
Marine Biology Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/smar20

# Evaluating the potential of 18S rDNA clone libraries to complement pyrosequencing data of marine protists with near full-length sequence information

Christian Wolf[a], Estelle Silvia Kilias[a] & Katja Metfies[a]
[a] Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
Published online: 22 Apr 2014.



To cite this article: Christian Wolf, Estelle Silvia Kilias & Katja Metfies (2014) Evaluating the potential of 18S rDNA clone libraries to complement pyrosequencing data of marine protists with near full-length sequence information, Marine Biology Research, 10:8, 771-780, DOI: 10.1080/17451000.2013.852685

To link to this article: http://dx.doi.org/10.1080/17451000.2013.852685

*Marine Biology Research*, 2014
Vol. 10, No. 8, 771–780, http://dx.doi.org/10.1080/17451000.2013.852685

**ORIGINAL ARTICLE**

# Evaluating the potential of 18S rDNA clone libraries to complement pyrosequencing data of marine protists with near full-length sequence information

CHRISTIAN WOLF*, ESTELLE SILVIA KILIAS & KATJA METFIES

*Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*

**Abstract**
Sequencing of 18S rDNA clone libraries and 454-pyrosequencing are valuable methods used to describe microbial diversity. The massively parallel 454-pyrosequencing generates vast amounts of ribosomal sequence data and has the potential to uncover more organisms, even rare species. However, the relatively short sequence lengths of ~500 bp are suboptimal for taxonomic annotation and phylogenetic analyses. In this study, we assessed the potential of 18S ribosomal clone libraries to complement corresponding 454-pyrosequencing data with near full-length sequence information. This involved a comparison of protist community compositions in five polar samples suggested by 18S rDNA clone libraries, with the corresponding community compositions suggested by 454-pyrosequencing. The study was conducted with four Arctic water samples, focusing on the eukaryotic picoplankton (0.4–3 μm), and with one sample collected in the Southern Ocean, examining the entire size spectrum (> 0.4 μm). For all individual samples, the protist community compositions suggested by the two different approaches showed significant similarities. Around 70% of the sequences detected by sequencing of clone libraries were also present in the 454-pyrosequencing data set. However, the clone library sequences reflected only ~20% of the abundant biosphere identified by 454-pyrosequencing and identified ribosomal sequences that were not detected in the 454-pyrosequencing data sets.

**Key words:** *18S rDNA near full-length clones, genetic diversity, polar regions*

## Introduction

Recently, a number of publications have shown that 454-pyrosequencing of ribosomal genes is an efficient tool for assessment of microbial communities (e.g. Sogin et al. 2006; Cheung et al. 2010; Stoeck et al. 2010; Comeau et al. 2013; Wolf et al. 2013). It is independent of the cloning step and allows high-resolution sequencing of microbial sequences (Margulies et al. 2005). In comparison to analysis of clone libraries, massive parallel pyrosequencing provides more sequences and uncovers more organisms with fewer costs (Huse et al. 2008). In respect of the vast microbial diversity, the greater sampling depth is advantageous and even allows elucidating the diversity of the rare biosphere (Sogin et al. 2006; Galand et al. 2009). However, one caveat of the pyrosequencing approach is the tendency to overestimate the number of rare phylotypes because of sequencing errors (Quince et al. 2009). Such errors will run the risk of inflating the diversity estimates, due to the fact that every single read is considered to represent a community member (Kunin et al. 2010). An additional caveat is the short sequence length of ~500 bp, which limits the reliability of phylogenetic analyses based on the 454-sequences. Thus, with respect to the length of the sequences and phylogenetic analyses, the analysis of ribosomal clone libraries (Diez et al. 2001; Lovejoy et al. 2006) is advantageous over the 454-pyrosequencing approach, because it allows sequencing of the whole 18S rDNA, which provides a better basis for phylogenetic analyses.

During the past two decades, numerous phylogenetic investigations of the eukaryotic protist diversity

*Corresponding author: Christian Wolf, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany. E-mail: Christian.Wolf@awi.de
All authors contributed equally to this work.

*Published in collaboration with the Institute of Marine Research, Norway*

*(Accepted 30 September 2013; Published online 22 April 2014; Printed 29 April 2014)*

based on the analysis of 18S rDNA clone libraries contributed significantly to elucidate eukaryotic phytoplankton diversity and community composition in the marine environment (Diez et al. 2001; Lovejoy et al. 2006). The characterization of eukaryotic microbial communities via sequencing of 18S rDNA clone libraries usually relies on limited sets of a maximum of 100–200 clones per sample. Thus, the method is not suited to provide a comprehensive view of the diversity in a sample (Diez et al. 2001; Lopez-Garcia et al. 2001; Moon-van der Staay et al. 2001). Nevertheless, sequencing of the 18S rDNA is a reliable approach for phylogenetic analyses of new environmental sequences and served during the past decades as a gold standard in molecular assessments of phytoplankton diversity (Diez et al. 2001; Lovejoy et al. 2006; Cheung et al. 2010; Lovejoy & Potvin 2011).

Concerning the advantage of clone libraries over 454-pyrosequencing in terms of phylogenetic analyses of new environmental sequences, in this study we addressed the potential of ribosomal clone libraries to complement ribosomal 454-pyrosequencing data with near full-length sequence information, at least for the abundant biosphere. This involves three major questions. (1) How does the community structure revealed by clone libraries compare to the community structure revealed by 454-pyrosequencing? (2) Do clone library data exclusively reflect the abundant biosphere? (3) Does the additional phylogenetic analysis of near full-length ribosomal genes improve the annotation of 454-sequences? To answer these questions, we analysed four samples from the Arctic Ocean, comprising the picoeukaryotic fraction (0.4–3 μm), and one sample from the Southern Ocean, comprising the whole size fraction (> 0.4 μm). We chose this sampling setup to exclude a possible bias induced by cell size or geographical background. Furthermore, we used different primer sets for the amplification of the ribosomal sequences in order to include the primer bias in the comparison.

