SCIENTIFIC REPORTS
nature research

OPEN

Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform

Jaeho Choi[1] & Jong Soo Park[1,2,3*]

Illumina sequencing is a representative tool for understanding the massive diversity of microbial eukaryotes in natural ecosystems. Here, we investigated the eukaryotic community in a pond (salinity of 2–4) on Dokdo (island) in the East Sea, Korea, using Illumina sequencing with primer sets for the V4 and V9 regions of 18S rDNA from 2016 to 2018 for the first time. Totally, 1,413 operational taxonomic units (OTUs) and 915 OTUs were detected using the V9 and V4 primer sets, respectively. Taxonomic analyses of these OTUs revealed that although the V4 primer set failed to describe the extant diversity for some major sub-division groups, the V9 primer set represented their diversity. Moreover, the rare taxa with <1% of total reads were exclusively detected using V9 primer set. Hence, the diversity of the eukaryotic community can vary depending on the choice of primers. The Illumina sequencing data of the V9 region of 18S rDNA may be advantageous for estimating the richness of the eukaryotic community including a rare biosphere, whereas the simultaneous application of two biomarkers may be suitable for understanding the molecular phylogenetic relationships. We strongly recommend both biomarkers be used to assess the diversity and phylogenetic relationship within the eukaryotic community in natural samples.

The advent of next-generation sequencing (NGS) led to a substantial change in the previous knowledge about the microbial diversity in natural ecosystems[1,2]. NGS targeting the 18S rDNA is usually used to evaluate the diversity within all domains of life and provides large quantities of sequencing data for individual investigators. The 454 platform is hardly used anymore and a lot of other platforms are currently used far more widely. The Illumina platforms are representative tools for the investigation of the microbial community although this method can read a relatively short fragment of sequence (200 bp–500 bp) due to the technical limitations[3,4]. Illumina platform is a cost-effective tool per base (priced ~100 times lower than the 454 platform) and can now read longer sequences (200 bp–300 bp) than the initial platform[5,6]. Furthermore, the error rate of the Illumina platform is lower than that of the 454 platform[7]. However, considering the revolutionary application of the Illumina platform, our knowledge about the diversity and phylogenetic relationship of eukaryotes remains poor in field surveys such as those for brackish water[8].

Several studies on the diversity of eukaryotes noted that the V1–V2, V3, V4, and V9 regions of 18S rDNA have been used for better understanding the massive diversity of microbial community[9–11]. The V4 (expected amplicon size, 270 bp–387 bp) and V9 (expected amplicon size, 96 bp–134 bp) regions are considered to be the popular for metabarcoding[12,13]. Based on an *in silico* analysis, the V9 region of 18S rDNA offers the advantage to reveal the extant diversity of eukaryotes, whereas the V4 region of 18S rDNA is commonly used for studying the phylogenetic relationship of eukaryotes[14]. Despite these advantages of both V4 and V9 regions of 18S rDNA, multiple primer sets have been employed rarely for environmental samples[15–17]. Furthermore, only a few specific eukaryotic groups (e.g., Chlorophyta, Trichomonads, Cercozoa, Radiolarians, and Ciliophora) have been elucidated for their diversity and phylogenetic relationship in environmental samples using NGS methods[10,18–21]. Subsequently, comparative analyses of the V4 and V9 regions have been uncharted in field surveys, and most

[1]Department of Oceanography, School of Earth System Sciences, Kyungpook National University, Daegu, 41566, Republic of Korea. [2]Research Institute for Dok-do and Ulleung-do Island, Kyungpook National University, Daegu, 41566, Republic of Korea. [3]Kyungpook Institute of Oceanography, Kyungpook National University, Daegu, 41566, Republic of Korea. *email: jongsoopark@knu.ac.kr

