The ISME Journal (2011) 5, 184–195

www.nature.com/ismej

ORIGINAL ARTICLE

Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean

Ramon Massana[1], Massimo Pernice[1], John A Bunge[2] and Javier del Campo[1]
[1]*Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain and* [2]*Department of Statistical Science, Cornell University, Ithaca, NY, USA*

Despite the ecological importance of marine pico-size eukaryotes, the study of their *in situ* diversity using molecular tools started just a few years ago. These studies have revealed that marine picoeukaryotes are very diverse and include many novel taxa. However, the amount and structure of their phylogenetic diversity and the extent of their sequence novelty still remains poorly known, as a systematic analysis has been seldom attempted. In this study, we use a coherent and carefully curated data set of 500 published 18S ribosomal DNA sequences to quantify the diversity and novelty patterns of picoeukaryotes in the Indian Ocean. Our phylogenetic tree showed many distant lineages. We grouped sequences in OTUs (operational taxonomic units) at discrete values delineated by pair-wise Jukes–Cantor (JC) distances and tree patristic distances. At a distance of 0.01, the number of OTUs observed (237/242; using JC or patristic distances, respectively) was half the number of sequences analyzed, indicating the existence of microdiverse clusters of highly related sequences. At this distance level, we estimated 600–800 OTUs using several statistical methods. The number of OTUs observed was still substantial at higher distances (39/82 at 0.20 distance) suggesting a large diversity at high-taxonomic ranks. Most sequences were related to marine clones from other sites and many were distant to cultured organisms, highlighting the huge culturing gap within protists. The novelty analysis indicated the putative presence of pseudogenes and of truly novel high-rank phylogenetic lineages. The identified diversity and novelty patterns among marine picoeukaryotes are of great importance for understanding and interpreting their ecology and evolution.
The ISME Journal (2011) **5,** 184–195; doi:10.1038/ismej.2010.104; published online 15 July 2010
Subject Category: microbial population and community ecology
Keywords: diversity; genetic distances; microdiversity; novelty; OTUs; picoeukaryotes

Introduction

Planktonic protists have fundamental roles in the functioning of marine ecosystems, both as primary producers and as microbial grazers (Sherr *et al.*, 2007). Early marine biologists were amazed by the large protist diversity in the plankton, a habitat apparently homogeneous and with a limited range of resources. This phenomenon was named as the paradox of the plankton (Hutchinson, 1961). Today it is assumed that biological and environmental factors interact continually, so that the plankton habitat never reaches an equilibrium, preventing competitive exclusion by a single species and promoting diversity (Scheffer *et al.*, 2003). Little was known for the smallest protists (picoeukaryotes, cells of 0.8–3 μm), which are hardly visible by inverted microscopy. Epifluorescence and flow-cytometry counts (Johnson and Sieburth, 1982; Olson *et al.*, 1985) revealed their abundance, ubiquity and ecological relevance, but still did not allow identification. This was made possible with the introduction of molecular tools to oceanography that provided a culturing and microscopic-independent assessment of microbial diversity (Giovannoni *et al.*, 1990). A series of seminal studies showed that marine picoeukaryotes were indeed very diverse, similar to what was observed for larger protists, and contained many novel lineages (Díez *et al.*, 2001; López-García *et al.*, 2001; Moon-van der Staay *et al.*, 2001). Comparable patterns were also observed in the first molecular surveys of freshwater systems (Lefranc *et al.*, 2005; Richards *et al.*, 2005).

The methodological improvements to retrieve phylogenetically informative genes from the environment have been paralleled by a growing understanding of the eukaryotic tree of life on the basis of

Correspondence: R Massana, Department of Marine Biology and Oceanography, Institut de Ciències del Mar, CSIC, Passeig Marítim de la Barceloneta 37-49, Barcelona, Catalonia 08003, Spain.
E-mail: ramonm@icm.csic.es
Received 26 March 2010; revised 31 May 2010; accepted 13 June 2010; published online 15 July 2010

the cultured organisms. Phylogenetic analyses have confirmed the taxonomic groups defined with cell ultrastructure studies. In addition, phylogenomic analyses have identified a few supergroups composed of eukaryotes with little morphological resemblance, but a common evolutionary origin (Baldauf, 2003). The eukaryotic tree of life was first delineated with eight supergroups, which have been further reduced to six (Simpson and Roger, 2004), or less (Burki *et al.*, 2008). For instance, the supergroup stramenopiles includes lineages as disparate as the diatoms, chrysophytes or bicosoecids and the supergroup opisthokonts includes the choanoflagellates, fungi and metazoans. The eukaryotic tree of life represents an optimal framework to assign environmental sequences to known lineages, or to define new ones if environmental sequences do not find a place. Thus, novel groups, such as marine stramenopiles (MASTs, Massana *et al.*, 2004), marine alveolates (MALVs, Guillou *et al.*, 2008) or picobiliphytes (Not *et al.*, 2007), have been defined on the basis of the environmental surveys. It has been shown that some members of these previously unnoticed lineages are ubiquitous marine grazers, parasites and algae, respectively.

Despite the numerous molecular surveys of marine picoeukaryotes (reviewed in Massana and Pedrós-Alió, 2008; Vaulot *et al.*, 2008), the knowledge about the extent of their diversity at different phylogenetic scales and the pattern of sequence novelty (that is, how different are the environmental sequences from a given study with respect to GenBank sequences) is still in its infancy. Few studies have reported the number of lineages observed grouping sequences at different clustering levels (Caron *et al.*, 2009). Parametric and non-parametric statistics have been used to estimate the total richness in different habitats, including picoeukaryotes from the marine plankton (Brown *et al.*, 2009). Moreover, little has been advanced in quantifying and representing the novelty patterns of sequences from environmental surveys. In this study, we are addressing these issues by using a coherent and curated data set of environmental sequences of picoeukaryotes (500 sequences of $\sim$800 bp). These sequences were just assigned to broad taxonomic groups in a general publication on small protists from the Indian Ocean (Not *et al.*, 2008), hence the diversity and novelty analyses proposed in this study are totally new. Specific questions that arise are the following: How many described taxonomic groups are detected? How many OTUs (operational taxonomic units) are observed when clustering sequences at different thresholds? Is the clustering method affecting the previous question? How many OTUs can be estimated? What is the novelty pattern of environmental sequences? Our study is an effort to describe the diversity and novelty of marine picoeukaryotes by using the data gathered by a classical 18S ribosomal DNA (rDNA) clone library approach, to set up a baseline with which to compare the massive amount of data that are just beginning to be available by high-throughput sequencing (Amaral-Zettler *et al.*, 2009; Brown *et al.*, 2009; Stoeck *et al.*, 2009).