## Materials and methods

### *Location and sampling*

The study area comprises four stations located in the Fram Strait (Arctic Ocean), as well as one station from the Southern Ocean (Figure 1). The coordinates of the four Arctic stations were: 6.1°E and 79.1°N (HG1), 4.2°E and 79.1°N (HG4), 4.5°E and 79.7°N (HGN4), and 5.1°E and 78.6°N (HGS3), sampled during the ARK XXIV/2 cruise onboard the *RV Polarstern* in July 2009. The

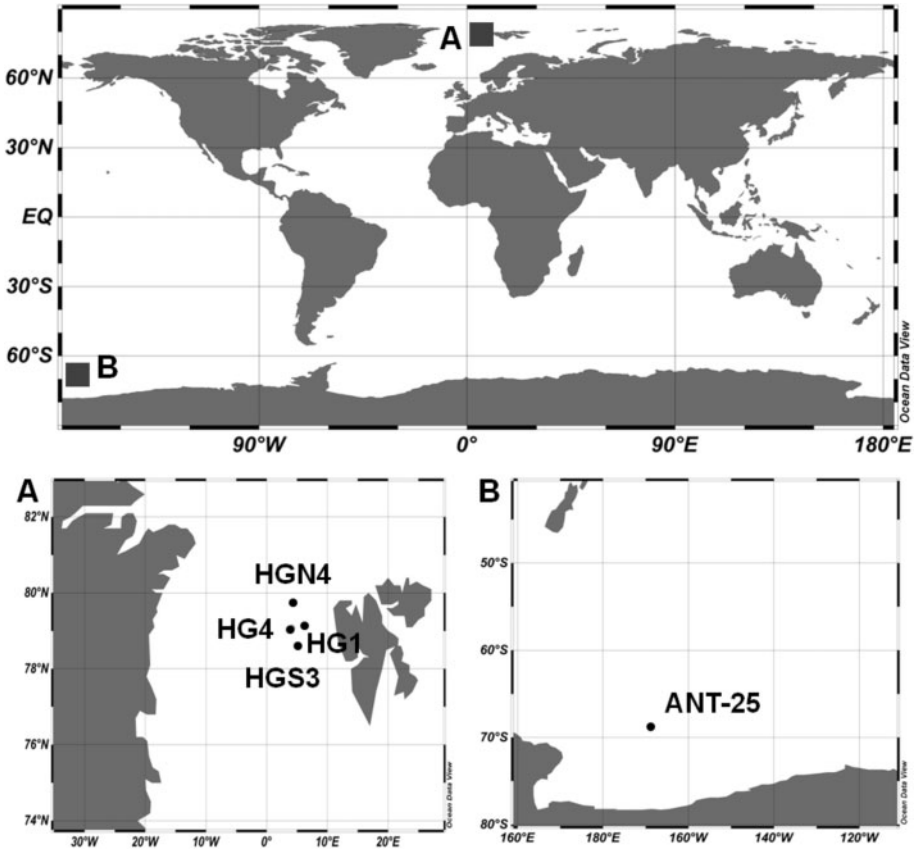


Figure 1. Map of the sampling stations located within (A) the long-term observatory 'Hausgarten' (Fram Strait, Arctic) and (B) the Southern Ocean.

sampling site in the Southern Ocean was located at 164.9°W longitude and 69°S latitude. Sampling took place during the *RV Polarstern* cruise ANT XXVI/3 in February 2010. The Arctic samples were collected from the subsurface maximum chlorophyll layer with Niskin bottles installed on a rosette system, equipped with depth, temperature, salinity, and fluorescence profilers. The Antarctic sample was collected using the ship pumping system (membrane pump), located at the bow at 8 m depth below the surface. The sampling depths differed because they were part of studies with different research aims. Because our study focused on a methodological comparison instead of ecological conclusions, the different sampling depths are not relevant. In both cases, 1.5 l of seawater were successively filtered at a pressure of 200 mbar onto Isopore Membrane Filters (Millipore, USA) with a pore size of 10, 3 and finally 0.4 µm. The filters were stored at −80°C until further treatment in the laboratory.

### *DNA extraction*

Extraction of genomic DNA from all samples was carried out with the E.Z.N.A. TM SP Plant DNA Kit (Omega Bio-Tek, USA) following the manufacturer's instructions. DNA concentration was determined with a NanoDrop 1000 system (Thermo Fisher Scientific, USA).

### *Clone library construction*

The 18S rDNA of the Arctic samples was amplified using the specific primers 82F (5′-GTA AAA CTG CGA ATG GCT CAT-3′) (Lopez-Garcia et al. 2001) and 1528R (5′-TGA TCC TTC TGC AGG TTC ACC TAC-3′) (modified after Elwood et al. 1985) and genomic DNA from the 0.4–3 µm fraction as template. The amplification of the Southern Ocean sample was conducted using the primer combination 300F (5′-AGG GTT CGA TTC CGG AG-3′) and 1200R (5′-CAG GTC TGT GAT GCC C-3′), because the former combination resulted in a poor PCR product. Both primers bind on eukaryotes, including all major taxonomic groups (tested with the SILVA database SSU Ref 108). Furthermore, the whole protist assemblage (> 0.4 µm) was used for the methodological comparison of the Southern Ocean sample. In this respect, the 18S rDNA of each fraction was amplified and equal volumes of each PCR product were pooled before the purification. The PCR reaction mixture contained 1 × HotMaster Taq Buffer containing 2.5 mM $Mg^{2+}$ (5 Prime, USA), 0.4 U of HotMaster Taq polymerase (5 Prime, USA), 10 mg/ml BSA, 10 mM dNTP-mix (Eppendorf, Germany), 10 µM of each Primer and 1 µl of template DNA in a final volume of 20 µl. PCR reactions were carried out in a Mastercycler (Eppendorf, Germany) under the following conditions: an initial denaturation at 94°C for 3 min, 35 cycles of denaturation at 94°C for 45 s, annealing at 55°C for 1 min and extension at 72°C for 3 min, and a final extension at 72°C for 10 min. The purification of the resulting PCR fragment was carried out with the Gel Purification Kit (Invitrogen, USA), following the manufacturer's protocol. Subsequently, the fragment was cloned into the pDrive Cloning Vector (QIAGEN, Germany) taking advantage of the PCR Cloning Kit (QIAGEN, Germany) and transformed into TOP10 chemo-competent *Escherichia coli* cells (Invitrogen, USA). Clones were sequenced from one direction using the 300F (see above) and 528F (5′-GCG GTA ATT CCA GCT CCA A-3′) (modified after Elwood et al. 1985) primer under the following conditions: an initial denaturation step at 96°C for 1 min, 25 cycles of denaturation at 96°C for 10 s, annealing at 50°C for 5 s and extension at 60°C for 4 min. The terminal sequencing was carried out on an ABI Prism 310 Genetic Analyzer (Applied Biosystems, USA).

