **Nielsen KM, Johnsen PJ, Bensasson D, Daffonchio D**. 2007. Release and persistence of extracellular DNA in the environment. Environ Biosafety Res. **6**:37–53.

57. **Stoeck T, Zuendorf A, Breiner HW, Behnke A**. 2007. A molecular approach to identify active microbes in environmental eukaryote clone libraries. Microb Ecol **53**:328–339.

58. **Not F, del Campo J, Balagué V, de Vargas C, Massana R**. 2009. New insights into the diversity of marine picoeukaryotes. PLoS One **4**:e7143.

59. **Dunthorn M, Klier J, Bunge J, Stoeck T**. 2012. Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. J. Eukaryot Microbiol **59**:185–7.

60. **Decelle J, Romac S, Sasaki E, Not F, Mahé F**. 2014. Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. PLoS One **9**: e104297.

Figure legends

Fig. 1. Comparison of total picoeukaryotic abundance (cells <3 µm) by DAPI counts and TSA-FISH counts using the eukaryotic probe EUK502 in all planktonic samples.

Fig. 2. Comparison of relative abundance of HTS reads against TSA-FISH cell counts in the 13 planktonic samples (9 samples for cDNA-V9 reads) for six picoeukaryotic taxa: MAST-4 (a), MAST-7 (b), *Minorisa minuta* (c), Pelagophyceae (d), *Micromonas* spp. (e) and MALV-II (f). Dark blue symbols indicate DNA-V4 reads, light blue cDNA-V4 reads and green cDNA-V9 reads. Regression lines are shown, and their statistics are presented in Table 3.

Fig. 3. Comparison of relative abundance of V9-Illumina reads and V4-454 reads (cDNA surveys in both cases) in 9 planktonic samples for six picoeukaryote taxa: MAST-4 (a), MAST-7 (b), *Minorisa minuta* (c), Pelagophyceae (d), *Micromonas* spp. (e) and MALV-II (f).

Fig. 4. Relative abundance of the different groups (among themselves) shown by the four approaches (TSA-FISH, cDNA-V4, DNA-V4, cDNA-V9) in all planktonic samples. Gray bars indicate the absence of the sample.

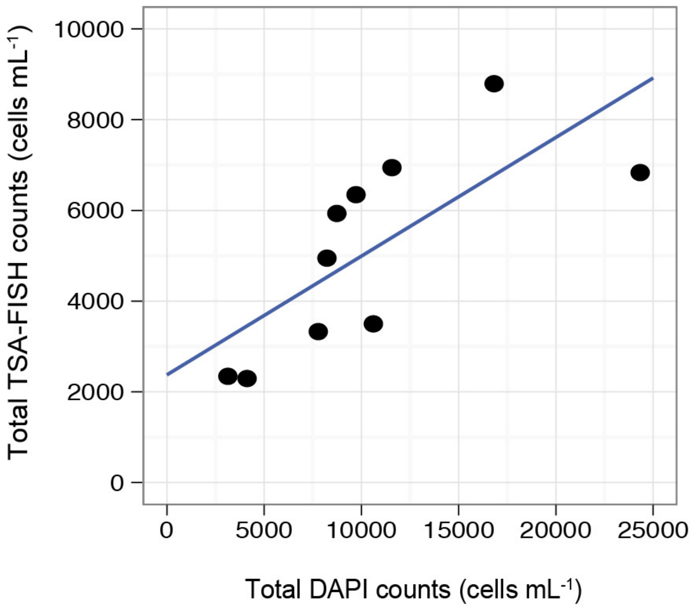

Fig. 1- Comparison of total picoeukaryotic abundance (cells <3 µm) by DAPI counts and TSA-FISH counts using the eukaryotic probe EUK502 in all planktonic samples.