Materials and methods

Sequence data set

Sequences were derived from a recent study conducted in the Indian Ocean (Not *et al.*, 2008). Eight clone libraries of the 18S rDNA genes from the picoplankton (0.2 to 3 μm) were prepared from surface and Deep Chlorophyll Maximum (DCM) samples from stations 01, 09, 18 and 23 (see Figure 1 in Not *et al.*, 2008). Station 01 was coastal, whereas the other three stations (representing 91% of the sequences) were offshore. Details of DNA extraction, PCR (with eukaryotic primers EukA and EukB) and cloning protocols can be found in the original publication. Clones were sequenced with the internal primer 528f, resulting in 572 sequences of around 850 bp each. The taxonomic affiliation of each sequence (including chimera detection) was done by BLAST (Altschul *et al.*, 1997) and KeyDNATools (http://www.keydnatools.com/) searches and compared with published phylogenetic trees. A final data set of 500 protist sequences was obtained after excluding 30 metazoan sequences, 33 chimeras and 9 sequences shorter than 500 bp or of low quality. All chromatograms were visually inspected to minimize the sequencing errors.

Phylogenetic analysis

Sequences were aligned with MAFFT using the slow and iterative refinement method FFT-NS-i (Katoh *et al.*, 2002). The alignment was checked manually and edited using Seaview 3.2 (Galtier *et al.*, 1996), to retain the longest region that is common in most sequences. The final alignment had 961 positions and $\sim$815 bp per sequence (the average size was 797 bp, indicating that most positions in the alignment were covered). Maximum likelihood (ML) phylogenetic trees were constructed with RAxML (Stamatakis, 2006) using the evolutionary model GTR + G + I that best fits our data following the ModelTest (Posada and Crandall, 1998). Phylogenetic analyses were performed in the freely available University of Oslo Bioportal (www.bioportal.uio.no). Repeated runs on distinct starting trees were carried out to select the tree with the best topology (the one having the best Likelihood of 1000 alternative trees). Bootstrap ML analysis was carried out using 1000 pseudo-replicates. Trees were edited with the online tool iTOL (Letunic and Bork, 2007).

Grouping sequences in OTUs

Pair-wise Jukes–Cantor (JC) distances among all sequences were computed with PAUP (Swofford,

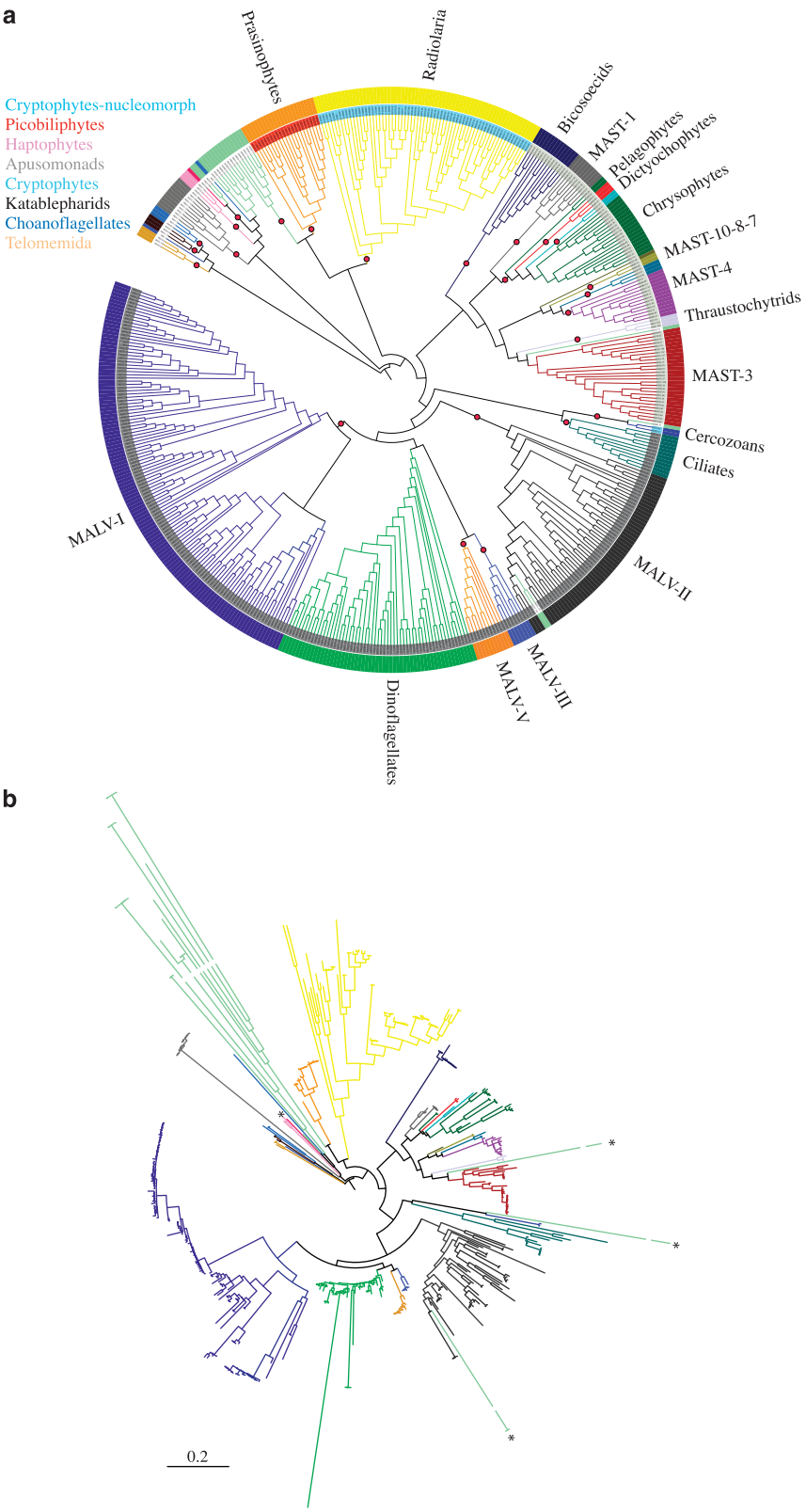

Figure 1 ML phylogenetic tree with 18S rDNA sequences of picoeukaryotes retrieved from the Indian Ocean. The tree was constructed using 500 sequences and ~815 bases in 961 positions. (**a**) Tree ignoring branch lengths and overlaid with colors with independent taxonomic assignations of sequences to an eukaryotic supergroup (inner ring) and to a given taxonomic group (colored branches and outer ring; names shown). Branches leading to groups with bootstrap values above 70% are marked with a red dot. (**b**) Same tree showing branch lengths, colored as before. The length of novel branches (light green) has been reduced to half, and those that have been phylogenetically placed are marked with an asterisk. The scale bar indicates 0.2 substitutions per position.

2002) using an alignment with unique sequences (398 sequences). The distance matrix was processed with DOTUR (Schloss and Handelsman, 2005) to group sequences in OTUs at different clustering distances. We used the rule of furthest neighbor and the highest precision ($P = 10000$). Heatmaps and Venn diagrams to compare samples were performed using the related application Mothur (http://www.mothur.org/). OTUs were also delineated using the online tool RAMI (Pommier *et al.*, 2009) that grouped sequences on the basis of their patristic distances (branch lengths). Rarefaction analyses were performed on both DOTUR and RAMI applications using the alignment with all 500 sequences.