### *454-pyrosequencing*

The hypervariable V4 region of the 18S rDNA was amplified taking advantage of the primer combination 528F and 1055R (5′-ACG GCC ATG CAC CAC CAC CCA T-3′) (modified after Elwood et al. 1985). The PCR mixtures were composed as described previously for the clone library construction. Reaction conditions were as following: an initial denaturation at 94°C for 3 min, 35 cycles of denaturation at 94°C for 45 s, annealing at 59°C for 1 min and extension at 72°C for 3 min, and a final extension at 72°C for 10 min. Subsequently, the amplicons were purified with the Mini Elute PCR Purification Kit (QIAGEN, Germany). In case of the Southern Ocean sample, equal volumes of PCR reaction of each size fraction were pooled and purified with the MinElute PCR purification kit (Qiagen, Germany) following the manufacturer's instructions. Pyrosequencing was performed on a Genome Sequencer FLX system (Roche, Germany) by GATC Biotech AG (Germany).

### *Data analysis*

The two sequences of each sequenced clone were assembled with the software Lasergene 10 (DNASTAR, USA) and a consensus sequence was built. All sequences (clone consensus sequences and 454-pyrosequencing reads) were checked for errors (reads with many unresolved bases) implied by the

sequencing process and sequences with more than one uncertain base (N) were removed. The remaining sequences were checked for possible chimera formation by applying the detecting software UCHIME 4.2.40 (Edgar et al. 2011; same reference database used as for pplacer, see below) and all sequences considered as being chimeric were excluded from further analysis. The remaining sequences were analysed using the Lasergene 10 software (DNASTAR, USA). They were clustered into operational taxonomic units (OTUs) at the 97% similarity level. The 97% similarity level has shown to be the most suitable to reproduce original eukaryotic diversity (Behnke et al. 2011) and has the effect of bracing most of the sequencing errors (Kunin et al. 2010). Furthermore, known intragenomic SSU polymorphism levels can range to 2.9% in dinoflagellate species (Miranda et al. 2012). OTUs based on only one sequence (singletons) were removed to avoid the analysis of artificial sequences and overestimation of the diversity (Behnke et al. 2011). Consensus sequences of the OTUs were aligned using the software HMMER 2.3.2 (Eddy 2011). Subsequently, taxonomical affiliation was determined by placing the consensus sequences into a reference tree, consisting of 1200 high-quality 18S rDNA sequences of Eukarya from the SILVA reference database (SSU Ref 108), using the software pplacer 1.0 (Matsen et al. 2010). The compiled reference database is available on request in ARB-format. Non-protist sequences originating from metazoans and fungi were removed.

A phylogenetic tree based on the 18S rDNA sequences (clone sequence and 454-pyrosequencing sequence) of one OTU was calculated using maximum likelihood under the implementation of the Jukes–Cantor model and 1000 bootstrap replications. Reference sequences were obtained from GenBank (NCBI).

The clone library sequences generated in this study have been deposited at GenBank under Accession No. JX840877–JX840942. The 454-pyrosequencing reads were part of other studies and were deposited at GenBank's Short Read Archive (SRA) under Accession No. SRA058841 (Arctic samples) and SRA056811 (Southern Ocean sample).

## Results

The five clone libraries analysed in this study resulted in 698 high-quality clones, while the number of clones per sample varied between 101 and 179 (Table I). Non-target sequences (metazoan and fungi sequences) were removed from the data sets. However, only the pooled ANT25 clone library (6%) contained non-target sequences. In contrast, chimeras were found in most clone libraries (6–19%), except in library HG4 (0%). Final clustering of the remaining sequences resulted in seven (HG4) to 24 (HGS3) different OTUs (Table I). The amount of sequence reads generated with 454-pyrosequencing was about three orders of magnitude higher than the number of sequence reads generated with the clone libraries. The high-throughput sequencing approach resulted in 79,865 raw reads. The number of raw reads per sample varied between 7539 (HG4) and 45,772 (ANT25). The quality filtering reduced the initial read number to a final range of 5220 (HGN4) to 30,561 (ANT25) reads. Based on a clustering at the 97% similarity level the raw reads clustered in 709 (HG4) to 1153 (ANT25) different OTUs. The subsequent analytical process revealed 2–6% of chimeric sequences in the clustered pyrosequencing data set. Both approaches suggested that the sample taken at HG4 had the least diversity, while the samples ANT25 and HGS3 contained the highest diversity.

Table I. Summary of recovered clones and 454-pyrosequencing reads.

| | Sampling site | | | | |
|---|---|---|---|---|---|
| | HG1 | HG4 | HGN4 | HGS3 | ANT25 |
| *Clone library* | | | | | |
| High-quality clones | 175 | 179 | 101 | 139 | 104 |
| OTUs (97%) | 16 | 7 | 13 | 24 | 19 |
| *454-pyrosequencing* | | | | | |
| Total reads | 9830 | 7539 | 7938 | 8786 | 45,772 |
| High-quality reads | 8154 | 5434 | 5220 | 7020 | 30,561 |
| OTUs (97%) | 754 | 709 | 829 | 1014 | 1153 |

### *Comparison of clone library and 454-pyrosequencing data set – Arctic*

In total, 47 different OTUs have been identified in the clone libraries generated from the Arctic samples (Table II). The number of OTUs obtained from a single sample ranged from 7 (HG4) to 24 (HGS3). The clone library analyses suggest a dominance of dinoflagellates in all Arctic samples in the data set. In the individual samples, most sequence reads were affiliated with dinoflagellates (3.5–64.3%), followed by chlorophytes (4.3–82.9%), stramenopiles (1–2.2%), cryptophytes (0.6–28.1%) and ciliates (0.7–9.7%). Neither haptophytes nor rhodophytes were detected in the clone libraries. A similar community structure is suggested by 454-pyrosequencing. As observed previously in the clone library data set, most reads generated from the individual samples were affiliated with dinoflagellates (18.2–51.4%), followed by chlorophytes (3.4–42.2%),

