Mol. Cells *34*, 383-391, October 31, 2012
DOI/10.1007/s10059-012-0169-0
pISSN: 1016-8478 eISSN: 0219-1032

Molecules and Cells
http://molcells.org

Development of Single-Nucleotide Polymorphism-Based Phylum-Specific PCR Amplification Technique: Application to the Community Analysis Using Ciliates as a Reference Organism

Jae-Ho Jung[1,7], Sanghee Kim[2,7], Seongho Ryu[3,7], Min-Seok Kim[1], Ye-Seul Baek[1], Se-Joo Kim[1,6], Joong- Ki Choi[4], Joong-Ki Park[5], and Gi-Sik Min[1,*]

Despite recent advance in mass sequencing technologies such as pyrosequencing, assessment of culture-independent microbial eukaryote community structures using universal primers remains very difficult due to the tremendous richness and complexity of organisms in these communities. Use of a specific PCR marker targeting a particular group would provide enhanced sensitivity and more in-depth evaluation of microbial eukaryote communities compared to what can be achieved with universal primers. We discovered that many phylum- or group-specific single-nucleotide polymorphisms (SNPs) exist in small subunit ribosomal RNA (SSU rRNA) genes from diverse eukaryote groups. By applying this discovery to a known simple allele-discriminating (SAP) PCR method, we developed a technique that enables the identification of organisms belonging to a specific higher taxonomic group (or phylum) among diverse types of eukaryotes. We performed an assay using two complementary methods, pyrosequencing and clone library screening. In doing this, specificities for the group (ciliates) targeted in this study in bulked environmental samples were 94.6% for the clone library and 99.2% for pyrosequencing, respectively. In particular, our novel technique showed high selectivity for rare species, a feature that may be more important than the ability to identify quantitatively predominant species in community structure analyses. Additionally, our data revealed that a target-specific library (or ciliate-specific one for the present study) can better explain the ecological features of a sampling locality than a universal library.

INTRODUCTION

Over the last few decades, molecular sequences have become a rich source of information for studying the diversity of microbial eukaryotes. The nuclear small subunit ribosomal RNA (SSU or 18S rRNA in eukaryotes) gene is often used in phylogenetic and molecular taxonomic studies because it is ubiquitous and highly conserved in eukaryotes. In particular, high conservation of the SSU rRNA gene facilitates the utility of this gene in studies of culture-independent microbial eukaryote communities using universal PCR primers (Medlin et al., 1988; Moon-van der Staay et al., 2001; Richards et al., 2005).

Recently, high-throughput sequencing methods such as pyrosequencing have been combined with a molecular survey to assess the entire spectrum of eukaryotes belonging to communities in environmental samples from various habitats (Blackwood et al., 2005; Nolte et al., 2010; Scheckenbach et al., 2010; Shanks et al., 2011). Conventional approaches use universal primers specific for an SSU rRNA gene targeting all organisms in an environmental sample. However, predominant organisms or ones possessing genes with sequences preferentially amplified by PCR will be detected before to rare species because the eukaryotic organism communities in most environmental samples are too complex and diverse to be thoroughly analyzed in a single study. For this reason, rare species or small taxonomic groups will not be evenly represented in the sampling data even if mass sequencing methods are used (Blackwood et al., 2005; Nolte et al., 2010; Scheckenbach et al., 2010; Shanks et al., 2011). Rare species or minor taxonomic groups, however, may be more important than quantitatively dominant species when analyzing community structures due to their narrow range of distribution and susceptibility to minute

[1]Department of Biological Sciences, Inha University, Incheon 402-751, Korea, [2]Korea Polar Research Institute (KORDI), Songdo Techno Park, Incheon 406-840, Korea, [3]Department of Cell and Developmental Biology, Weill Cornell Medical College, New York, NY 10065, USA, [4]Department of Oceanography, Inha University, Incheon 402-751, Korea, [5]Department of Parasitology and Graduate Program in Cell Biology and Genetics, College of Medicine, Chungbuk National University, Cheongju 361-763, Korea, [6]Present address: Deep-sea and Seabed Resources Research Division, Korea Institute of Ocean Science and Technology, Ansan 426-744, Korea, [7]These authors contributed equally to this work.
*Correspondence: mingisik@inha.ac.kr

Received June 26, 2012; revised July 26, 2012; accepted July 27, 2012; published online September 6, 2012

Keywords: ciliate, community analysis, phylum-specific PCR, pyrosequencing, SNP, SSU rRNA

Springer

habitat alterations or ecological disturbances (Blackwood et al., 2005).

The use of primers specific for higher taxonomic groups in rRNA gene surveys would provide enhanced sensitivity and more in-depth assessments of microbial eukaryote communities compared to what can be achieved with universal primers. Recently, many efforts have been made to develop phylum- or higher taxonomic group-specific PCR primers (Blackwood et al., 2005; Dopheide et al., 2008; Scheckenbach et al., 2010; Stoeck et al., 2009). The main problem encountered in these previous studies is a difficulty in designing higher taxonomic group-specific primers that discriminate between target groups among all eukaryotes that exist in an environmental sample. Due to the highly conserved nature of the SSU rRNA gene, it is difficult to find regions that are specifically retained among all members of a target group that are distinct from those in other groups. To overcome this obstacle, a combination of several primer sets specific for as many members as possible from a certain phylum or class were developed (Dopheide et al., 2008; Holzmann et al., 2003; Lin et al., 2006).

Single-nucleotide polymorphisms (SNPs) are a common type of genetic variation which are widely used as molecular markers to identify closely related phylogenetic groups generally at levels lower than genera, such as species, population, or even individual organisms (Morita et al., 2007; Sachidanandam et al., 2001; Zou et al., 2010). However, we found that many SNPs in the SSU rRNA gene are well conserved in higher taxonomic groups such as classes, phyla, and even levels higher than phyla. Combining this knowledge with a known simple allele-discriminating PCR (SAP) method (Bui and Liu, 2009), we developed a technique for identifying organisms of a specific phylum (or higher taxonomic group) among diverse eukaryote populations based only on a single nucleotide difference. We named this new method a SNP-based phylum-specific PCR amplification technique (SPAT). Most previously developed genotyping techniques, including SAP, use SNPs for allele discrimination among the lower taxonomic groups (Haliassos et al., 1989; Konieczny and Ausubel, 1993). In contrast, SPAT targets higher taxonomic groups such as phyla, a feature that may help popularize its use. In the present study, we evaluated the reliability of SPAT using two complementary methodologies, pyrosequencing and clone library screening, for identifying ciliates as a reference organism.