Estimating the total number of OTUs

The total number of OTUs (defined at discrete clustering levels) was estimated by applying a set of statistical models to the observed OTU abundance. Parametric methods apply a model to the frequency distribution of OTUs and then project the distribution to estimate how many OTUs have been missed (Jeon *et al.*, 2006), whereas non-parametric methods, such as Chao1 or ACE, just apply a simple equation (Chao and Lee, 1992). Several parametric and non-parametric estimators (under different competing models and assumptions) were run at every possible right-truncation point of the frequency-count data, that is, omitting outliers (highly abundant taxa in the sample) with the beta version of the program CatchAll built at the Department of Statistical Science, Cornell University. The best parametric model was selected as the one providing the best compromise with a high goodness of fit, low standard error and maximal use of high frequency counts. The non-parametric method was chosen on the basis of the coefficient of variation of the estimate (Shen *et al.*, 2003).

Novelty analysis

Two values were recorded on the basis of a BLAST search of each environmental sequence against the nucleotide collection (nr/nt) database of NCBI (search on March 2010). The first value was the similarity with the closest environmental sequence in the BLAST output list (similarity CEM (closest environmental match)), excluding clones from the same library or study. The second was the similarity with the closest cultured organism (similarity CCM (closest cultured match)), which was the first entry in the list that was taxonomically classified. In a few cases, environmental sequences were so divergent that BLAST calculated the similarity using only a fragment, overestimating the similarity value. This occurred in 21 cases with the CEM and 38 cases with the CCM. In these instances, environmental and GenBank sequences were aligned with MAFFT, and the similarity was calculated using the uncorrected *p*-distance computed in PAUP. The novelty analysis reported the similarities of the environmental sequences against CEM and CCM in histograms or in dispersion plots (del Campo and Massana, submitted).

Results

Phylogenetic reconstruction of the diversity of marine picoeukaryotes

A ML phylogenetic tree with all 500 sequences provided a detailed picture of the diversity of Indian Ocean picoeukaryotes (Figure 1). In this tree, colored branches and external rings are based on the classification of sequences using BLAST and KeyDNATools. Both independent approaches, tree phylogeny and BLAST/KeyDNATools classification, were remarkably concordant (Figure 1a). The main supergroups (inner ring) were well represented and were divided into taxonomic groups roughly at the class level (outer ring), most of them with high-bootstrap values. Alveolates (dark gray in the inner ring) accounted for most clones in the data set, in particular dinoflagellates, MALV-I and MALV-II (47% of clones). Stramenopiles (light gray in the inner ring) followed in clonal abundance (19% of clones) and were dominated by several MAST lineages, chrysophytes and bicosoecids. Rhizaria (blue in the inner ring) were formed mostly by radiolarians (13% of clones). Two cercozoan sequences were closer to ciliates than to radiolarians, representing the only example for obvious incorrect phylogenetic placement. Archaeplastida (red in the inner ring) were formed, exclusively, by prasinophytes and accounted for 4% of the clones. The remaining groups (white in the inner ring) contained few badly resolved sequences, including typical marine groups, such as haptophytes, cryptophytes, katablepharids, picobiliphytes, telonemida and choanoflagellates. The same tree with real branch lengths (Figure 1b) gave a general impression of the unequal variability contained in each taxonomic group.

This highly supported tree was pivotal to place very divergent sequences that could not be identified by BLAST and KeyDNATools searches (21 sequences shown in light green branches). These novel sequences showed very long branches in the tree (Figure 1b) and interestingly some affiliated within a given taxonomic group (marked with an asterisk in Figure 1b). Thus, two divergent sequences were related to MALV-II, one to cercozoans, two to picobiliphytes and one to MAST (the three first cases supported by high bootstrap values). Nevertheless, 15 novel sequences could still not be related to any taxonomic group, not even to a supergroup, and occupy highly unique branches in this phylogenetic analysis.

Number of OTUs observed at varying clustering distances

Identical sequences were removed, which resulted in 398 unique sequences that represented the

number of OTUs at null distance. Unique sequences were then grouped into OTUs at distinct thresholds on the basis of JC pair-wise distances and patristic distances shown in the ML tree. The number of OTUs showed the largest decrease with the initial clustering relaxation (Figure 2a). Thus, the initial 398 OTUs were reduced to 237/242 (JC/Patristic grouping) at a distance of 0.01 (equivalent to 99% similarity; Figure 2a), meaning that 40% of the unique sequences collapse at this low distance (Figure 2b). We observed that this phenomenon occurred in all phylogenetic groups. After this significant initial decline, the number of OTUs continuously decreases with an increase in the clustering distance. JC and patristic distances grouped OTUs similarly up to a distance of 0.10, and above this value patristic distances delineated more OTUs (Figure 2a). This cannot be caused by the evolution model, as pair-wise JC and ML distances gave similar values (slope = 1.0253; $R^2 = 0.9993$; 500 sequences). Instead, these differences appear when the distances are calculated on