Table II. Phylogenetic affiliations of the Arctic clone OTUs and their relative abundance in the libraries and the 454-pyrosequencing data set at the four sampling sites (HG1, HG4, HGN4, HGS3); x, absent.

| OTU | Closest match (Maximum identity %) | Taxonomic group | Clones (%)/454 (%) | | | |
|---|---|---|---|---|---|---|
| | | | HG1 | HG4 | HGN4 | HGS3 |
| ARK_1 | *Bolidomonas pacifica* (92) | Stramenopiles | 1.1/2.8 | x/0.7 | 1.0/0.2 | x/0.5 |
| ARK_2 | Clone EU793918.1 (99) | Syndiniales | 0.6/0.8 | x/0.5 | x/0.5 | x/2.6 |
| ARK_3 | Clone HM135092.1 (98) | Dinophytes | 0.6/0.4 | x/0.4 | 4.0/0.9 | 7.2/0.2 |
| ARK_4 | Clone JF791003.1 (98) | Syndiniales | x | x | 5.0/x | x |
| ARK_5 | Clone GU819790.1 (98) | Syndiniales | x/0.24 | 1.1/0.2 | 3.0/0.5 | 1.4/1.7 |
| ARK_6 | *Micromonas pusilla* (99) | Chlorophytes | 2.9/1.4 | x/1.9 | x/0.3 | x/1.3 |
| ARK_11 | Clone HQ438132.1 (94) | Syndiniales | x | x/0.2 | 8.9/0.7 | x/0.1 |
| ARK_12 | Syndiniales EU793925.1 (95) | Syndiniales | x/0.2 | x/0.1 | x/0.8 | 26.6/0.6 |
| ARK_13 | *Gyrodinium* AB120001.1 | Dinophytes | x | x | 4.0/x | 5.0/x |
| ARK_14 | *Geminigera cryophila* (99) | Cryptophytes | 0.6/0.4 | x/0.1 | x | 23.7/0.1 |
| ARK_15 | *Micromonas pusilla* (99) | Chlorophytes | 77.1/14.4 | 65.9/1.1 | 47.5/0.4 | x/1.1 |
| ARK_16 | Clone AY295399.1 (91) | Ciliates | 8.0/1.2 | x/0.8 | x/0.4 | 0.7/0.2 |
| ARK_17 | Clone EU682572.1 (97) | Ciliates | 1.7/0.3 | x/0.3 | x/0.2 | x/0.1 |
| ARK_20 | Clone HQ43812.9 (98) | Dinophytes | 1.1/x | x/<0.1 | x | x/0.1 |
| ARK_21 | Clone JN934892.1 (95) | Picobiliphytes | 1.1/<0.1 | x/<0.1 | x/<0.1 | x/<0.1 |
| ARK_25 | *Gyrodinium* sp. (98) | Dinophytes | 0.6/0.4 | x | x/0.1 | x |
| ARK_26 | *Woloszynskia* sp. (99) | Dinophytes | 0.6/0.4 | x/<0.1 | x/0.1 | x/0.2 |
| ARK_29 | *Micromonas pusilla* (99) | Chlorophytes | 0.6/3.0 | 0.6/1.1 | x/0.1 | x/0.6 |
| ARK_30 | Clone HQ222463.1 (98) | Picobiliphytes | 0.6/x | x/<0.1 | x | x/<0.1 |
| ARK_31 | *Micromonas pusilla* (99) | Chlorophytes | 2.3/x | x | x | x |
| ARK_33 | Clone AJ420693.1 (96) | Rhodophytes | 0.6/x | x | x | x |
| ARK_37 | Clone AF290067.2 (98) | Syndiniales | x | 0.6/x | x | x |
| ARK_38 | Clone EU682636.1 (97) | Chlorophytes | x | 1.7/x | x | x |
| ARK_46 | *Micromonas pusilla* (91) | Chlorophytes | x | 0.6/x | x | x |
| ARK_47 | Syndiniales EU793375.1 (90) | Syndiniales | x | 29.6/x | x | x |
| ARK_58 | Clone EU793946.1 (88) | Syndiniales | x | x/0.1 | 10.9/x | 0.7/x |
| ARK_60 | Clone EU793957.12 (92) | Syndiniales | x | x | 2.0/x | x |
| ARK_62 | Clone EU682577.1 (98) | Dinophytes | x | x | 6.9/x | x |
| ARK_68 | Clone EF172940.1 (98) | Syndiniales | x | x/0.1 | 4.0/0.1 | x/<0.1 |
| ARK_69 | Clone JF826365.1 (91) | Syndiniales | x/0.1 | x/0.3 | 2.0/0.1 | x/0.1 |
| ARK_70 | Clone HQ438143.1 (95) | Syndiniales | x/<0.1 | x/0.1 | 1.0/0.1 | 1.4/0.1 |
| ARK_72 | Clone EU793201.1 (98) | Syndiniales | x | x | x | 0.7/x |
| ARK_76 | Clone EU793383.1 (90) | Syndiniales | x/0.1 | x/0.1 | x | 2.2/0.2 |
| ARK_78 | Clone EF195735.1 (90) | Cryptophytes | x/0.3 | x/0.3 | x/0.2 | 0.7/0.1 |
| ARK_82 | Clone EU793221.1 (94) | Syndiniales | x | x | x | 1.4/<0.1 |
| ARK_83 | Clone EU793700.1 (94) | Dinophytes | x/0.2 | x/1.0 | x/1.1 | 2.2/1.3 |
| ARK_86 | Clone EU793708.1 (92) | Syndiniales | x/0.1 | x/<0.1 | x/<0.1 | 0.7/0.1 |
| ARK_87 | Clone HM561117.1 (95) | Dinophytes | x | x | x/<0.1 | 5.0/x |
| ARK_90 | Clone HQ222399.1 (95) | Syndiniales | x | x | x | 1.4/0.1 |
| ARK_91 | Clone FJ537539.1 (92) | Syndiniales | x | x | x | 2.9/x |
| ARK_92 | *Bolidomonas pacifica* (95) | Stramenopiles | x/0.1 | x | x/0.1 | 2.2/0.1 |
| ARK_93 | *Bathycoccus prasinos* (98) | Chlorophytes | x/8.0 | x/5.3 | x/1.3 | 4.3/2.2 |
| ARK_97 | Clone JF826393.1 (91) | Syndiniales | x | x | x | 1.4/x |
| ARK_100 | Clone AF290050.2 (95) | Dinophytes | x/0.5 | x/0.5 | x/0.4 | 1.4/0.4 |
| ARK_102 | Clone GU819971.1 (95) | Syndiniales | x | x/<0.1 | x/<0.1 | 3.6/x |
| ARK_103 | Clone EU818505.2 (97) | Syndiniales | x | x/<0.1 | x/<0.1 | 0.7/x |
| ARK_104 | Clone EU793381.1 (96) | Syndiniales | x/0.2 | x/<0.1 | x/<0.1 | 2.2/0.4 |