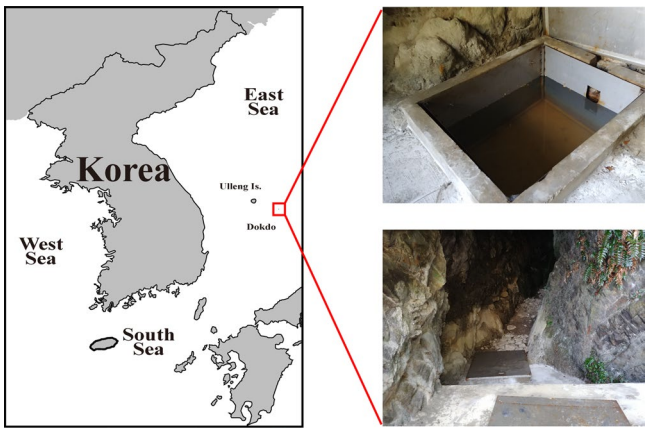

Figure 1. Study area and sampling site (the Mulgol pond) on Dokdo (red square box), Republic of Korea. Note that a map was created using the Mapping Toolbox in MATLAB (MATLAB R2019b). Is.: Island.

important eukaryotic groups remain unexplored using the Illumina platform for both the V4 and V9 regions[22]. Further, most NGS studies had been focused mainly on the diversity of dominant eukaryotic groups rather than on the rare eukaryotic taxa that play a crucial role in the eukaryotic community in natural ecosystems[23,24]. Therefore, fundamental interrogations remain unanswered regarding the region of 18S rDNA that is appropriate to describe the diversity and phylogenetic relationship of the dominant or rare eukaryotic groups using the Illumina platform in field surveys, particularly of brackish water.

In the present study, we investigate the eukaryotic community in a brackish small pond on Dokdo (island) using the Illumina sequencing platform with the V4 and V9 regions of 18S rDNA from August 2016 to June 2018. The relative abundance of the major eukaryotes at the Class level in the original supergroups 'Amoebozoa', 'Archaeplastida', 'Chromalveolata including Stramenopiles, Cryptista, Haptista, and Alveolata', 'Excavata', 'Opisthokonta', and 'Rhizaria', represent inconsistent results based on the sequencing data of the V4 and V9 regions sequencing data of 18S rDNA. The V9 region can reveal most of the important eukaryotes or rare eukaryotic taxa, whereas the V4 region remains poorly covered in field samples. In the phylogenetic analyses of enormous eukaryotes, the V4 region has a much better resolution than the V9 region. Although the molecular detection of eukaryotes in the field samples is still far from complete, the assignment of environmental sequences depends on the choice of primer regions at the class (mostly) or higher level. Furthermore, the V4 region that is longer than the V9 region is appropriate to explain the previous evolutionary relationship of eukaryotes.

Results

Characteristic of the Mulgol pond in Dokdo (island). The water temperature of the Mulgol pond was between 14.5 °C and 17.6 °C in August 2016, September 2017, and April and June 2018. The salinities of the pond ranged from 1.49 to 2.70, indicating that this pond contained brackish water at the time of sampling. Light intensity of surface water was measured to be ~4.5 lux. However, this value could be much lower because of the metal lid blocking light source from the Mulgol pond (Fig. 1). The concentration of chlorophyll-*a* ranged from 0.07 to 0.50 μg L^{-1} during the whole study period.

Overall Illumina sequencing data. A total of 780,706 V9 reads and 874,604 V4 reads were obtained by Illumina sequencing within a total of ~24 L of subsamples (6 L × 4 surveys in two years) of brackish water (Table 1). Sequence reads were filtered when the sequences had an ambiguous base, low quality (quality score offset of 33), chimera, and short reads (less than 36 bp) (Table 1). Total read count was 534,846 for the V9 region of 18S rDNA and 584,264 for the V4 region of 18S rDNA (Table 1). The length of the V9 sequence fragments ranged from 123 bp to 215 bp (average, 141 bp), whereas the length of the V4 sequence fragments was between 146 bp and 564 bp (average, 300 bp). These reads were assigned to OTUs at the same level of sequence identity. Individual sequences were clustered at 97% identity threshold. This study used a 97% identity threshold to cluster sequences to assign OTUs because of an agreement with several recent studies that used these primer sets[14,16,25]. With a 97% identity threshold, 1,632 V9 OTUs and 1,122 V4 OTUs were obtained (Table 1). After filtering OTUs assigned to prokaryotes, the unambiguous numbers of eukaryotic OTUs were 1,413 V9 region sequences and 915 V4 region sequences (Table 1).