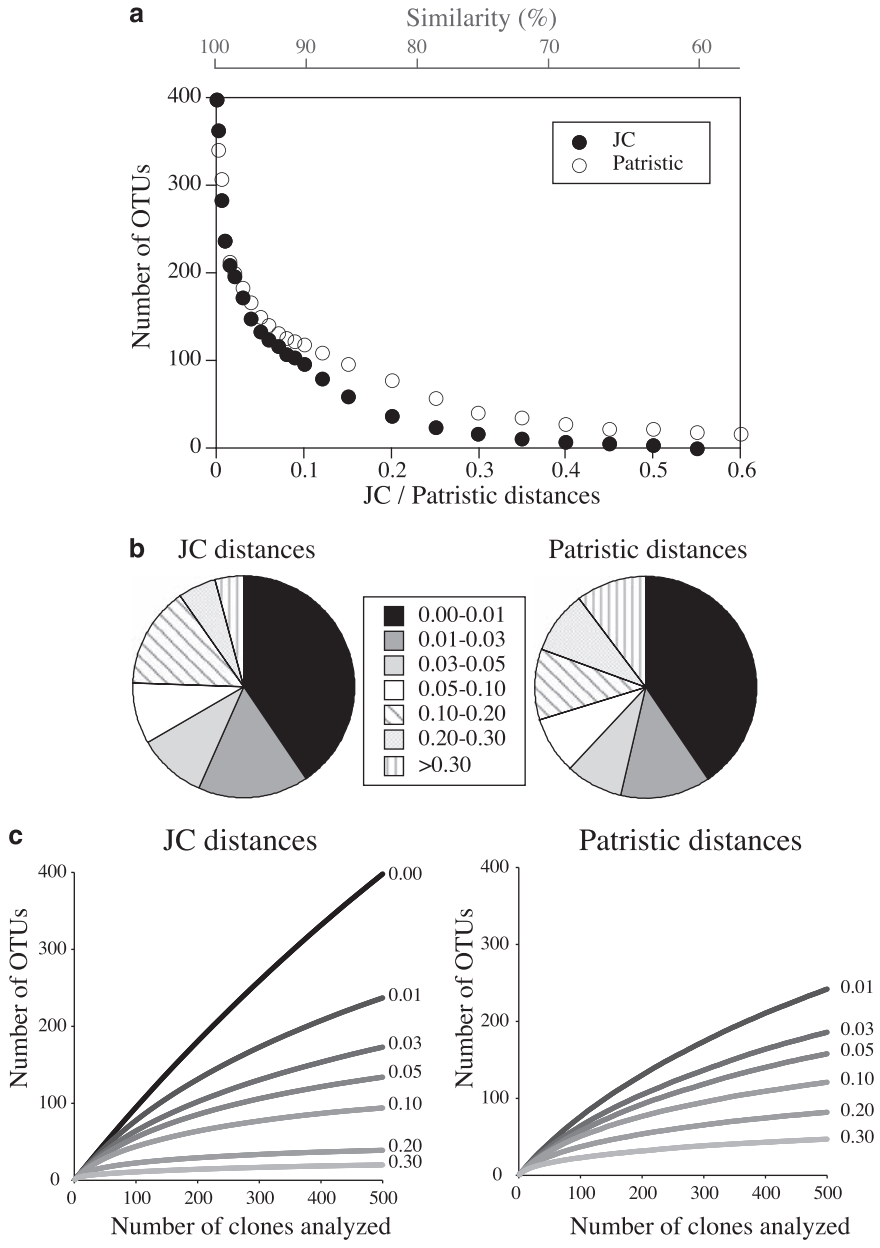

Figure 2 (**a**) Number of OTUs observed after grouping the 398 unique sequences from the Indian Ocean at different clustering levels on the basis of Jukes–Cantor or patristic distances. The correspondence between JC distance and sequence similarity is shown at the top of the graph for comparative purposes. (**b**) Distribution of the number of OTUs in distance classes for both grouping approaches. The area in each class represents the difference in OTUs observed at the two limits of the class (so the OTUs decrease when relaxing the clustering conditions between the two limits). (**c**) Rarefaction curves (OTUs observed versus clones analyzed) at discrete clustering distance levels (from 0.00 to 0.30) for both grouping approaches.

the basis of the phylogenetic tree. For instance, at a distance of 0.20 roughly separating taxonomic classes, JC distances delineate 39 OTUs, whereas patristic distances delineate 82. These differences are also evident in the distribution of OTUs in distance classes (Figure 2b).

Rarefaction curves were then constructed to relate the number of OTUs to the sequencing effort. The rarefaction curve with OTUs grouped at null JC distance did not show any sign of saturation (Figure 2c). Rarefaction curves constructed using OTUs that clustered at increasing distances showed a progressively better coverage, with plateaus starting to be evident at levels of 0.2 and 0.3 for both grouping methods. This indicated a severe undersampling to retrieve OTUs defined stringently, but at the same time suggested that the higher-rank phylogenetic groups were moderately well represented in the sequence data set.

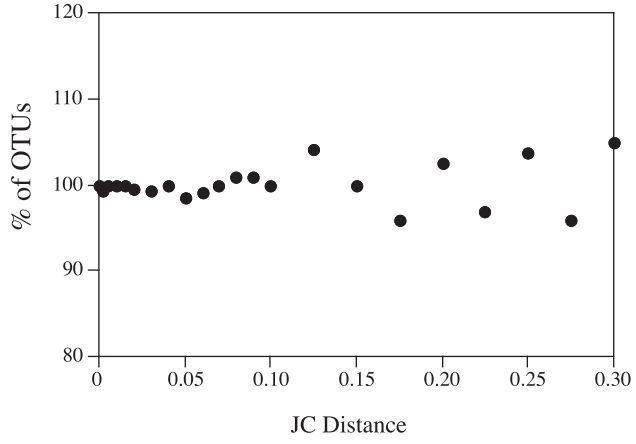

Figure 3 Percentage of total OTUs (estimated with the whole data set) that are recovered in 23 analyses with defined phylogenetic groups and adding up the counts for each separate group. This comparison was performed at 22 discrete clustering distance levels.

Our data set included a huge sequence variability (Figure 1b) and raised doubts about the accuracy of the alignment used for calculating pair-wise distances and the ML tree. In addition, hypervariable regions, kept to report the variability at all scales, were inevitably ambiguously aligned. Thus, we expected that doing separate analyses for coherent phylogenetic groups would yield better OTU counts. We prepared sequence data sets with the taxonomic groups shown in Figure 1a and redid the OTU counting (alignment, JC distances and DOTUR) for the 23 separate sets. And then, the number of OTUs in each set were added up and compared with the number observed with the whole data set. To our surprise, both approaches gave similar OTU numbers at clustering distance levels up to 0.30, being almost identical at all levels tested up to 0.10 (Figure 3). This exercise sustains the accuracy of the OTU counts and the ML tree obtained using the whole, and highly variable, data set.

Number of OTUs estimated at varying clustering distances

The rarefaction curves clearly showed that our data set underestimated diversity, particularly when OTUs were defined at low genetic distances. To estimate the 'total' number of OTUs, we applied several statistical methods on their frequency distribution (Table 1). Parametric models tend to predict higher estimates than non-parametric indices, and this was also observed in this study. The best parametric estimate obtained at null distance was 1951 (±193), and a distance of 0.01 was 731 (±150; JC grouping) or 803 (±188; patristic grouping). OTU estimates at increasing distances decrease in parallel with the decrease in the observed number, although observed and estimated values get closer at high distances. Thus, at 0.01 the observed OTUs represent 32–30% of the estimated

Table 1 Observed and estimated number of OTUs defined at discrete clustering levels (based on JC and patristic distances) within the 500 sequences of picoeukaryotes from the Indian Ocean

Distance	*JC-distance grouping*							*Patristic distance grouping*						
	Observed	*Parametric estimate*			*Non-parametric estimate*			*Observed*	*Parametric estimate*			*Non-parametric estimate*		
0.00	**398**	**1951**	*193*	SE	**1320**	*162*	AC							
0.01	**237**	**731**	*150*	ME	**609**	*91*	A1	**242**	**803**	*188*	ME	**700**	*117*	A1
0.02	**197**	**624**	*160*	ME	**552**	*96*	A1	**205**	**617**	*120*	ME	**646**	*122*	A1
0.03	**173**	**472**	*116*	ME	**396**	*64*	A1	**186**	**710**	*311*	ME	**685**	*155*	A1
0.04	**149**	**312**	*45*	ME	**243**	*25*	AC	**170**	**557**	*175*	ME	**440**	*79*	A1
0.05	**134**	**251**	*34*	ME	**203**	*19*	AC	**158**	**486**	*151*	ME	**394**	*73*	A1
0.07	**117**	**224**	*33*	ME	**176**	*18*	AC	**135**	**357**	*84*	ME	**306**	*56*	A1
0.10	**94**	**158**	*21*	ME	**129**	*12*	AC	**121**	**257**	*49*	ME	**223**	*35*	A1
0.12	**78**	**147**	*26*	ME	**132**	*24*	A1	**113**	**231**	*37*	ME	**177**	*20*	AC
0.15	**58**	**91**	*15*	ME	**77**	*9*	A1	**99**	**184**	*22*	ME	**151**	*18*	AC
0.20	**39**	**61**	*14*	ME	**61**	*15*	A1	**82**	**159**	*25*	ME	**151**	*29*	A1
0.30	**20**	**27**	*4*	SE	**35**	*14*	A1	**47**	**85**	*16*	ME	**93**	*26*	A1

The estimated number of OTUs was calculated under several parametric and non-parametric methods, showing the estimated value (bold), the standard error (italics) and the best-fitting model or index (SE: Single Exponential; ME: Two Mixed Exponential; AC: ACE; A1: ACE1).

value, whereas at 0.20 they represent 63–51% (Table 1). Hence, we are missing many more low-rank taxa than high-rank lineages.