haptophytes (16.3–33.1%), stramenopiles (14.7–16.8%), and cryptophytes (0.5–2.2%). Ciliates (1–3.4%) and rhodophytes (0–0.9%) were detected by 454-pyrosequencing but, analogous to the clone library data, they appeared to be minor contributors to the respective protist communities.

Around 70% of the clone library OTUs (34/47) also occurred in the 454-pyrosequencing data set (Figure S1, supplementary material). The clone library OTUs reflected ~27% of the abundant biosphere (number of sequences ≥ 1% of total sequences) of the 454-pyrosequencing data of samples HG1, HG4 and HGS3. Moreover, the clone library of sample HGN4 reflected none of the abundant 454-pyrosequencing OTUs (Table III).

Table III. Coverage of the abundant biosphere (≥ 1%; 454-pyrosequencing) by the clone library sequences; x, absent.

| 454 OTU | Closest match (Maximum identity %) | Taxonomic group | Clones (%)/454 (%) | | | |
|---|---|---|---|---|---|---|
| | | | HG1 | HG4 | HGN4 | HGS3 |
| ARK_1 | Clone FO082268.1 (97) | Chlorophytes | x/8.0 | x/5.3 | x/1.3 | 4.3/2.2 |
| ARK_2 | Clone DQ025753.1 (98) | Chlorophytes | 0.6/3 | 0.6/1.1 | x | x |
| ARK_3 | Clone JN934683.1 (97) | Chlorophytes | 77.1/14.4 | 65.9/1.1 | x | x/1.1 |
| ARK_4 | Clone AY955010.1 (99) | Chlorophytes | 2.9/1.4 | x/1.9 | x | x/1.3 |
| ARK_5 | Clone AY955010.1 (99) | Chlorophytes | x/11.9 | x/2.2 | x | x/1.2 |
| ARK_6 | Clone AF182114.1 (98) | Haptophytes | x/1.7 | x/5.1 | x/2.2 | x/2.8 |
| ARK_7 | Clone AJ278036.1 (100) | Haptophytes | x/8.2 | x/23.1 | x/6.6 | x/12.3 |
| ARK_8 | Clone AF182114.1 (99) | Haptophytes | x | x | x | x/2.4 |
| ARK_9 | Clone AF182114.1 (100) | Haptophytes | x | x/1.1 | x | x |
| ARK_10 | Clone AF182114.2 (97) | Haptophytes | x | x | x/3.2 | x |
| ARK_11 | Clone JX840906.1 (99) | Stramenopiles | 1.1/2.8 | x | x | x |
| ARK_12 | Clone HQ867845.1 (99) | Stramenopiles | x | x/1.7 | x | x |
| ARK_13 | Clone FJ431721.1 (99) | Stramenopiles | x | x/1.0 | x | x |
| ARK_14 | Clone FJ032664.1 (99) | Stramenopiles | x | x | x/1.8 | x |
| ARK_15 | Clone HM561124.1 (100) | Dinophytes | x/2.7 | x/4.0 | x/13.8 | x/4.5 |
| ARK_16 | Clone EU793918.1 (100) | Syndiniales | x | x | x | x/2.6 |
| ARK_17 | Clone HQ869207.1 (98) | Syndiniales | x | x/1.1 | x | x |
| ARK_18 | Clone EU793175.1 (99) | Syndiniales | x | x | x/1.0 | x/1.0 |
| ARK_19 | Clone FN598275.1 (98) | Syndiniales | x | x | x/2.0 | x |
| ARK_20 | Clone JN832755.1 (99) | Syndiniales | x | x | x/1.0 | x |
| ARK_21 | Clone EU793383.1 (99) | Syndiniales | x | x | x | 1.4/1.7 |
| ARK_22 | Clone FJ431832.1 (99) | Syndiniales | x | x | x/2.2 | x |
| ARK_23 | Clone DQ186528.1 (96) | Syndiniales | x | x | x | x/1.0 |
| ARK_24 | Clone EU793554.1 (98) | Syndiniales | x | x | x | x/2.7 |
| ARK_25 | Clone EU793928.1 (99) | Syndiniales | x | x/1.3 | x | x |
| ARK_26 | Clone EU793700.1 (94) | Syndiniales | x | x/1.0 | x/1.1 | 2.2/1.3 |
| ARK_27 | Clone FJ032674.1 (98) | Ciliates | 8/1.2 | x | x | x |
| ARK_28 | Clone FJ824125.1 (98) | Cercozoa | x | x | x/1.2 | x |
| ARK_29 | Clone JF698748.1 (98) | Cercozoa | x/1.4 | x/1.3 | x | x |
| ARK_30 | Clone HM561276.1 (99) | Cercozoa | x | x | x/1.1 | x |



*Comparison of clone library and 454-pyrosequencing data set – Southern Ocean*

The clone library sequences generated from the Southern Ocean sample (ANT25) clustered into 19 different OTUs (Table IV). The majority of these OTUs affiliated in the phylogenetic tree with dinoflagellates (~42%), followed by haptophytes (~39%), stramenopiles (~8%), cryptophytes (~1%), syndiniales (~1.9%), picobiliphytes (~5%) and ciliates (~5%). The clone libraries did not contain sequences related to rhodophytes or chlorophytes.