MATERIALS AND METHODS

Sample collection

Estuary offshore field samples were collected using a planktonic net (20-μm mesh size) at a location 3.2 km off the coast of Incheon Sea Port (N 37° 25′, E 126° 34′), Incheon, South Korea. Collected seawater was filtered using a 200-μm sieve to remove detritus and metazoans, and then the filtered seawater was absorbed with 3-μm nitrocellulose filter papers (white SSWP, 47 mm). Top sediment layer was removed with a spatula from the tidal mudflats in Janghwa-ri (N 37° 37′, E 126° 22′), Incheon, South Korea. Cultured strains of two ciliates, *Euplotes minuta* and *Dysteria* sp., and two dinoflagellates, *Alexandrium tamarense* and *Cochlodinium polykrikoides*, were isolated from the field samples collected from seashores in South Korea. Two cultured diatom strains of *Stellarima microtrias* and *Fragilaria* sp., and two green algae strains, *Chlorella* sp. and *Ulva pertusa*, were kindly provided by Dr. H. Choi at the Korea Polar Research Institute (KOPRI).

Identification of group-specific SNPs in the SSU rRNA gene

To find group-specific SNP sites in the SSU rRNA gene, diverse eukaryotic sequences were retrieved from the NCBI database. Several representative sequences from all six supergroups of eukaryotes (Parfrey et al., 2006), including those of all 11 ciliate classes (Lynn, 2003), were included. Most representative members of the groups or supergroups analyzed in the alignment were phylogenetically distinct (Supplementary Table 1). As shown in Fig. 1, multiple alignments were generated using ClustalX (Thompson et al., 1997), and the SNP sites were estimated manually in Bioedit (Hall, 1999).

To calculate the conserved rate of a ciliate-specific SNP, all complete or nearly complete SSU rDNA sequences of ciliates identified at species-level were retrieved from the SILVA database, version 98 (Pruesse et al., 2007), and then aligned with ClustalX. The numbers of sequences containing a conserved SNP were counted. Based on these numbers, the SNP coverage rates in ciliate group were estimated at both the sequence and species levels.

Primers

A ciliate-specific reverse SPAT primer, referred to as spCiliV4-1 (Table 1), was designed based on a ciliate-specific SNP, CiliV4 (Supplementary Table 2). We located a thymine (T) at the 3′-end position of the primer that was complimentary to the CiliV4 nucleotide 'A'. At the second position of the 3′-end of the primer, 'A' instead of 'T' as a mismatched base was added (Figs. 2A and 2B). From the third position, other sequences in the primer were complementary to the universal sequences of eukaryotes. Some degenerated sequences were incorporated to improve the universality of the primer for all eukaryotes.

We used three eukaryote universal primers. EukA is located at 5′ ends of the eukaryote SSU rRNA gene (Medlin et al., 1988). The reverse primer uniEukaV4 had same sequence as spCiliV4-1, but two nucleotides of the 3′-end (SNP site) was omitted. The reverse primer uniEuka-945 was designed for clone sequencing (Table 1).

For pyrosequencing, three primers were designed to amplify the entire V4 region of the SSU rRNA gene using a 454 FLX titanium sequencer (Roche). The primers contained three different types of sequences: 454 Life Science A or B sequencing adapters (Stoeck et al., 2010), 10-bp multiplex identifier (MID) tags, and taxon-specific sequences. The taxon-specific sequence of the ciliate-specific reverse primer (CiliPyro-R) was designed with the SPAT protocol using CiliV4 as the 3′-end ciliate SNP. Both a universal forward (Pyro-F, located between V3 and V4) and a reverse universal primer (EukaPyro-R, between V4 and V5) were designed based on the comparison among the diverse eukaryote sequences in Table 1.

DNA extraction, PCR, cloning, and sequencing

Total genomic DNA was extracted with 2× CTAB as previously described (Doyle and Doyle, 1987). PCR was performed using the EukA and spCiliV4-1 (for construction of the ciliate-specific library and genomic DNA test), and EukA and uniEukaV4 (eukaryote universal library construction and genomic DNA testing) primer pairs. PCR reactions (30 μl each) were prepared containing 1 μl of genomic DNA, 200 μM each of dATP, dCTP, dGTP, and dTTP; 0.5 μl of each primer (20 pmol), 1× AccuPrime PCR Buffer II, and 1.0 U AccuPrime *Taq* DNA High-Fidelity polymerase (Invitrogen). PCR was performed in a GenAmp PCR System 2700 (ABI) with the following program: one cycle of 3 min at 92°C, then 30 cycles of 10 s at 95°C, 20 s at 52°C, and 2 min at 68°C, followed by a final extension at

Table 1. Primers used for library construction

Primer	Sequence	Direction	Position[1]	Application	Reference
Clone library					
EukA	5′-CTGGTTGATYCTGCCAGT-3′	Forward	4-21	Forward universal primer	Medlin et al. (1988)
spCiliV4-1	5′-ACGAYGGTATCTRATCRTCTAT-3′	Reverse	968-989	Ciliate specific SPAT PCR	The present study
uniEukaV4	5′-ACGAYGGTATCTRATCRTCT-3′	Reverse	969-989	Eukaryote universal PCR	The present study
uniEuka-945	5′-ATCCCCTAACTTTCGTTCTTG-3′	Reverse	946-966	Clone sequencing	The present study
Pyrosequencing					
Pyro-F	5′-AGCAGCCGCGGTAAYHCC-3′	Forward	560-577	Forward universal primer	The present study
CiliPyro-R	5′-ACGAYGGTATCTRATCRTCTAT-3′	Reverse	968-989	Ciliate specific SPAT PCR	The present study
EukaPyro-R	5′-ACGAYGGTATCTRATCRTCT-3′	Reverse	970-989	Eukaryote universal PCR	The present study

[1]Position in the *Tetrahymena canadensis* sequence (M26359).

72°C for 7 min. Each PCR product (5 μl) was separated in 1.0% agarose gel and visualized under UV light. To construct the clone library, PCR fragments from a bulked offshore environmental sample were cloned without purification using a TOPO TA cloning kit (Invitrogen) according to the manufacturer's instructions. Clones were randomly selected and sequenced with a uniEuka-945 primer using an automated ABI 310 DNA sequencer (Perkin Elmer). All the representative sequences of each operational taxonomic unit (OTU) were deposited in the GenBank database (JF727580-JF727636).