Novelty analysis of marine picoeukaryotes

For each sequence, the similarity against the CEM and the CCM was recorded. The average CEM similarity (97.9%) was much higher than the average CCM similarity (91.9%). The similarity distribution against CEM was skewed towards the highest values, with a marked peak at 99%, whereas CCM similarity distributed well from 85 to 100%, with minor peaks at 87, 92 and 99% (Figure 4a). A dispersion plot of both similarity values showed that few sequences were closer to CCM than to CEM with most dots at the 1:1 line or below (Figure 4b). A notable exception was ten sequences close to *Amastigomonas debruynei*, and only 80% similar to a marine clone. Dots were shaded depending the neighbors they have, unveiling two dense areas (Figure 4b). The first was limited by CEM and CCM similarities above 98% (17% of sequences) and included sequences close to cultured organisms and marine clones. The second dense area was limited by CEM scores above 98% and CCM similarities between 87 and 93% (42% of sequences) and included sequences close to marine clones but distant to cultured organisms. The plot also highlighted novel sequences. Dots below 80% similarity in both axes indicated very divergent sequences never found before. Some sequences were only 75% similar to all sequences in GenBank except a few marine clones. Thus, three related clones were 98% similar to a single sequence from the Mediterranean Sea, whereas three other clones were 95–99% similar to sequences of a few Sargasso and Mediterranean Seas.

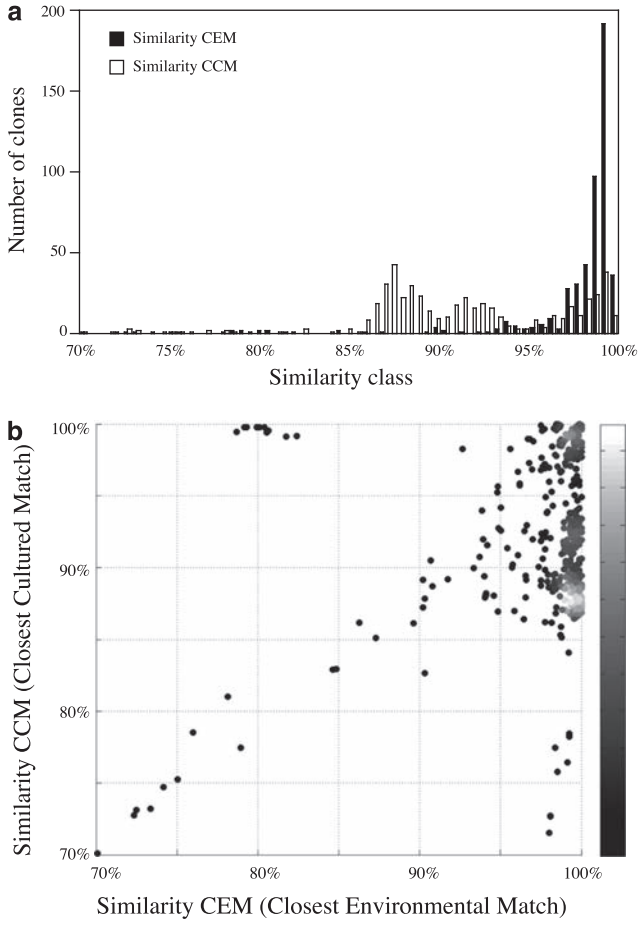

Figure 4 Novelty analysis of the 500 sequences of picoeukaryotes retrieved from the Indian Ocean. (**a**) Histogram showing the distribution of similarities against CEM and CCM of all sequences, in 0.5% similarity classes. (**b**) Dispersion plot of the CEM and CCM similarities for each sequence, with dots shaded depending on the number of neighbors (light gray dots indicate a dense area, whereas black dots indicate a disperse area).

Each particular phylogenetic group might exhibit a different novelty pattern, as exemplified with the supergroups alveolates and stramenopiles (Figure 5). In both cases most sequences were placed in the area with high CEM similarities (>98%), and had a particular behavior with respect to CCM. Thus, some stramenopile groups are at the top of the graph with high CCM scores (bicosoecids, dictyochophytes, pelagophytes), chrysophytes show an intermediate position with 90–95% CCM similarities, whereas MASTs and thraustochytrids have CCM similarities below 90% (Figure 5a). A similar distribution can be described for alveolates, with dinoflagellates at the top of the graph, followed by MALV-III and -V at an intermediate position and MALV-I and -II with lowest CCM scores (Figure 5b).

Comparing the diversity among samples

The protist composition in different samples was compared using their OTU content defined at 0.01 distance, roughly corresponding to species, and at 0.20 distance, roughly corresponding to a taxonomic level of class. Data were shown in heatmaps that quantify the pair-wise difference among samples, and Venn diagrams that show the number of unique and shared OTUs. At low distance, samples strongly differed among each other (Figure 6a), as expected due to the undersampling shown in rarefaction curves and statistical estimates. Still, some ecologically sound information was derived from the maps: the coastal sample was the most different, and the closest pairs were the two offshore surface samples (58 and 70) and the two offshore DCM samples (33 and 72). Venn diagrams were then constructed to compare coastal, surface and DCM samples. At low distance level, only a few OTUs were shared and unique OTUs were as high as 64% (coastal), 78% (surface) and 72% (DCM). As expected, OTUs grouped at a higher distance gave a different picture (Figure 6b). Heatmaps showed a homogenization of samples and Venn diagrams showed fewer unique OTUs (8% in coastal; 37% in surface; 33% in DCM), suggesting a rather coherent high-rank diversity among the samples analyzed.