The data set generated by 454-pyrosequencing contained 1153 different OTUs. As observed previously for the Arctic samples, the community structure suggested by 454-pyrosequencing is similar to the one suggested by the clone library data. Highest relative contributions of sequence reads were constituted by dinoflagellates (~24%), stramenopiles (~32%) and haptophytes (~31%), while cryptophytes (~1.6%), syndiniales (~2.5%), rhodophytes (~1.1%), and ciliates (~6.8%) were the only minor contributors to the data set. In contrast, picobiliphytes were not detected by 454-pyrosequencing. The majority of OTUs clustering with haptophytes were closely related to the genus *Phaeocystis*. The phylogenetic analysis of the corresponding clone library sequence for an OTU of *Phaeocystis* suggests that the near full-length sequences allow a species specific identification of the OTU as *Phaeocystis antarctica* Karsten. This was not possible based on the shorter 454-pyrosequencing read, suggesting that the cloning approach allows for a more reliable taxonomic annotation, even down to species level (Figure 2).

The overlap of clone library OTUs and 454-pyrosequencing data for the Southern Ocean sample were in a similar range as observed for the Arctic samples. More than 70% of the clone library OTUs were found in both data sets (Figure S1, supplementary material), while the clone library OTUs covered 35.7% of the abundant OTUs identified by 454 pyrosequencing (Table V).

Table IV. Phylogenetic affiliations of the Southern Ocean clone OTUs and their relative abundance in the library and the 454-pyrosequencing data set at the sampling site ANT25; x, absent.

| OTU | Closest match (Maximum identity %) | Taxonomic group | Clones (%)/454 (%) |
|---|---|---|---|
| ANT_1 | Clone SGPX577 (98) | Dinophytes | 2.9/x |
| ANT_2 | *Gyrodinium fusiforme* (99) | Dinophytes | 2.9/x |
| ANT_3 | Clone SIF_2C7 (99) | Dinophytes | 3.9/<1 |
| ANT_4 | Clone B16 (98) | Dinophytes | 1.0/<1 |
| ANT_5 | Clone SHAX878 (95) | Dinophytes | 2.9/<1 |
| ANT_6 | Clone CNCIII51_20 (99) | Dinophytes | 20.2/8.7 |
| ANT_7 | *Azadinium spinosum* (99) | Dinophytes | 7.7/1.0 |
| ANT_8 | *Gyrodinium rubrum* (96) | Dinophytes | 1.0/<1 |
| ANT_9 | DH147-EKD20 (94) | Syndiniales | 1.9/<1 |
| ANT_10 | *Salpingella acuminata* (99) | Ciliates | 4.8/3.9 |
| ANT_11 | Clone KRL01E30 (87) | Picobiliphytes | 2.9/<1 |
| ANT_12 | *Geminigera cryophila* (99) | Cryptophytes | 1.0/<1 |
| ANT_13 | Clone B1 (99) | Haptophytes | 21.2/19.9 |
| ANT_14 | Clone B1 (99) | Haptophytes | 18.3/4.3 |
| ANT_15 | Clone F11N10 (91) | Diatoms | 1.0/<1 |
| ANT_16 | *Hemiaulus sinensis* (96) | Diatoms | 1.0/<1 |
| ANT_17 | Clone RA070625T.073 (96) | Stramenopiles | 2.9/x |
| ANT_18 | Clone CNCIII05_73 (93) | Stramenopiles | 1.9/x |
| ANT_19 | Clone 14H3Te6QW (95) | Stramenopiles | 1.0/<1 |

## Discussion

Although culture-independent methods like the well-established analyses of ribosomal clone libraries and latest 454-pyrosequencing are commonly used for screening microbial community structures (Diez et al. 2001; Lovejoy et al. 2006; Cheung et al. 2010), studies that directly compare both approaches are scarce. To our knowledge, the majority of these studies focused on the genetic diversity of prokaryotes (Zhang et al. 2011). Here, we compared information on the protist community composition suggested by sequencing of clone libraries with information on the community composition generated by 454-pyrosequencing of ribosomal genes. This was done to address whether ribosomal clone libraries can complement 454-pyrosequencing data sets with near full-length information. The availability of near full-length sequences would provide a better basis for phylogenetic analyses of environmental sequences. This study is based on the analyses of Arctic picoeukaryotic protist communities and the analysis of the whole protist assemblage (micro-, nano- and picoplankton) in a sample collected in the Southern Ocean. This study involves the application of different primer sets for the amplification of the 18S rDNA fragments for the clone libraries and the 454-pyrosequencing. The use of different primer sets is crucial to amplify on one hand near full-length 18S rDNA fragments for the clone library approach and on the other hand shorter fragments of the V4 region of the 18S rDNA, that are suited for 454-pyrosequencing.

### *How does the community structure revealed by clone libraries compare to the community structure revealed by 454-pyrosequencing?*

Independent of the size fractionation, geographical location or primer usage, the two methods showed significant similarities in respect of the community composition and the recovery of clone library sequences in the 454-pyrosequencing data set. A previous study involved a similar approach

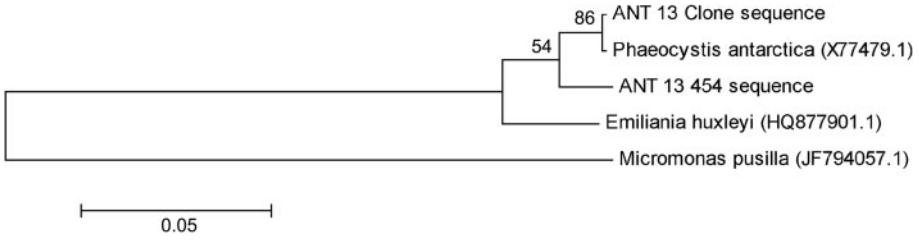


Figure 2. Phylogenetic tree based on the 18S rDNA sequences (clone sequence and 454-pyrosequencing sequence) of the ANT_13 OTU. Reference sequences were obtained from GenBank (NCBI). Accession numbers are given in parentheses. Calculation of the tree was performed with maximum likelihood under the implementation of the Jukes–Cantor model and 1000 bootstrap replications.