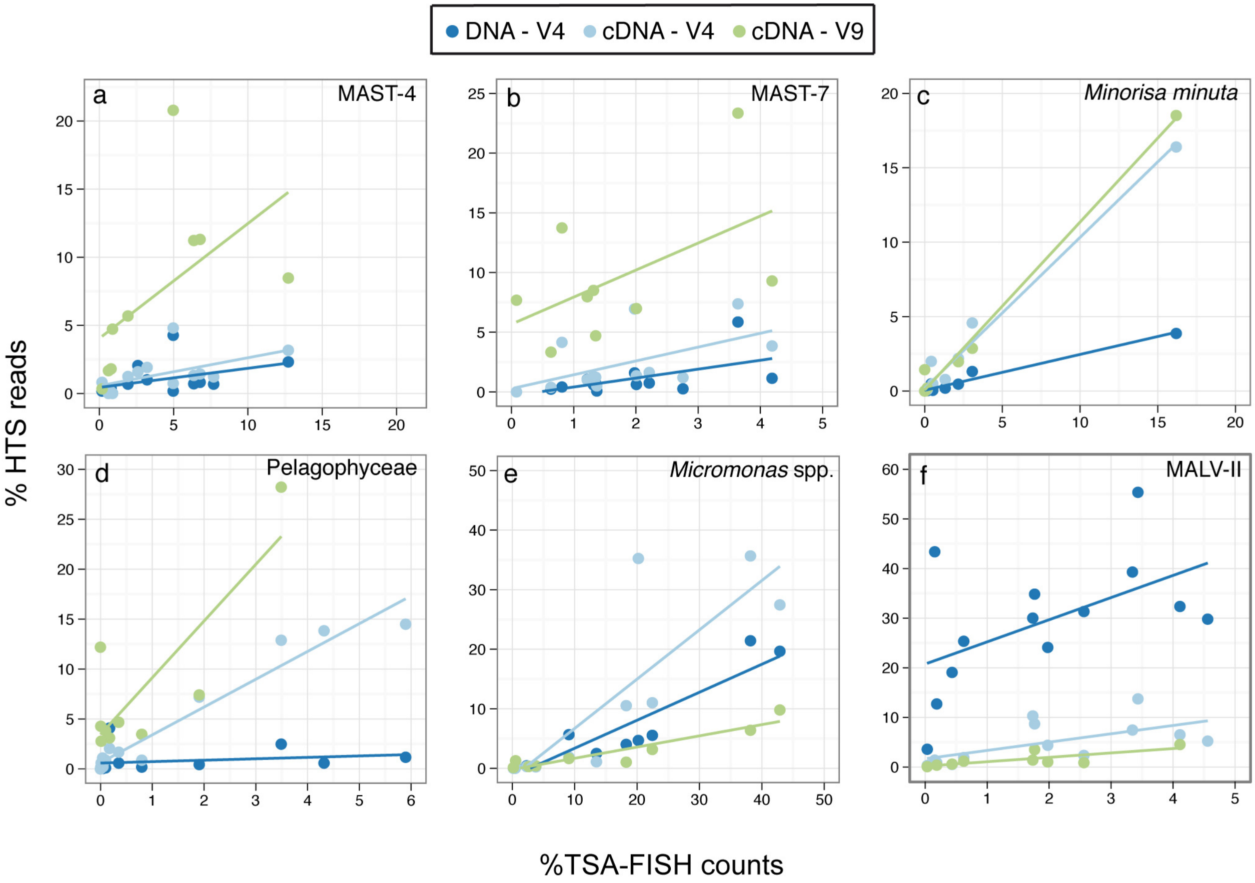

Fig. 2- Comparison of relative abundance of HTS reads against TSA-FISH cell counts in the 13 planktonic samples (9 samples for cDNA-V9 reads), for six picoeukaryotic taxa: MAST-4 (a), MAST-7 (b), *Minorisa minuta* (c), Pelagophyceae (d), *Micromonas* spp. (e) and MALV-II (f). Dark blue symbols indicate DNA-V4 reads, light blue cDNA-V4 reads and green cDNA-V9 reads. Regression lines are shown, and their statistics are presented in Table 3.