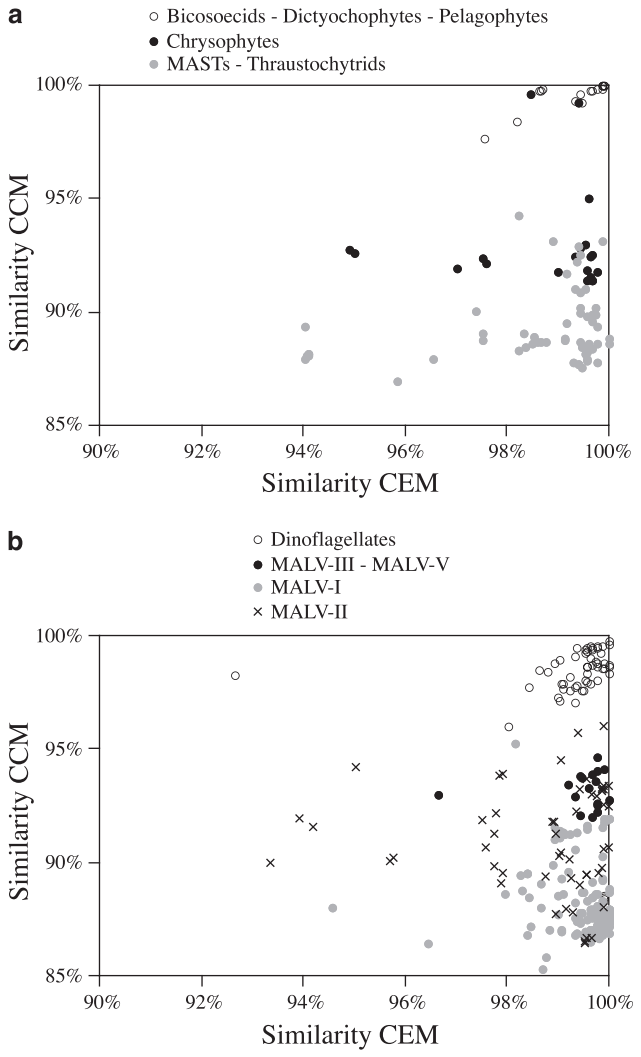

Figure 5 Dispersion plot of the CEM and CCM similarities for sequences affiliating to stramenopiles (**a**) and alveolates (**b**) separated in several taxonomic groups.

Discussion

Taxonomic groups detected

We used a data set of 500 18S rDNA sequences published before (Not *et al.*, 2008) to describe and quantify the diversity and novelty of picoeukaryotes from the Indian Ocean. 18S rDNA sequences were not complete (they were almost half of the gene, >800 bp), hence they contained insufficient positions for sound phylogenies. Moreover, it was not clear whether an alignment including highly divergent sequences could retrieve the proper relationships among them. The alignment was first used for a ML phylogenetic tree that recovered the main supergroups and most taxonomic groups (Figure 1). In fact, the tree-independent sequence classification (by BLAST and KeyDNATools) was concordant with the tree. Second, we compared the OTU number computed from the whole alignment or by adding the values of 23 separate alignments. Again, the results were satisfactory, as minor differences were found at all clustering levels tested (Figure 3). These exercises indicated that MAFFT could deal with highly variable sequence inputs and that our partial sequences were long enough for proper phylogenies, as was shown for bacterial 16S rDNA partial sequences (Stackebrandt and Rainey, 1995). These tests add consistency to the results presented in this study.

The ML tree showed a large diversity at different phylogenetic scales, pointing out that the seemingly homogeneous picoeukaryotic assemblages, seen by epifluorescence microscopy or flow cytometry, are formed by cells with highly divergent evolutionary histories. As in all studies based on size-fractionated biomass, it is possible that some of these sequences do not derive from picoeukaryotes, but are derived from larger cells broken during the filtration or detrital DNA, hence the picoeukaryote diversity that we present in this study might be overestimated. The high-rank diversity observed in this study, both in terms of eukaryotic supergroups detected and the presence and relative abundance of specific lineages, was typical of molecular surveys of marine picoeukaryotes (Massana and Pedrós-Alió, 2008; Vaulot *et al.*, 2008). Alveolates and stramenopiles were the most common groups. Rhizaria and archaeplastida appeared on a second level and were represented by a single lineage each. A unique choanoflagellate sequence (fungi were absent) represented the opisthokonts, whereas excavates and amoebozoa were not detected in this particular data set. Several reasons might explain the absence (or very low abundance) of some lineages in our libraries. First, some groups, such as many excavates, are likely unable to thrive on the marine plankton. Second, some cells could be excluded during the prefiltration step, such as larger loricated choanoflagellates (Leakey *et al.*, 2002) or particle-living amoebas (Rogerson *et al.*, 2003). Third, some lineages could be too scarce to be detected (Pedrós-Alió, 2007), a concern that can be partially solved by high-throughput sequencing, or by enriching the sample with the cells of interest using flow cytometry cell sorting (Shi *et al.*, 2009). Finally, some lineages could remain undetected because of inefficient DNA extraction or biased PCR amplification (Wintzingerode *et al.*, 1997). This will only be solved by modifying DNA extraction protocols and applying new universal and group-specific PCR primers.

Observed and estimated richness at different clustering levels

An important issue when quantifying the diversity of a natural assemblage is how to define the countable units. Ideally, units are biological species that work reasonably well for macroorganisms, but are impractical in the microbial world, particularly

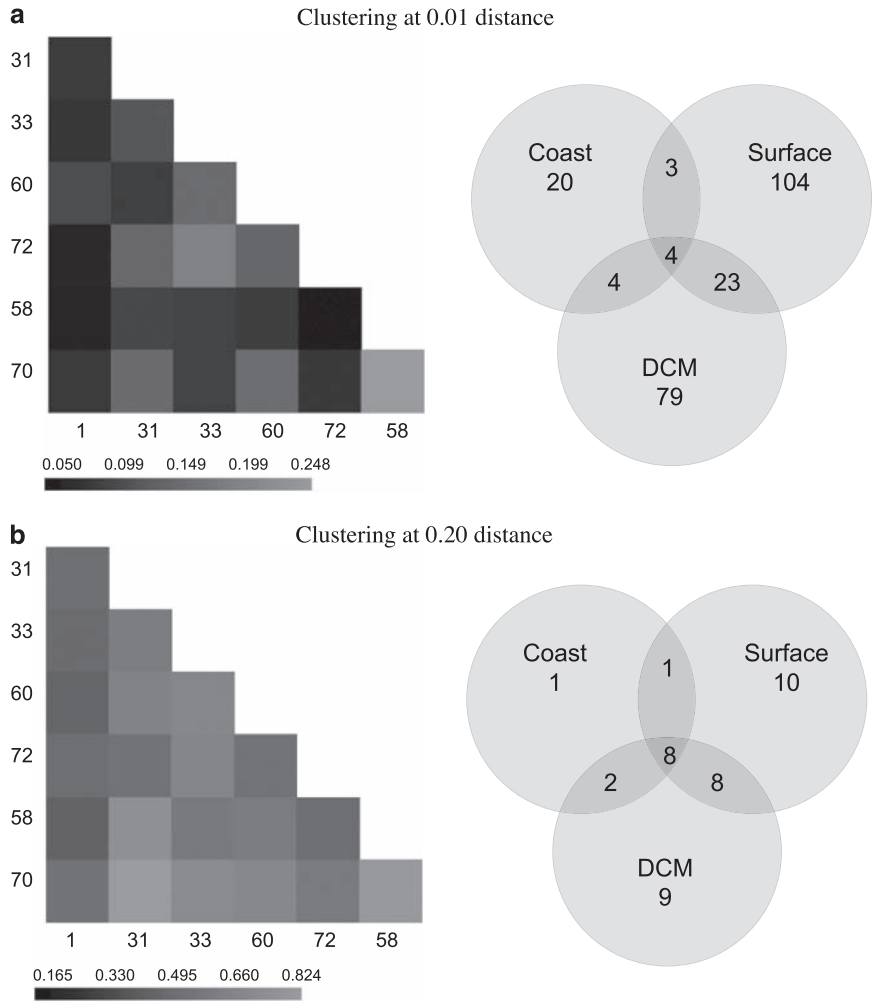

Figure 6 Heatmaps (left) and Venn diagrams (right) comparing the diversity of marine picoeukaryotes among samples, using the shared or unique OTUs defined at clustering JC distances of 0.01 (**a**) or 0.20 (**b**). Samples were derived from the coast (1), offshore surface (31, 58 and 70) or offshore DCM (33, 60 and 72).