Diversity analyses. Rarefaction analysis was conducted to determine whether OTUs in data has been sufficiently covered (Fig. 2). The saturation phase of the two rarefaction curve for the V4 and V9 regions of 18S rDNA indicated that the coverage of Illumina sequencing was sufficient during the whole study period. Eukaryotic OTUs were categorized into major sub-division groups (mostly class level) as reported by Adl *et al.*[26]. (see appendix 3. table 3.1). Of total reads, 57,997 V9 reads and 131,413 V4 reads could not be assigned to the class level, but these unclassified class reads could be successfully assigned to their supergroups at the highest rank level. Thus, 99.99% of total V9 data and 99.95% of total V4 data reads could be successfully assigned to their supergroups. The remnants of the unknown reads, which do not fall into a specific supergroup, were assigned to the 'non-assigned'

Sequence description	V4	V9
Total bases	378,053,256	138,975,040
Read count	874,604	780,706
Filtered read count	584,264	534,846
Ambiguous	0	78
Low-quality	2,876	5
Chimera	20,040	25,204
Other (non-sequencing error)	267,424	220,573
OTUs (Total)	1,122	1,632
OTUs (Eukaryotic reads)	915	1,413

Table 1. Summary of Illumina sequence data from the V4 and V9 regions. The ambiguous indicates filtered sequences with ambiguous base calls. The low-quality indicates filtered sequences with low-quality bases (Quality score offset 33). The term 'other' was defined as a non-sequencing error, which indicates query coverage and identity percentage with < 85%.

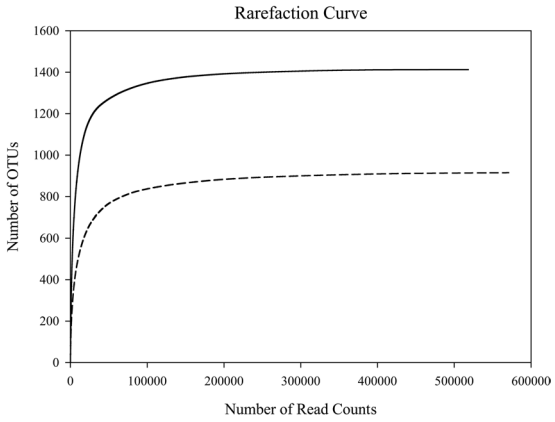

Figure 2. Diversity rarefaction curve of the V4 (dotted line) and V9 (solid line) operational taxonomic units (OTUs) with 97% identity threshold. The saturation of curve indicates that the diversity of OTUs in the present study was mostly covered by the V4 and V9 sequences.

group (Supplementary Table 1). Only a small fraction of total reads (0.01% of V9 data and 0.05% of V4 data) was characterized in this group.

The Chao1 values ranged from 103 to 735.9 (average, 480.2) for the V9 region, and from 47 to 471 (average, 326.8) for the V4 region (Fig. 3). The observed OTUs were able to cover from 99.6% to 100% of the richness of the estimated species from Chao1 analysis. The Shannon diversity index represented the evenness of species that varied from 1.33 to 6.52 (average, 4.56) for the V9 region and from 3.42 to 6.23 (average, 4.99) for the V4 region (Fig. 3). The inverse Simpson index represents the probability of two randomly selected taxa belonging to the same species. In this study, the index varied from 0.32 to 0.97 (average, 0.79) for the V9 region and from 0.81 to 0.97 (average, 0.90) for the V4 region (Fig. 3). The Good's coverage provides how well the sample represents the environment, and the coverages in this study was from 0.99 to 1 for both regions (Fig. 3). This illustrates that the generated reads from Illumina sequencing estimated the completeness of eukaryotic diversity in the samples.