Table V. Coverage of the abundant biosphere (≥ 1%; 454-pyrosequencing) by the clone library sequences; x, absent.

| 454 OUT | Closest match (Maximum identity %) | Taxonomic group | Clones (%)/454 (%) |
|---|---|---|---|
| ANT_1 | Clone B1 (99) | Haptophytes | 21.2/19.9 |
| ANT_2 | Clone B1 (99) | Haptophytes | 18.3/4.3 |
| ANT_3 | *Fragilariopsis curta* (99) | Diatoms | x/4.1 |
| ANT_4 | *Hemiaulus sinensis* (95) | Diatoms | x/2.6 |
| ANT_5 | *Chaetoceros peruvianus* (99) | Diatoms | x/2.3 |
| ANT_6 | Clone ANT37-10 (99) | Diatoms | x/1.9 |
| ANT_7 | *Chaetoceros* sp. (99) | Diatoms | x/1.4 |
| ANT_8 | *Thalassiothrix longissima* (99) | Diatoms | x/1.1 |
| ANT_9 | Clone CNCIII51_20 (99) | Dinophytes | 20.2/8.7 |
| ANT_10 | Clone SCM38C10 (99) | Dinophytes | x/1.2 |
| ANT_11 | Clone SHAC491 (98) | Dinophytes | x/1.0 |
| ANT_12 | *Azadinium spinosum* (99) | Dinophytes | 7.7/1.0 |
| ANT_13 | *Salpingella acuminata* (99) | Ciliates | 4.8/3.9 |
| ANT_14 | Gonyaulacales clone (99) | Alveolates | x/1.5 |

comparing the clone library and the 454-pyrosequencing approach to assess ciliate communities (Bachy et al. 2013). They found that both molecular approaches revealed similar phylogenetic structures of the tintinnid community. This is in accordance with our findings, although they focused only on ciliates. In the present study, ~70–80% of the clone library sequences were also discovered in the corresponding 454-pyrosequencing data set, while the clone library sequences covered ~25–35% of the abundant biosphere.

The clone libraries and the 454-pyrosequencing of the Arctic samples were in good accordance for the observation that OTUs related to larger taxonomic groups, as dinoflagellates and chlorophytes, contribute the majority of sequences observed in the analysis. Furthermore, both methods agree that the picoeukaryote community was least diverse at station HG4 and most diverse at HGS3. For the Southern Ocean sample, there was also good agreement between the two methods in respect to the read abundance of the larger taxonomic groups (haptophytes, cryptophytes, syndiniales and ciliates). For the Southern Ocean sample, the two methods even agree on the OTU with the highest relative read abundance. The relatively good accordance between the two methods was rather surprising, because different primer sets were used for the amplification of the 18S rDNA. There are several studies, analysing different taxonomical levels, reporting that different primer sets applied on the same sample resulted in different diversity and abundance patterns (Jeon et al. 2008; Potvin & Lovejoy 2009; Stoeck et al. 2010).

The primer sets used in this study were checked for binding errors using the SSU Ref 108 database. All primer sets covered all major taxonomic groups. Nevertheless, we observed that the primer set used for the Arctic samples might be biased against haptophytes and the primer set used for the Southern Ocean sample against diatoms. We annotated our sequences on a higher taxonomical level and a deeper taxonomical resolution might result in more differences. Haptophytes were not detected in the Arctic clone library data sets, while the 454-data set suggests that haptophytes dominate these samples. In general, it is possible to clone haptophytes, because numerous *Phaeocystis* sp. clones have been found in this study in the Southern Ocean library. The library of the Southern Ocean sample was based on the usage of the forward primer 300F, while the 454-pyrosequencing data were based on an amplification of the 18S rDNA with 528F and 1055R. These primers might be better suited for the amplification of haptophytes than the primer 82F, even though all of them have no mismatch to known haptophytes.

### *Do clone library data exclusively reflect the abundant biosphere?*

A recent assumption of previous studies (Pedros-Alio 2006) is that clone libraries cover at least the abundant biosphere of protist communities. In this study, clone library data covered only ~25–30% of the abundant biosphere identified by 454-pyrosequencing. Thus, the occurrence of a sequence in a clone library does not entail that it is present in the abundant biosphere of a sample. Furthermore, our observations suggest that the cloning approach is even suited to retrieve taxa from the rare biosphere of the 454-pyrosequencing data set (< 1% in total). On one hand, these data suggest that sequences retrieved from clone libraries could be a random selection of ribosomal sequences in a sample. However, on the other hand, most of the sequences of the abundant biosphere

that were not detected in the clone libraries were affiliated with *Phaeocystis* (Arctic samples) or diatoms (Southern Ocean sample). These findings suggest that the primer sets used for the amplification of the near full-length 18S rDNA sequence are not optimally suited for the amplification of haptophytes and diatoms. Thus, the primer set used in this study could account for the absence of sequences of the abundant biosphere in the clone libraries (Caron et al. 2004; Countway et al. 2005). One could speculate that the coverage of the abundant biosphere could be higher, if optimized primer sets for the amplification of the near full-length ribosomal sequence were used to generate the clone libraries. Even though there is significant overlap in the OTUs recovered by clone libraries and 454-pyrosequencing, it has to be acknowledged that ~30% of the OTUs identified in the clone libraries were not present in the 454-pyrosequencing data set at all. Again, primer usage could be the reason for this observation. The primer set used in this study for 454-pyrosequencing might not be suited to amplify all sequences that are amplified by the primer set used for the generation of the clone libraries. However, the data also suggest that the number of sequence reads obtained in this study were not sufficient to analyse the diversity to saturation. Rarefaction curves are presented as a supplement (Figure S2, supplementary material). However, the calculation of rarefaction curves based on ribosomal sequences is only meaningful if all organisms have the same copy number of ribosomal genes in their genomes. A significant amount of variability in the copy number of ribosomal genes was reported for protists (Zhu et al. 2005). This makes the calculation of rarefaction curves or diversity indices doubtful.