within picoplankton, where diversity is determined using DNA sequence data. To create tractable units, sequences above a given distance threshold are pragmatically grouped into OTUs. When carried out at discrete clustering levels, this provides the number of OTUs at different phylogenetic scales and yields information on the genetic structure of microbial assemblages (Acinas *et al.*, 2004; Shaw *et al.*, 2008). This analysis has seldom been carried out with marine picoeukaryotes (Caron *et al.*, 2009). The most stringent criteria using null distance would be supported by laboratory studies, which show that only strains with identical rDNA gene sequences are sexually compatible (Amato *et al.*, 2007). In our data set, only 20% of the sequences are not contributing to a new OTU at null distance, highlighting the large diversity of the data set.

The largest decrease in OTU number occurred with the initial relaxation of the clustering conditions. This OTU collapse was caused by the presence of a substantial number of very similar ($\geqslant$99%) but seldom identical sequences. This microdiversity could be explained by a combination of methodological, biological and ecological factors. First, PCR or sequencing errors might account for part of these minute differences. In our data set, chromatograms were visually inspected to confirm the high quality of the reads and remove ambiguous positions, so few sequencing errors would be expected. Second, the rDNA gene in eukaryotes appears typically in tandem repeats varying from a few to several thousand copies depending on the taxa (Zhu *et al.*, 2005). Copies are generally homogenized by concerted evolution (Dover, 1982), but this process is not always complete and minor differences can be found within the same genome (Alverson and Kolnick, 2005). Third, in absence or low frequency of sexual reproduction, a plausible scenario for many protist species (Weisse, 2008), marine picoeukaryotes could experience similar evolutionary processes as bacteria and reveal equivalent microdiverse clusters (Acinas *et al.*,

2004). These clusters would be generated by neutral mutations (their genetic and functional diversity would be neutral), and could be regarded as natural taxonomic units or ecological species (Cohan, 2006).

The number of OTUs kept decreasing with the increasing distances. At distances up to 0.10, the grouping using JC or patristic distances showed a good correspondence, whereas above 0.10 both clustering methods deviate significantly, with patristic distances delineating more OTUs at a given clustering level. This is the expected and described trend (Pommier *et al.*, 2009), and occurs because patristic distances among two sequences, especially if they are divergent, are systematically larger than JC distances. OTUs produced by patristic distances are based on genetic change and would result in a more accurate and evolutionary robust clustering, but this is not yet a common practice in microbial ecology. The high number of OTUs detected at large distances results from the combination of a remarkable high-rank diversity (many taxonomic groups and supergroups) and the presence of very long branches at different positions of the tree (within well-defined groups or forming novel high-rank lineages).

Comparing observed and estimated OTU values allowed evaluation of the undersampling of our data. Parametric estimators, which are known to work better with low coverage data sets, such as ours (Epstein and López-García, 2008), predicted 1951 OTUs at null distance and 731/803 OTUs at 0.01 distance (JC/Patristic grouping, respectively). Thus, we only retrieved a glimpse of picoeukaryotic diversity (20% and 32–30%, respectively). By increasing the clustering distance level, the diversity coverage also increased (consistent with the rarefaction analysis), signifying that we started to miss less lineages. Our estimates ranked among the highest detected in surveys of microbial eukaryotes using clone libraries. At a similarity clustering level of 99% (distance of 0.01), our estimate was higher than the 398 OTUs from marine anoxic samples (Jeon *et al.*, 2006), 107 OTUs from hypersaline deep samples (Alexander *et al.*, 2009) or 605 OTUs from hydrothermal vent samples (Stoeck *et al.*, 2007). In surface marine samples, 572 OTUs were estimated at a clustering level of 95% (Countway *et al.*, 2007), a number slightly higher than ours. The unique high-throughput sequencing study with estimates from marine surface samples gave much higher values: 56292 OTUs at 100% similarity, 9231 at 99% and 3765 at 95% (Brown *et al.*, 2009). This study was based on early 454 technologies, which sequenced a very short amplicon (>50 bp) and could overestimate diversity due to low-frequency errors. Thus, although the actual numbers have to be regarded with caution, it seems clear that this study (and the many more to come with improved technologies) will significantly raise the higher limit of protistan diversity. Overall, marine picoeukaryotes appeared as highly diverse assemblages.

Novelty analysis of environmental sequences

Novelty of environmental sequences was inferred on the basis of their similarity with the GenBank database. At the time of the first eukaryotic molecular surveys, only the similarity against CCM could be calculated, yielding generally low values (Díez *et al.*, 2001). This situation changed after years of molecular surveys and thousands of deposited sequences. In present studies, marine environmental sequences have generally high CEM scores (with clones from other marine studies), whereas CCM scores still remain low. Hence, the large sequencing effort on marine picoeukaryotes during the last 10 years has not been paralleled by a significant culturing success, as revealed by the still uncultured MAST or MALV groups. Overall, our data highlight the huge culturing gap existing for the dominant marine picoeukaryotes.

The novelty analysis pointed out very divergent sequences that appeared in the area of the dispersion plot with very low CCM and/or CEM values. These sequences formed very long branches in the ML phylogenetic tree generally with an unresolved position, although some could be robustly placed in a taxonomic group based on the tree (see stars in Figure 1b). Nine sequences showed very low CCM and CEM scores, meaning that they are very distant to any existing sequence. We speculate that these unique sequences could be pseudogenes (Thornhill *et al.*, 2007), and this could be confirmed by secondary structure models. If they were pseudogenes, then they would not have any ecological implication and would not have stood as separate biological units. Six sequences were extremely divergent from any sequence except for a few marine clones. It is unlikely (but not impossible) that sequences retrieved thousands of kilometers apart are pseudogenes. Instead, these could represent high-rank novel phylogenetic lineages and are obvious candidates for further research. Retrieving additional sequences, constructing sound phylogenies and visualizing the target cells by Fluorescent in situ hybridization will identify if they are truly novel taxonomic units.

Concluding remarks

In this study, we explored the diversity and novelty patterns of marine picoeukaryotes using 18S rDNA clone libraries and Sanger sequencing. Our observations and the new exploratory approaches presented in this study can be adapted to facilitate the analysis of the massive amounts of data from the next generation sequencing technologies. It should be pointed out that although far from saturation, clone libraries can still provide longer sequences and of very high quality as compared with current next generation sequencing methods. In this study, we showed that picoeukaryotes from the Indian Ocean were highly diverse at distinct phylogenetic scales. In fact, we are only seeing the tip of the iceberg of

their diversity and it is expected that next generation sequencing will allow investigation of this underexplored space. Our data also indicated the presence of microdiverse clusters similar to those found in bacteria, but it is early to explain them by ecological factors or by biological or methodological factors. Most sequences from the Indian Ocean were highly similar to environmental sequences from other marine sites, indicating a widespread distribution of similar lineages, and many were far from cultured organisms, revealing a significant culturing gap. We also highlighted very divergent sequences, and we speculated that some could be pseudogenes and others could be novel high-rank phylogenetic lineages. From an ecological perspective, our quantitative sequence analysis would help to address fundamental questions of what generates, maintains and structures the large diversity observed, and what are the functional implications of this large diversity at different scales. From an evolutionary perspective, we are faced with very divergent sequences that could account for new, unexpected and fascinating evolutionary lineages.