Sequence analysis of the clone libraries

Multiple alignments were generated using the ClustalX program, and pair-wise distances among the sequences were calculated using MEGA version 4.0 software (Kumar et al., 2004) with *p*-distance and pair-wise gap deletion options. A chimeric test for clone libraries was conducted using the Bellerophon program (Huber et al., 2004). However, the chimeric sequence was not sorted properly. Through manual examination of the sequences, four chimera sequences were found and removed from the data set. After applying a 1% distance cutoff to each clone library, sequences within this distance were clustered in an OTU. A single sequence was then randomly selected from each OTU for a BLAST search using the 'exclude uncultured/environmental sample sequences' option and the maximum identical sequence if more than 95% existed was retrieved from the NCBI result for each query sequence. By gathering NCBI sequences and representative OTUs, a new data set was prepared for phylogenetic analysis. Regions V2 through V4 (nucleotides at position 63 to 838 in *Tetrahymena canadensis*; 776 bp in length) were used for comparison.

Phylogenetic analyses were performed using neighbor-joining (NJ), maximum likelihood (ML), and Bayesian methods. NJ analyses were carried out using the *p*-distance method with the pair-wise gap deletion option by MEGA version 4.0. PhyML version 3.0 (Guindon and Gascuel, 2003) was used to perform the maximum likelihood analysis. To determine the appropriate DNA substitution model for phylogenetic analysis of the maximum likelihood and Bayesian inference analyses, the Akaike information criterion (AIC) was used to identify the best model of evolution that fit our data using the Modeltest program (Posada and Crandall, 1998) in the PAUP 4.0 (Swofford, 2002) software package. The model selected for the data set shown in Fig. 3 was GTR + I + G with an assumed proportion of invariable sites of 0.1806 and a gamma distribution shape parameter of 0.8278. Confidence in the resulting relationship was assessed using the bootstrap procedure with 100 replications for ML. A Bayesian inference assessment was performed with MrBayes 3.0 by simulating a Markov chain for 1,000,000 cycles, 300,000 of which were discarded as burn-in.

Pyrosequencing and analysis

To evaluate SPAT efficiency in the high-throughput sequencing technique, a bulked offshore environmental sample, which was also used to construct the clone library, was analyzed with pyrosequencing. PCR was conducted using two primer pairs: Pyro-F and CiliPyro-R (ciliate-specific) or Pyro-F and Euka Pyro-R (eukaryote universal) used for the environmental sample. Each PCR reaction (30 μl) contained 1 μl of genomic DNA and was prepared using an AccuPrime *Taq* DNA High-Fidelity polymerase kit (Invitrogen). The optimized PCR conditions were as follows: denaturation at 92°C for 3 min followed by 35 cycles of denaturation at 95°C for 10 s, annealing at 52°C for 20 s, and extension at 68°C for 1 min; then a final extension step at 72°C for 7 min. Using an identical primer pair, eight PCR reactions were carried out for each sample and the products were pooled together to prevent random amplification of specific taxa (Lahr and Katz, 2009). These PCR products were purified with the QIAquick® PCR Purification Kit (Qiagen) and the purified products were then concentrated up to 100 ng/μL using a Centricon YM-3 (Millipore). The pooled DNA was sequenced using a 454 FLX titanium sequencer. Files containing our sequences and their quality scores are available from the NCBI Short Read Archive (accession No. SRA030827).

Preliminary filtering criteria for the low-quality reads were as follows: 1) complete bearing MID sequences in both ends, 2) minimum sequence length of 300 bp (with PCR primers), 3) no Ns, 4) > 25 average quality score, and 5) no low-quality sequences of the alignment with the SILVA eukaryotes using mothur software (Schloss et al., 2009). The hypervariable V4 region was extracted from the trimmed sequences that aligned with the SILVA eukaryote reference. Chimeric sequences were detected and removed using Perseus (Quince et al., 2011). The extracted sequences were carried out to define the OTU by 1% and 3% *p*-distance cutoffs using mothur software. The rarefaction analysis was performed based on the 3% cutoff OTUs using mothur software. After performing a BLASTN search using the 'exclude uncultured/environmental sample sequences' option, stand-alone results with non-redundant sequences (Altschul et al., 1990) were transferred to MEGAN version 3.7.2, and compared to a multiple metagenomic data set (Huson et al., 2009). The LCA parameters of the MEGAN program were used

Table 2. Conservation of CiliV4, a ciliate-specific SNP site found in each ciliate class. All of the complete or nearly complete SSU rDNA sequences in the nominated ciliate sequences were retrieved from the SILVA-ARB database (version 98) and analyzed to determine the conservation of adenine (A), a ciliate-specific SNP nucleotide in the site.

Class	Sequence			Species		
	Total	Conserved		Total	Conserved	
	No.	No.	Ratio (%)	No.	No.	Ratio (%)
Armophorea	16	16	100.0	8	8	100.0
Colpodea	29	29	100.0	26	26	100.0
Heterotrichea	31	30	96.8	22	21	95.5
Karyorelictea	11	9	81.8	6	4	66.7
Litostomatea	98	92	93.9	66	64	97.0
Nassophorea	7	7	100.0	7	7	100.0
Oligohymenophorea	203	201	99.0	142	140	98.6
Phyllopharyngea	24	24	100.0	11	11	100.0
Plagiopylea	6	6	100.0	5	5	100.0
Prostomatea	8	8	100.0	8	8	100.0
Spirotrichea	337	328	97.3	155	149	96.1
Total (Mean)	770	750	(97.4)	456	443	(97.1)

with the following options: MinScore = 250; TopPercent = 0; WinScore = 0; and MinSupport = 1. To create a normalized data set between the two amplicon libraries, extra reads of the larger data set were cut off.

RESULTS

Identification of ciliate-specific SNPs in the SSU rRNA gene

From the aligned sequences which included representatives of all six eukaryotic supergroups and all 11 ciliate classes listed in Supplementary Table 1, we identified nine ciliate-specific SNPs. Eight were located in conserved regions of the SSU rRNA gene and one was located within a variable region of V9 (Supplementary Table 2). Among theses ciliate SNPs, one (referred to as CiliV4 in this study) located between the variable regions of V4 and V5 in the eukaryote SSU rRNA gene was further examined. Most ciliate sequences had 'A', and only 13 species out of 456 (20 sequences out of 770) had different nucleotides at this SNP site. Therefore, this SNP showed a high conservation rate of 97.1% among all ciliate species from the database (Table 2).