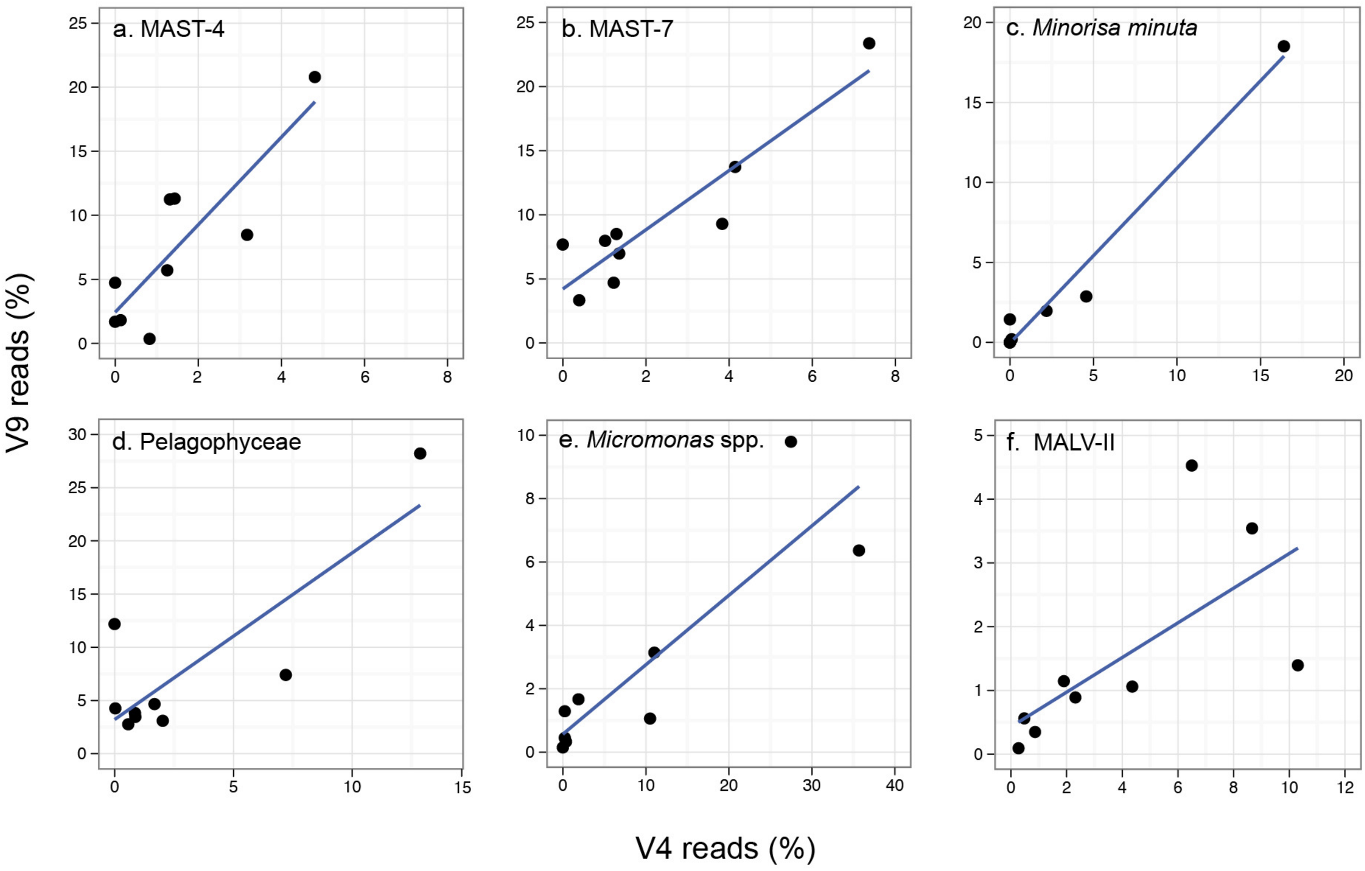

Fig. 3 - Comparison of relative abundance of V9-llumina reads and V4-454 reads (cDNA surveys in both cases) in 9 planktonic samples for six picoeukaryote taxa: MAST-4 (a), MAST-7 (b), *Minorisa minuta* (c), Pelagophyceae (d), *Micromonas* spp. (e) and MALV-II (f).