Acknowledgements

This study was supported by the projects GEMMA (CTM2007–63753-C02–01/MAR, MEC), the NSF grant DEB-0816638 and the European Funding Agencies from the ERA-net program BiodivERsA, under the BioMarKs project. We thank Marco Álvarez for help in BLAST analysis, Miguel Lurgi and Baptiste Mourre for help in MatLab and Jose Castresana and Ramiro Logares for useful advices on phylogeny.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Alexander E, Stock A, Breiner H-W, Behnke A, Bunge J, Yakimov MM *et al.* (2009). Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. *Environ Microbiol* **11**: 360–381.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Alverson AJ, Kolnick L. (2005). Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *J Phycol* **41**: 1248–1257.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009). A method for studying protistan diversity using massively parrallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**: e6372.

Amato A, Kooistra WHCF, Ghiron JHL, Mann DG, Pröschold T, Montresor M. (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* **158**: 193–207.

Baldauf SL. (2003). The deep roots of eukaryotes. *Science* **300**: 1703–1706.

Brown MV, Philip GK, Bunge JA, Smith MC, Bisset A, Lauro FM *et al.* (2009). Microbial community structure in the North Pacific Ocean. *ISME J* **3**: 1374–1386.

Burki F, Schalchian-Tabrizi K, Pawlowski J. (2008). Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biology Lett* **4**: 366–369.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD *et al.* (2009). Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**: 5797–5808.

Chao A, Lee S-M. (1992). Estimating the number of classes via sample coverage. *J Amer Stat Assoc* **87**: 210–217.

Cohan FM. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Phil Trans R Soc B* **361**: 1985–1996.

Countway PD, Gast RJ, Dennet MR, Savai P, Rose JM, Caron DA. (2007). Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ Microbiol* **9**: 1219–1232.

Díez B, Pedrós-Alió C, Massana R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol* **67**: 2932–2941.

Dover GA. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299**: 111–117.

Epstein S, López-García P. (2008). 'Missing' protists: a molecular prospective. *Biodivers Conserv* **17**: 261–276.

Galtier N, Gouy M, Gautier C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**: 543–548.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60–63.

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al.* (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ Microbiol* **10**: 397–408.

Hutchinson GE. (1961). The paradox of the plankton. *Am Nat* **95**: 137–145.

Jeon S-O, Bunge J, Stoeck T, Barger KJA, Hong S-H, Epstein SS. (2006). Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72**: 6578–6583.

Johnson PW, Sieburth JMcN. (1982). *In-situ* morphology and occurrence of eucaryotic phototrophs of bacterial size in the picoplankton of estuarine and oceanic waters. *J Phycol* **18**: 318–327.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

Leakey RJG, Leadbeater BSC, Mitchell E, McCready SMM, Murray AWA. (2002). The abundance and biomass of choanoflagellates and other nanoflagellates in waters of contrasting temperature to the north-west of South Georgia in the Southern Ocean. *Eur J Protistol* **38**: 333–350.

Lefranc M, Thénot A, Lepère C, Debroas D. (2005). Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microbiol* **71**: 5935–5942.

Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127–128.

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.

Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A *et al.* (2004). Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol* **70**: 3528–3534.

Massana R, Pedrós-Alió C. (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213–218.

Moon-van der Staay SY, De Wachter R, Vaulot D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**: 607–610.

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Töbe K *et al.* (2007). Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315**: 252–254.

Not F, Latasa M, Scharek R, Viprey M, Karleskind P, Balagué V *et al.* (2008). Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep Sea Res I* **55**: 1456–1473.

Olson RJ, Vaulot D, Chisholm SW. (1985). Marine-phytoplankton distributions measured using shipboard flow-cytometry. *Deep Sea Res* **32**: 1273–1280.

Pedrós-Alió C. (2007). Dipping into the rare biosphere. *Science* **315**: 192–193.

Pommier T, Canbäck B, Lundberg P, Hagström Å, Tunlind A. (2009). RAMI, a tool for identification and characterization of phylogenetic clusters in microbial communitites. *Bioinformatics* **25**: 736–742.

Posada D, Crandall KA. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA. (2005). The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* **7**: 1413–1425.

Rogerson A, Anderson OR, Vogel C. (2003). Are planktonic naked amoebae predominantly floc associated or free in the water column? *J Plankton Res* **25**: 1359–1365.

Scheffer M, Rinaldi S, Huisman J, Weissing FJ. (2003). Why plankton communities have no equilibrium: solutions to the paradox. *Hydrobiologia* **491**: 9–18.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining Operational Taxonomic Units and estimating especies richness. *Appl Environ Microbiol* **71**: 1501–1506.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JBH. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Shen TJ, Chao A, Lin CF. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**: 798–804.

Sherr BF, Sherr EB, Caron DA, Vaulot D, Worden AZ. (2007). Oceanic protists. *Oceanography* **20**: 130–134.

Shi XL, Marie D, Jardillier L, Scanlan DJ, Vaulot D. (2009). Groups without cultured representatives dominate eukaryotic picophytoplankton in the oligotrophic South East Pacific Ocean. *PLoS ONE* **4**: e7657.

Simpson AGB, Roger AJ. (2004). The real 'kingdoms' of eukaryotes. *Curr Biol* **14**: R693–R696.

Stackebrandt E, Rainey FA. (1995). Partial and complete 16S rDNA sequences, their use in generation of 16S rDNA phylogenetic trees and their implications in molecular ecological studies. In: Akkermans ADL, van Elsas JD, de Bruijn FJ (eds). *Molecular Microbial Ecology Manual*. Kluwer Academic: Dordrecht. pp 1–17.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stoeck T, Kasper J, Bunge J, Leslin C, Ilyin V, Epstein S. (2007). Protistan diversity in the Arctic: a case of paleoclimate shaping modern biodiversity? *PLoS ONE* **2**: e278.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A *et al.* (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology* **7**: 72.

Swofford DL. (2002). PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates: Sunderland, Mass.

Thornhill DJ, Lajeunesse TC, Santos SR. (2007). Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Mol Ecol* **16**: 5326–5340.

Vaulot D, Eikrem W, Viprey M, Moreau H. (2008). The diversity of small eukaryotic phytoplankton ($\leqslant 3$ μm) in marine ecosystems. *FEMS Microbiol Rev* **32**: 795–820.

Weisse T. (2008). Distribution and diversity of aquatic protists: and evolutionary and ecological perspective. *Biodivers Conserv* **17**: 243–259.

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.

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.