Validation of the ciliate-specific primer specificity

We evaluated the specificity of the spCiliV4-1 primer using genomic DNA extracted from members of diverse eukaryote groups and bulk environmental samples. When we used the universal primer pair, EukA and uniEukaV4, amplified bands were present in all samples. When we used the EukA and ciliate-specific spCiliV4-1 primer pair, only DNA from ciliates and environmental samples was successfully amplified. However, a weak band was observed in the DNA sample from *Chlorella* sp. (green algae), presumably due to non-specific PCR amplification (Fig. 2C). This result concurs with our expectation and indicates that the spCiliV4-1 primer was highly specific for ciliate sequences.

Application to a clone library screen

In the ciliate-specific library, 94.6% (175 out of 185) resulted in ciliate sequences as the most similar sequences in a NCBI BLAST search using the 'exclude uncultured/environmental sample sequences' option (Table 3). In contrast, hit ratio for the ciliate sequences in the universal library dropped dramatically, and only 14.5% (24 out of 166) were matched to ciliates. By applying a 1% *p*-distance variation as the OTU cutoff, 19 out of 23 OTUs (82.6%) were matched to ciliates in the ciliate-specific library. Meanwhile, only four among 33 OTUs (12.1%) were identified as ciliates in the universal library (Table 3).

In the ciliate-specific clone library, three putative planktonic ciliate OTUs were the most predominant, accounting for 69.7% of all ciliate sequences. The most duplicated OTU was CiliV4-224, which showed 100% similarity to *Codonellopsis nipponica* (NCBI accession No. FJ196072) as presented in Fig. 3. The CiliV4-224 accounted for 43.4% of all ciliate sequences. The second and third most duplicated OTUs were CiliV4-93 (18.9% occurrence frequency rate and 100% matched with *Parastrombidinopsis shimi,* AJ786648) and CiliV4-14 (7.4% occurrence frequency rate and 99% matched with *Strombidinopsis acuminata,* FJ790207), respectively. Interestingly, apart from one OTU (EukaV4-89) represented by only one sequence, the other three ciliate OTUs detected in the universal clone library were clustered with OTUs of the ciliate-specific clone library with a 1% distance cutoff. Specifically, two OTUs represented 91.7% of all ciliate sequences in the universal clone library. The most predominant OTU, EukaV4-148, accounted for 70.8% of all ciliate sequences and was identical (100% similarity) to *C. nipponica,* FJ196072. The second most predominant OTU, EukaV4-17, represented 20.8% of the ciliate sequences and clustered with *P. shimi,* AJ786648 (Fig. 3).

Application to high-throughput pyrosequencing analysis

Universal and ciliate-specific pyrosequencing data sets containing the entire V4 region of the SSU rRNA gene were also obtained. From the quality-filtering process, we removed 42 and 21 chimeric sequences from ciliate-specific and universal data set, respectively. Finally, 2,983 reads for the ciliate-specific and 4,554 for the universal data sets were used for further analysis. BLASTN was used to match these reads to known sequences in the NCBI database using the 'exclude uncultured/environmental sample sequences' option to exclude possible hits for

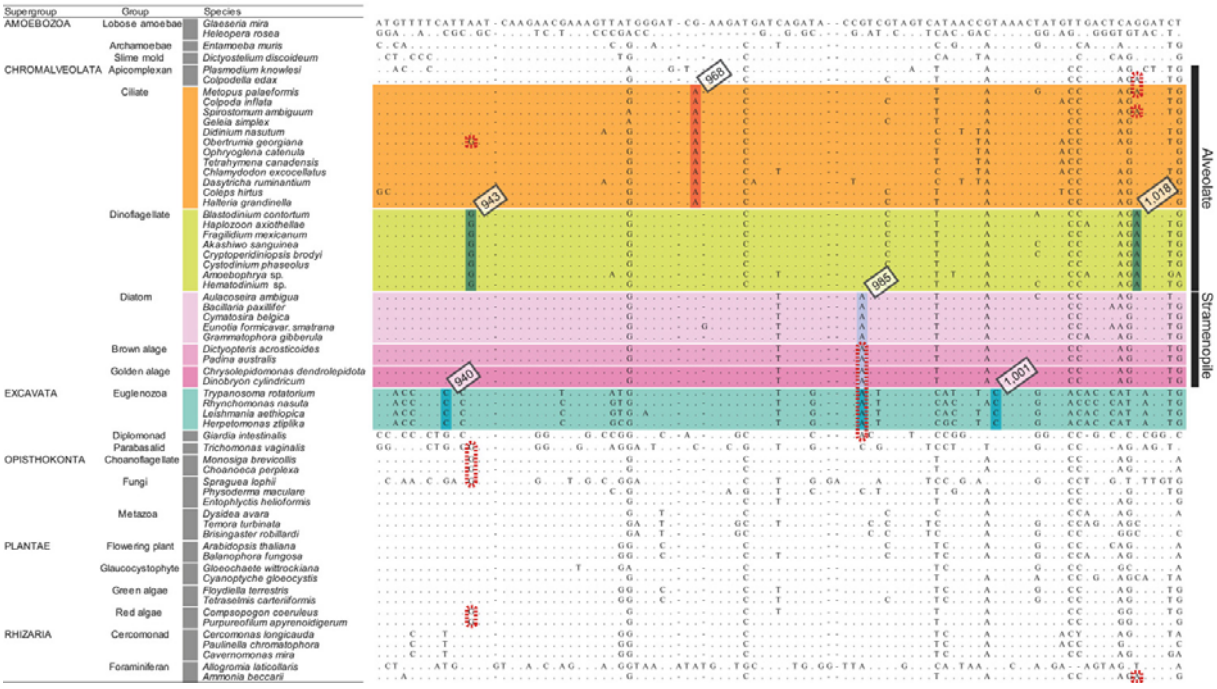

Fig. 1. Group-specific SNPs located in the conserved area between the V4 and V5 variable regions of the SSU rRNA gene were identified based on a multiple alignment retrieved from the NCBI database. Colored narrow boxes indicate the specific SNPs of the ciliates (adenine at the 968th position in *T. canadensis*), dinoflagellates (guanine and adenine at the 943rd and 1,018th positions, respectively), diatoms (adenine at the 985th position), and euglenoids (cytosine and cytosine at the 940th and 1,001st positions, respectively). Detailed information for each species is listed in Supplementary Table 1.