### *Does the additional phylogenetic analysis of near full-length ribosomal genes improve the annotation of 454-sequences?*

In this study we demonstrate that the use of 18S rDNA near full-length sequences have the potential for a deeper taxonomical resolution. It demonstrated that the phylogenetic annotation of a *Phaeocystis* sp. OTU observed in the 454-pyrosequencing data set from the Southern Ocean could be refined by the use of the near full-length sequence. Using the near full-length sequence, it was possible to annotate the OTU species specifically as *Phaeocystis antarctica*, which was not possible based on the V4-sequence.

## Concluding remarks

In this study, ~70–80% of the OTUs retrieved via the cloning of 18S rDNA sequences from different samples collected in the Arctic and the Southern Ocean were also found in the corresponding 454-pyrosequencing data sets. This observation suggests that clone libraries have a certain potential to complement 454-pyrosequencing data with near full-length sequence information that could contribute to a refined phylogenetic analysis of new environmental sequences. However, the clone libraries covered only ~25–35% of the abundant biosphere in the 454-pyrosequencing data sets, while 70–80% of the clone library sequences even matched with the rare biosphere and sequences that were not present in the pyrosequencing data sets at all. Therefore, the likelihood of finding a sequence of interest retrieved by 454-pyrosequencing among the corresponding clone library sequences is rather limited. However, it might be increased by optimizing the amplification and cloning efficiency related to the generation of the 18S rDNA clone libraries to generate more clones. The work related to the analysis of increased numbers of clones could be minimized by pre-screening the clones on an agar plate for the presence of a certain ribosomal sequence read by a colony hybridization with a molecular probe derived from the ribosomal 454-sequence of interest. This would circumvent sequencing of hundreds of clones and make the idea of complementing 454-pyrosequencing data with 18S rDNA near full-length sequence information for improved phylogenetic analyses of new environmental sequences more feasible. However, it has to be kept in mind that longer sequences are one important step towards an improved phylogenetic annotation of environmental ribosomal sequences. The other step would be an optimization of ribosomal databases by complementing them with near full-length ribosomal sequences obtained from newly isolated and cultured protists.

## Acknowledgements

We thank the captain and crew of the RV *Polarstern* for their support during the cruises. We are very grateful to Stephan Frickenhaus, Fabian Kilpert and Bank Beszteri for their bioinformatical support. We also want to thank Annika Schroer, Anja Nicolaus and Kerstin Oetjen for excellent technical support in the laboratory. We are grateful to Steven Holland for providing access to the program Analytic Rarefaction 1.3.

## Funding

This study was accomplished within the Young Investigator Group PLANKTOSENS (VH-NG-500), funded by the Initiative and Networking Fund of the Helmholtz Association.

## Supplementary material

Supplementary material for this article is available via the Supplemental tab of the article's online page at http://dx.doi.org/10.1080/17451000.2013.852685.

## References

Bachy C, Dolan JR, Lopez-Garcia P, Deschamps P, Moreira D. 2013. 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 Journal 7:244–55.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T. 2011. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environmental Microbiology 13:340–49.

Caron DA, Countway PD, Brown MV. 2004. The growing contributions of molecular biology and immunology to protistan ecology: Molecular signatures as ecological tools. Journal of Eukaryotic Microbiology 51:38–48.

Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK. 2010. Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. Isme Journal 4:1053–59.

Comeau AM, Philippe B, Thaler M, Gosselin M, Poulin M, Lovejoy C. 2013. Protists in Arctic drift and land-fast sea ice. Journal of Phycology 49:229–40.

Countway PD, Gast RJ, Savai P, Caron DA. 2005. Protistan diversity estimates based on 18S rDNA from seawater incubations in the western North Atlantic. Journal of Eukaryotic Microbiology 52:95–106.

Diez B, Pedros-Alio C, Massana R. 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Applied and Environmental Microbiology 67:2932–41.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Computational Biology 7(10):e1002195. 16 pages.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–200.

Elwood HJ, Olsen GJ, Sogin ML. 1985. The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata*. Molecular Biology and Evolution 2:399–410.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proceedings of the National Academy of Sciences 106: 22427–32.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genetics 4(11):e1000255. 10 pages.

Jeon S, Bunge J, Leslin C, Stoeck T, Hong SH, Epstein SS. 2008. Environmental rRNA inventories miss over half of protistan diversity. BMC Microbiology 8:222. 13 pages.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology 12:118–23.

Lopez-Garcia P, Lopez-Lopez A, Moreira D, Rodriguez-Valera F. 2001. Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. FEMS Microbiology Ecology 36:193–202.

Lovejoy C, Potvin M. 2011. Microbial eukaryotic distribution in a dynamic Beaufort Sea and the Arctic Ocean. Journal of Plankton Research 33:431–44.

Lovejoy C, Massana R, Pedros-Alio C. 2006. Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. Applied and Environmental Microbiology 72:3085–95.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–80.

Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538. 16 pages.

Miranda LN, Zhuang YY, Zhang H, Lin S. 2012. Phylogenetic analysis guided by intragenomic SSU rDNA polymorphism refines classification of '*Alexandrium tamarense*' species complex. Harmful Algae 16:35–48.

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.

Pedros-Alio C. 2006. Marine microbial diversity: Can it be determined? Trends in Microbiology 14:257–63.

Potvin M, Lovejoy C. 2009. PCR-based diversity estimates of artificial and environmental 18S rRNA gene libraries. Journal of Eukaryotic Microbiology 56:174–81.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6:639–41.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. 2006. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. Proceedings of the National Academy of Sciences 103:12115–20.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, et al. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular Ecology 19:21–31.

Wolf C, Frickenhaus S, Kilias ES, Peeken I, Metfies K. 2013. Regional variability in eukaryotic protist communities in the Amundsen Sea. Antarctic Science 25:741–51.

Zhang XJ, Yue SQ, Zhong HH, Hua WY, Chen RJ, Cao YF, et al. 2011. A diverse bacterial community in an anoxic quinoline-degrading bioreactor determined by using pyrosequencing and clone library analysis. Applied Microbiology and Biotechnology 91:425–34.

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 Microbiology Ecology 52:79–92.

*Editorial responsibility: Hongyue Dang*