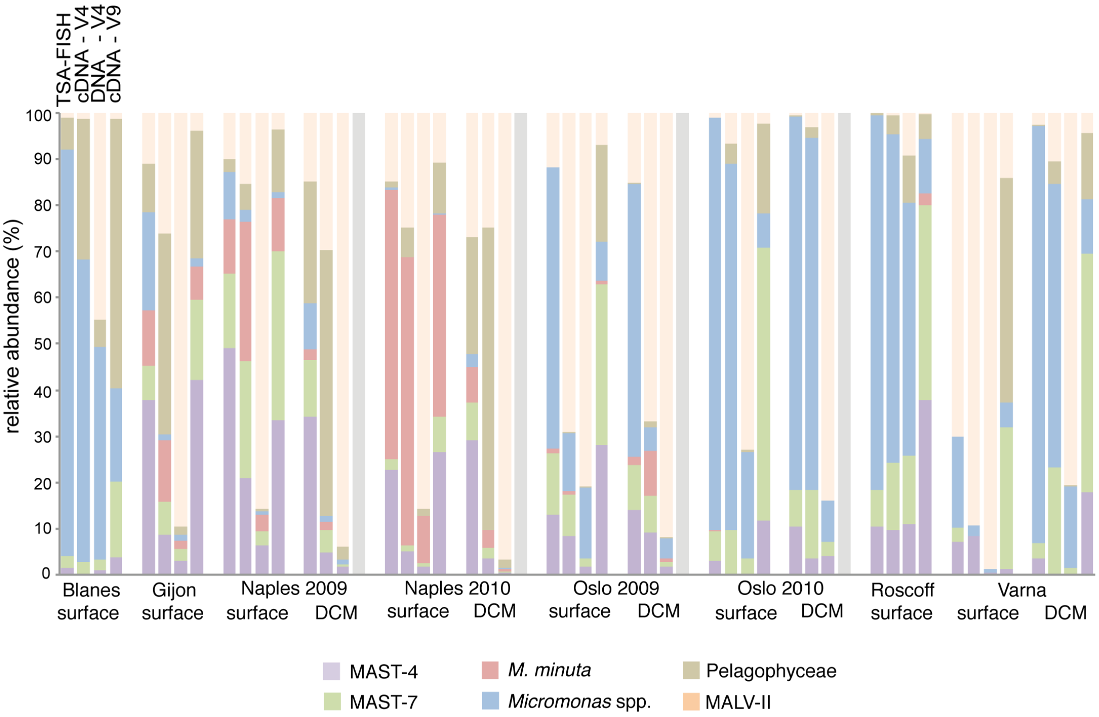

Fig. 4 - Relative abundance of the different groups (among themselves) shown by the four approaches (TSA-FISH, cDNA-V4, DNA-V4, cDNA-V9) in all planktonic samples. Gray bars indicate the absence of the sample.

Table 1. Planktonic samples analyzed (sampling site, date, depth and seawater temperature) and cell counts (cells ml^{-1}) in these samples: total picoeukaryote abundance (cells ≤3 µm) determined by DAPI (phototrophs and heterotrophs), and photosynthetic picoeukaryote abundance determined by flow cytometry (FC). The last two columns show the percentage of phototrophic and heterotrophic cells explained by the probes used.