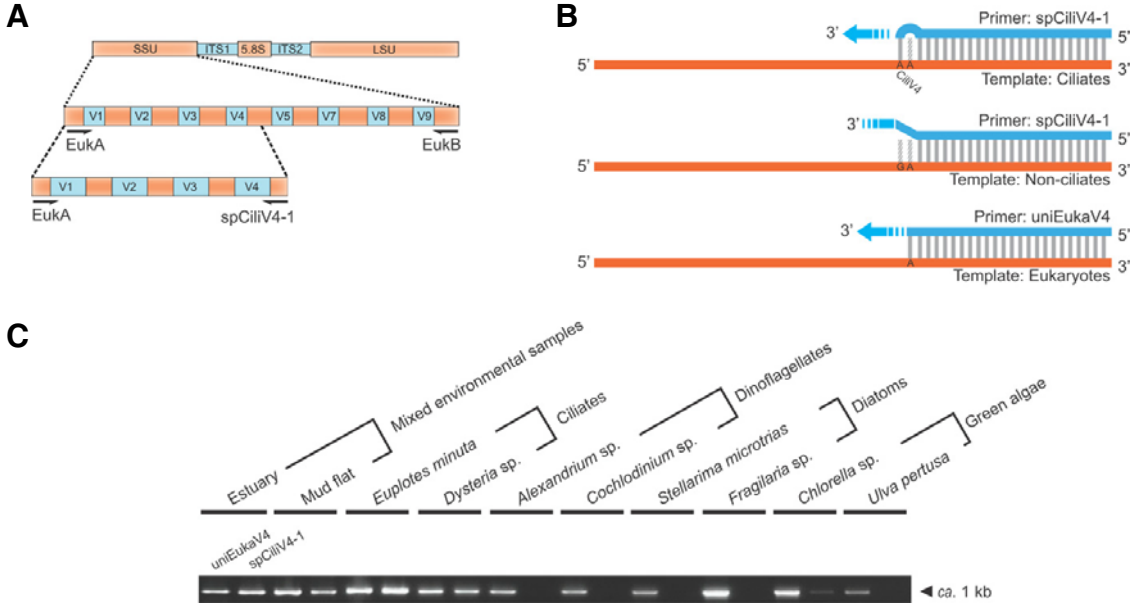

Fig. 2. SPAT strategy. (A) Schematic representation of the universal eukaryotic rRNA gene unit and primer locations used in this study. (B) Outline of the SPAT strategy. The reverse primer for ciliate-specific PCR amplification, spCiliV4-1, has one mismatched base compared to the ciliate template at the second position from the 3′-end of the primer. EukA was used as the forward primer (upper). Two nucleotides are mismatched at the first and second positions from the 3′-end of the primer compared to the non-ciliate group templates; these could not be used for PCR amplification (middle). Normal PCR amplification was expected for both the ciliate and non-ciliate groups when the two universal primers (EukA and uniEukaV4) were used (bottom). (C) Evaluation of spCiliV4-1 specificity using genomic DNA extracted from microorganisms belonging to diverse groups of eukaryotes, and bulk environmental samples. PCR amplification was performed using ciliate-specific (EukA and spCiliV4-1) and universal (EukA and uniEukaV4) primer pairs.

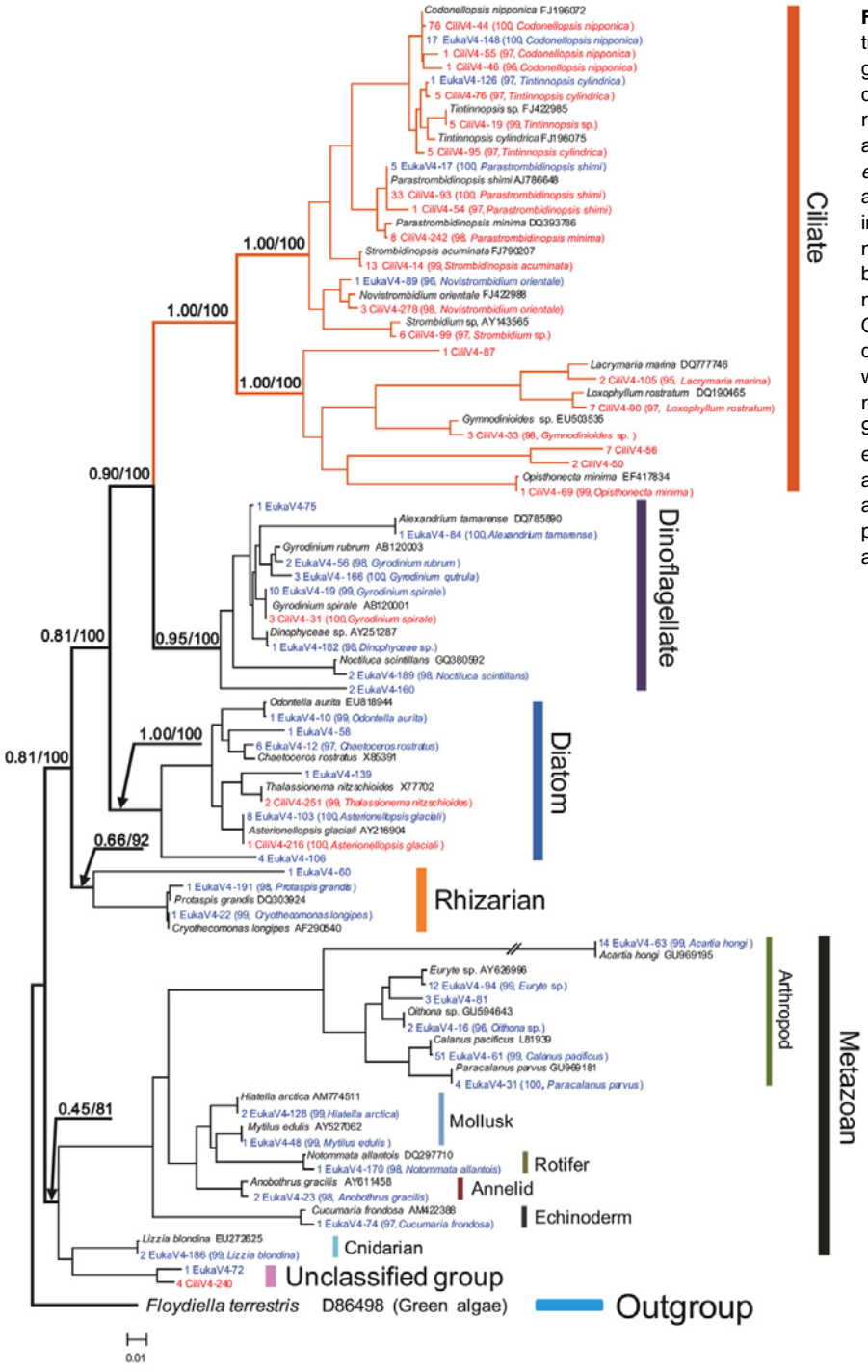

Fig. 3. A neighbor-joining phylogenetic tree constructed based on the V2-V5 region of the SSU rRNA gene including sequences from the NCBI (black), and representative OTUs from ciliate-specific (red) and universal (blue) clone libraries. *Floydiella terrestris* (green algae) was used as an outgroup. NCBI sequences are shown in the tree and labeled with the species name plus the GenBank accession number. OTUs are listed in the following order: number of sequences included in each OTU, names of the representative sequence of each OTU, and parenthesis in which the similarity (%) and species name retrieved from NCBI, if showing more than 95% similarity with the OTU sequences exist. Numbers appearing at the branches are the bootstrap values obtained from ML analysis (right) and Bayesian posterior probability (left). The scale bar represents a distance of 0.01 substitutions per site.

unidentified environmental sample sequences. For the MEGAN analysis based on the BLASTN results, 2,803 reads from the ciliate-specific and 4,431 from the universal data sets were assigned to known eukaryote taxa. In the ciliate-specific data set, 99.2% reads (2,780 out of 2,803) were identified as ciliates while only 11.3% (499 out of 4,431) clustered with known ciliates in the universal data set (Supplementary Fig. 2). When we applied 1% and 3% distance cutoffs for OTU determination using the mothur program, 436 and 191 from the ciliate-specific and 59 and 38 OTUs from the universal data sets were identified as ciliate OTUs (Table 3). The number of OTUs increased with the number of reads, and a plot of OTUs *vs.* the number of sequences resulted in rarefaction curves that did not approach plateaus, in both of ciliate-specific and universal pyrosequencing libraries (Supplementary Fig. 2).

In a multiple-comparative tree based on NCBI taxonomy constructed by MEGAN, 64 from the ciliate-specific and 152 from the universal data sets were designated as nominated eukaryotic species. In the ciliate-specific data set, 85.9% (55 out of 64) resulted in ciliates. For the universal data set, only

Table 3. Summary of the sequences and OTUs identified using clone libraries and pyrosequencing with CiliV4-based ciliate-specific or eukaryote universal primers

	Clone library				Pyrosequencing					
	Specific primer		Universal primer		Specific primer			Universal primer		
	No. of seq. (%)	No. of OTUs* (%)	No. of seq. (%)	No. of OTUs* (%)	No. of seq. (%)	No. of OTUs* (%)	No. of OTUs** (%)	No. of seq. (%)	No. of OTUs* (%)	No. of OTUs** (%)
Ciliate										
Litostomatea	9 (4.7)	2 (8.7)			41 (1.5)	13 (2.9)	10 (5.0)	7 (0.2)	2 (0.3)	2 (0.6)
Oligohymenophorea	4 (2.1)	2 (8.7)			63 (2.2)	32 (7.1)	25 (12.5)	2 (0.0)	2 (0.3)	2 (0.6)
Phyllopharyngea					8 (0.3)	7 (1.6)	5 (2.5)	1 (0.0)	1 (0.2)	1 (0.3)
Prostomatea					71 (2.5)	29 (6.5)	21 (10.5)			
Spirotrichea	157 (82.6)	13 (56.5)	24 (14.5)	4 (12.1)	2,587 (92.3)	350 (78.1)	126 (63.0)	488 (11.0)	53 (9.1)	32 (9.8)
Unidentified ciliates	10 (5.3)	2 (8.7)			10 (0.4)	5 (1.1)	4 (2.0)	1 (0.0)	1 (0.2)	1 (0.3)
Total ciliates	180 (94.7)	19 (82.6)	24 (14.5)	4 (12.1)	2,780 (99.2)	436 (97.3)	191 (95.5)	499 (11.3)	59 (10.2)	38 (11.7)
Non-ciliates										
Dinoflagellate	3 (1.6)	1 (4.4)	22 (13.3)	8 (24.2)	7 (0.2)	3 (0.7)	2 (1.0)	867 (19.6)	122 (21.0)	64 (19.7)
Diatom	3 (1.6)	2 (8.7)	21 (12.6)	5 (15.2)	6 (0.2)	3 (0.7)	3 (1.5)	2,633 (59.4)	248 (42.7)	101 (31.1)
Rhizarian			3 (1.8)	3 (9.1)				112 (2.5)	61 (10.5)	57 (17.5)
Metazoan			95 (57.2)	12 (36.4)	2 (0.1)	2 (0.4)	2 (1.0)	280 (6.3)	62 (10.7)	38 (11.7)
Other groups	4 (2.1)	1 (4.4)	1 (0.6)	1 (3.0)	8 (0.3)	4 (0.9)	2 (1.0)	40 (0.9)	29 (5.0)	27 (8.3)
Total non-ciliates	10 (5.3)	4 (17.4)	142 (85.5)	29 (87.9)	23 (0.8)	12 (2.7)	9 (4.5)	3,932 (88.7)	522 (89.8)	287 (88.3)
Total	190	23	166	33	2,803***	448	200	4,431***	581	325

*1% and **3% cutoffs were used for OTU identification.
***180 and 123 reads were excluded from the ciliate-specific and universal data sets, respectively, because these reads were designated as 'Not assigned' or 'No hits' in the analysis with the MEGAN software.