Sampling site	**Date**	**Depth (m)**	**Temp. (ºC)**	**DAPI counts**		**FC counts**	**% Phototr.**	**% Heterotr.**
				Phototr.	Heterotr.	Phototr.		
Blanes	Feb. 2010	1 (Surf.)	12.5	9273	445	9215	48.6	53.7
Gijon	Sep. 2010	1 (Surf.)	20.2	1606	2503	2990	14.5	20.2
Naples	Oct. 2009	1 (Surf.)	22.8	*	*	2714	-	-
		26 (DCM)	22.4	*	*	2049	-	-
	May 2010	1 (Surf.)	19.2	4376	4372	4700	1.1	54.6
		34 (DCM)	15.5	1808	1331	1802	8.3	28.8
Oslo	Sep. 2009	1 (Surf.)	15.0	12342	4470	9540	12.4	21.9
		20 (DCM)	15.0	8773	2807	8930	17.9	38.4
	Jun. 2010	1 (Surf.)	15.0	7727	2893	13295	25.5	7.9
		10 (DCM)	12.5	21523	2823	17900	22.9	40.7
Roscoff	Apr. 2010	1 (Surf.)	9.9	7203	1034	8240	43.9	68.9
Varna	May 2010	1 (Surf.)	21.5	*	*	3861	-	-
		40 (DCM)	9.5	7043	731	9487	24.9	24.6

* DAPI counts were not performed, so picoeukaryotes could not be differentiated between phototrophs and heterotrophs. In these samples, total picoeukaryote counts were done on FISH filters and were: 4272 cells ml^{-1} in Naples-2009 Surf, 1834 cells ml^{-1} in Naples-2009 DCM, and 4656 cells ml^{-1} in Varna Surf. These values were used in the correlations.

Table 2. List of oligonucleotide FISH probes used and effectiveness of the probes against reads from this study (% reads-probe). The table shows the number of 454 reads from each phylogenetic group extracted from the OTU table or from raw reads by local BLAST. The last column shows the percentage of raw reads in each group that have the probe target region with 0 mismatches.

Probe Name	Target group	Probe sequence (5' – 3')	Probe reference	Num. of reads per Taxa		% reads - probe
				In OTU table	From the raw reads	
NS4	MAST-4	TACTTCGGTCTGCAAACC	Massana *et al.,* 2002	2082	2082	98.0
NS7	MAST-7	TCATTACCATAGTACGCA	This study	2842	2833	95.7
CRN02	*Minorisa minuta*	TACTTAGCTCTCAGAACC	del Campo *et al.*, 2012	1853	1853	99.8
PELA01	Pelagophyceae	ACGTCCTTGTTCGACGCT	Not *et al*., 2002	4440	3169	98.5
MICRO01	*Micromonas* spp.	AATGGAACACCGCCGGCG	Not *et al.,* 2004	11,166	-	-
ALV01	MALV-II	GCCTGCCGTGAACACTCT	Chambouvet *et al*., 2008	35,359	29,894	83.0
EUK502	Eukaryotes	GCACCAGACTTGCCCTCC	Lim *et al.*, 1999	-	-	-

Table 3. Statistics (R^2, slope value, and p-value) of the correlations between relative abundance of reads and cells in the three molecular surveys: 454 DNA-V4 (Fig. 2, dark blue), 454 cDNA-V4 (Fig. 2, light blue) and Illumina cDNA-V9 (Fig. 2, green). The fourth statistics (p1) compares the slopes against the desired value of 1 (i.e. "ns" indicates that the slope is not significantly different from 1).

	V4 - 454 survey								**V9 - Illumina survey**			
	DNA				**cDNA**				**cDNA**			
	R^2	**slope**	**p-value**	**p1**	**R^2**	**slope**	**p-value**	**p1**	**R^2**	**slope**	**p-value**	**p1**
MAST-4	0.18	0.14	ns	-	0.31	0.21	*	***	0.3	0.84	ns	-
MAST-7	0.33	0.75	*	ns	0.31	1.16	*	ns	0.36	2.79	ns	-
Minorisa minuta	0.97	0.24	***	***	0.98	1.01	***	ns	0.99	1.13	***	***
Pelagophyceae	0.06	0.14	ns	-	0.94	2.78	***	***	0.68	5.68	**	**
Micromonas spp.	0.87	0.47	***	***	0.73	0.83	***	ns	0.87	0.2	***	***
MALV-II	0.29	4.46	*	*	0.39	1.68	*	*	0.60	0.89	*	ns

Significance codes: ***: <0.001; **: 0.001–0.01; *: 0.01–0.05; ns: no significant