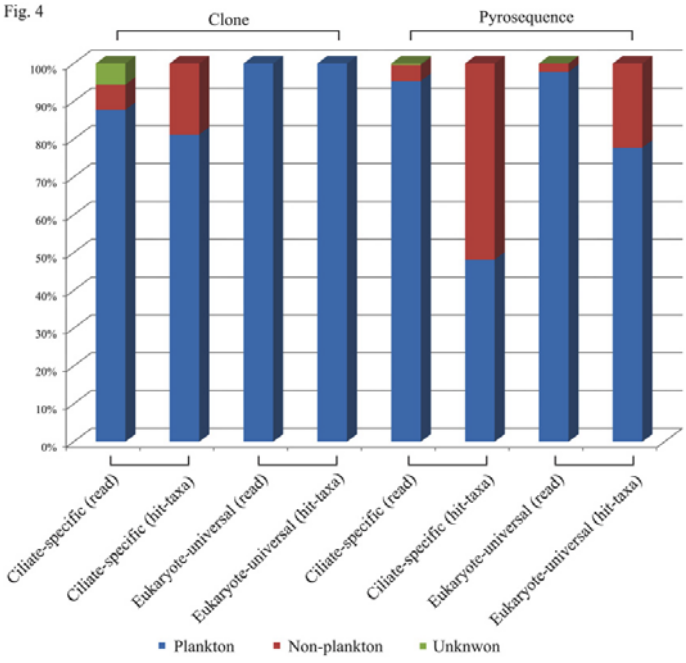

Fig. 4. Relative histogram of planktonic and non-planktonic ciliates in the clone (Fig. 3) and pyrosequencing libraries (Supplementary Fig. 1). Life forms of the ciliates were identified based on the criteria established by Lynn (2008).

12.5% of the species (19 out of 152) were matched to ciliates. Sixteen species were shared by both data sets; therefore, 70.9% (39 out 55) of species in the ciliate-specific data set were unique while only three species (15.8%) in the universal data

set were unique (Supplementary Table 3). Similar to the findings in the clone libraries, two planktonic spirotricheans, *C. nipponica* and *P. shimi,* were the most predominant species and occupied more than 67% of all ciliates from both data sets (Supplementary Table 3; 67.3% in ciliate-specific data set and 75.8% in the universal data set).

DISCUSSION

This study was initiated after observing that many phylum- or higher taxonomic group-specific SNPs exist in the SSU rRNA genes from diverse groups of eukaryotes. We found nine ciliate-specific SNPs in the SSU rRNA gene. Two of them, CiliV2 and CiliV4, are unique to ciliates while the other seven SNPs are shared by ciliates and other groups of microorganisms (Supplementary Table 2). Along with the ciliate-specific SNPs, we identified many unique or shared group-specific SNPs in other higher eukaryote taxonomic groups, such as dinoflagellates, diatoms, and euglenoids, in the conserved regions of the complete SSU rDNA sequences. This enabled us to design specific primers for each of these monophyletic groups (unpublished data).

SPAT reliability was verified both by analytic and experimental approaches. The coverage rate used in this study was defined as the number of species in the target group that will be targeted by SPAT. This rate was calculated by estimating the rate of SNP conservation in all known complete or nearly complete ciliate sequences in the SILVA database (version 98). CiliV4, a ciliate-specific SNP used to develop the ciliate-specific SPAT primers in this study, showed a high coverage rate of 97.1%, which indicates that this proportion of extant ciliate species will most likely be targeted by SPAT (Table 2). When we used SPAT primers that were designed using CiliV4 as an SNP site, the specificities of these primers for ciliates in a bulk estuary environmental sample were 94.6% for the clone library and 99.2% for pyrosequencing. These values are higher than that (73%) determined with recently developed ciliate-specific primers (Dopheide et al., 2008).

For most other group-specific PCR methods, especially those involving higher taxonomic groups, one of the most difficult tasks is perhaps identifying the priming regions of eukaryote SSU rDNA sequences. It is not easy to find a priming region with a large degree of similarity among targeted groups. For this reason, several candidate primers were initially developed in previous studies, and the most efficient primer pair was then chosen through a tedious and time-consuming experimental process (Dopheide et al., 2008). However, this is not an issue with SPAT. Since phylum or higher taxonomic group-specific SNPs are abundant and distributed evenly throughout the conserved regions of the complete SSU rRNA gene sequences, diverse SPAT primers can be easily developed for most groups of monophyletic eukaryotes. This would be especially beneficial for next-generation sequencing (NGS) technologies, such as pyrosequencing, because the length restriction of reads is rather stringent due to technical limitations. Furthermore, selecting target regions from pyrosequencing is much more difficult compared to conventional sequencing methodologies (Mashayekhi and Ronaghi, 2007; Nossa et al., 2010). Nevertheless, priming sites that include the most informative region with a moderate length can be easily selected if the SPAT protocol is used when designing the pyrosequencing primers.

More recent studies have claimed that rare species from minor taxonomic groups identified by community structural analysis are more important than quantitatively predominant species (Carrino-Kyker and Swanson, 2008; Nolte et al., 2010; Richards and Bass, 2005; Scheckenbach et al., 2010; Shanks et al., 2011). Generally, it is difficult to discover rare species sequences unless the data set is sufficiently large. In a comparative analysis of ciliate-specific and universal clone libraries (Fig. 3 and Table 3), 19 ciliate OTUs out of 180 sequences were identified in the ciliate-specific library while only four ciliate OTUs out of 24 sequences were identified in the universal library. Except for one OTU that included a single sequence (EukaV4-89), three other ciliate OTUs, including 23 sequences from the universal clone library, were clustered with predominant OTUs from the ciliate-specific clone library (Fig. 3). This tendency was also noted in larger data sets such as ones used for pyrosequencing (Table 3). Based on the results from our study, we can conclude that SPAT as a group-specific PCR amplification method is a promising technique for detecting diverse species, and is especially efficient for discovering rare species or OTUs in bulk materials such as environmental samples.

Our data also revealed that a SPAT library (or a ciliate-specific one in the present study) library can better elucidate the ecological features of a sampling locality than a universal library. Environmental samples used for the community analysis of this study were collected from the surface water 3.2 km off the coast of Incheon Sea Port (at a depth of approximately 30 m). Therefore, planktonic ciliates were expected to be the predominant microorganisms. However, the benthic environment of this sampling site is likely disturbed by high tide. Benthic ciliates can thus be readily found in the surface water because the sampling site has a large tidal range and can be affected by the Han River as well. We can therefore hypothesize that many benthic ciliate species are transported from benthic sediment and introduced into the surface water by tidal disturbances. This potential ecological scenario was reflected by the SPAT libraries in this study. In our study, most ciliate sequences (or reads) in both the clone and pyrosequencing libraries were best matched with planktonic ciliates in the NCBI database (Lynn, 2008) that accounted for 88.2-100% of the total ciliates (Supplementary Table 3). While base on hit-taxa, the number of hit-taxa in non-planktonic (or benthic) ciliates was significantly increased in the ciliate-specific clone and pyrosequencing libraries. In particular, more than 50% of the ciliate hit-taxa in the pyrosequencing library were non-planktonic (Supplementary Table 3 and Fig. 4). Additionally, it is worth noting that all non-planktonic hit-taxa are rare because each accounts for less than 1% of whole ciliate reads.

It is certain that SNPs can be found rather abundantly in a diverse number of organisms at various levels, from shallow phylogenetic groups such as populations and species (Morita et al., 2007; Sachidanandam et al., 2001; Zou et al., 2010) to phyla or above, as suggested by the results of our study. However, most previously developed SNP genotyping methods, including SAP, have focused on lower taxonomic groups (Haliassos et al., 1989; Konieczny and Ausubel, 1993). The SPAT results in this study clearly demonstrated that SNPs can be successfully used for the analysis of higher taxonomic groups as well. Considering the efficacy and simple primer design strategy associated with SPAT along with the abundance of SNPs in diverse eukaryotic organisms at various taxonomical levels, SPAT can be broadly used for various studies in numerous fields including molecular phylogenetics and evolution as well as community ecology.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

We thank Dr. Rob DeSalle and Dr. Mark Siddall (American Museum of Natural History) for their valuable input and comments on earlier drafts of this manuscript. This work was partially supported by the Basic Science Research Program (2012R1A1A2006835) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (G.-S. Min), and the Basic Research Program of the Korea Polar Research Institute project (PE12030; S. Kim).

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

Blackwood, C.B., Oaks, A., and Buyers, J.S. (2005). Phylum- and class-specific PCR primers for general microbial community analysis. Appl. Environ. Microb. *71*, 6193-6198.

Bui, M., and Liu, Z. (2009). Simple allele-discriminating PCR for cost-effective and rapid genotyping and mapping. Plant Methods *5*, 1.

Carrino-Kyker, S.R., and Swanson, A.K. (2008). Temporal and spatial patterns of eukaryotic and bacterial communities found in vernal pools. Appl. Environ. Microb. *74*, 2554-2557.

Dopheide, A., Lear, G., Stott, R., and Lewis, G. (2008). Molecular characterization of ciliate diversity in stream biofilms. Appl. Environ. Microb. *74*, 1740-1747.

Doyle, J.J., and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. *19*, 11-15.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696-704.

Haliassos, A., Chomel, J.C., Grandjouan, S., Kruh, J., Kaplan, J.C., and Kitzis, A. (1989). Detection of minority point mutations by modified PCR technique: a new approach for a sensitive diagnosis of tumor-progression markers. Nucleic Acids Res. *17*, 8093-8099.

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. *41*, 95-98.

Holzmann, M., Habura, A., Giles, H., Bowser, S.S., and Pawlowski, J. (2003). Freshwater foraminiferans revealed by analysis of environmental DNA samples. J. Eukaryot. Microbiol. *50*, 135-139.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics *20*, 2317-2319.

Huson, D.H., Richter, D.C., Mitra, S., Auch, A.F., and Schuster, S.C. (2009). Methods for comparative metagenomics. BMC Bioinformatics *10*, S12.

Konieczny, A., and Ausubel, F.M. (1993). A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J. *4*, 403-410.

Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. *5*, 150-163.

Lahr, D.J.G., and Katz, L.A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques *47*, 857-863.

Lin, S.J., Zhang, H., Hou, Y., Miranda, L., and Bhattacharya, D. (2006). Development of a dinoflagellate-oriented PCR primer set leads to detection of picoplanktonic dinoflagellates from long island sound. Appl. Environ. Microb. *72*, 5626-5630.

Lynn, D.H. (2003). Morphology or molecules: how do we identify the major lineages of ciliates (Phylum Ciliophora)? Eur. J. Protistol. *39*, 356-364.

Lynn, D.H. (2008). The ciliated protozoa: characterization, classification, and guide to the literature (New York: Springer).

Mashayekhi, F., and Ronaghi, M. (2007). Analysis of read length limiting factors in Pyrosequencing chemistry. Anal. Biochem. *363*, 275-287.

Medlin, L., Elwood, H.J., Stickel, S., and Sogin, M.L. (1988). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene *71*, 491-499.

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

Morita, A., Nakayama, T., Doba, N., Hinohara, S., Mizutani, T., and Soma, M. (2007). Genotyping of triallelic SNPs using TaqMan® PCR. Mol. Cell. Probes *21*, 171-176.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwalder, B., Boenigk, J., and Schlotterer, C. (2010). Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol. Ecol. *19*, 2908-2915.

Nossa, C.W., Oberdorf, W.E., Yang, L.Y., Aas, J.A., Paster, B.J., DeSantis, T.Z., Brodie, E.L., Malamud, D., Poles, M.A., and Pei, Z.H. (2010). Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. World J. Gastroenterol. *16*, 4135-4144.

Parfrey, L.W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D.J., and Katz, L.A. (2006). Evaluating support for the current classification of eukaryotic diversity. PLoS Genet. Genet. *2*, 2062-2073.

Posada, D., and Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics *14*, 817-818.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W.G., Peplies, J., and Glockner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. *35*, 7188-7196.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011). Removing noise from pyrosequenced amplicons. BMC Bioinformatics *12*, 38.

Richards, T.A., and Bass, D. (2005). Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. Curr. Opin. Microbiol. *8*, 240-252.

Richards, T.A., Vepritskiy, A.A., Gouliamova, D.E., and Nierzwicki-Bauer, S.A. (2005). The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. Environ. Microbiol. *7*, 1413-1425.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-933.

Scheckenbach, F., Hausmann, K., Wylezich, C., Weitere, M., and Arndt, H. (2010). Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. Proc. Natl. Acad. Sci. USA *107*, 115-120.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microb. *75*, 7537-7541.

Shanks, O.C., Kelty, C.A., Archibeque, S., Jenkins, M., Newton, R.J., McLellan, S.L., Huse, S.M., and Sogin, M.L. (2011). Community structures of fecal bacteria in cattle from different animal feeding operations. Appl. Environ. Microbiol. *77*, 2992-3001.

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M.J., Chistoserdov, A., Orsi, W., and Edgcomb, V.P. (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol. *7*, 72.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., and Richards, T.A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. *19*, 21-31.

Swofford, D. (2002). PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. *25*, 4876-4882.

Zou, F., Lee, S., Knowles, M.R., and Wright, F.A. (2010). Quantification of population structure using correlated SNPs by shrinkage principal components. Hum. Hered. *70*, 9-22.