Comparative analyses of the V9 and V4 sequences data in field samples. In this study, according to a recent classification and nomenclature of eukaryotes[26], 94 and 70 different major sub-division groups were assigned from the V9 and V4 region of 18S rDNA sequences, respectively (total 102 separate groups, Supplementary Table 1). It appears that all sub-division groups described here belonging to the highest rank groups 'Ancyromonadida', 'Cryptista', 'CRuMs', 'Haptista', and 'Telonemia' could be recovered by both biomarkers V9 and V4 (Fig. 4 and Supplementary Table 1). Although a fraction of the major sub-division groups in total reads depended on the type of groups, except for the supergroup 'Opisthokonta', the major sub-division groups at the top rank in all supergroups were detected with both biomarkers. For instance, the sub-division group Thecofilosea in the supergroup 'Rhizaria' was 38.6% and 1.6% of the total reads in the V9 and V4 region sequences, respectively (Supplementary Table 1). Further, the sub-division group Centroplasthelida in the highest rank group 'Haptista' showed 86.1% and 0.8% of the total reads in the V9 and V4 region sequences, respectively (Supplementary Table 1). The unclassified subgroups showed usually a higher proportion of total reads at the supergroup or highest rank levels, regardless of the V9 and V4 regions. In the cases of 'Alveolata', 'Amoebozoa', 'Excavata', 'Rhizaria', 'Stramenopiles', and 'Haptista' all unclassified subgroups in the V4 region sequencing data appeared to a higher proportion than those in the V9 region sequencing data. On the bases of total V9 and V4

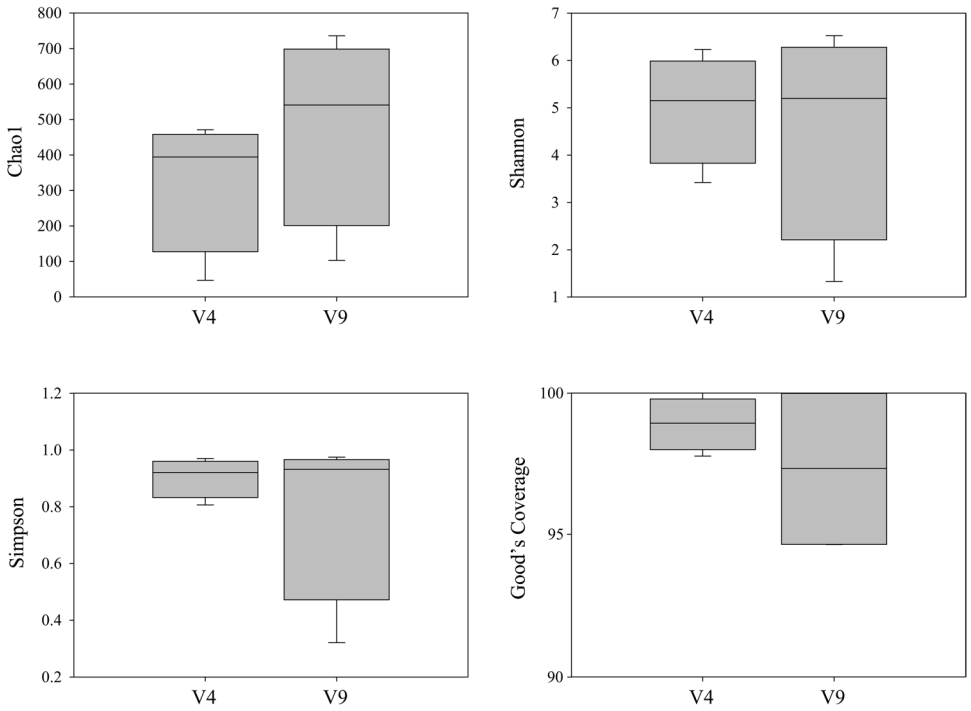

Figure 3. Alpha diversity indices for Chao1, Shannon diversity, inverse Simpson, and Good's coverage in V4 and V9 region datasets. Note that whiskers above and below the box plot (interquartile range) represent the 90th and 10th percentiles. The lines within the boxes indicate the medians.

reads, the most considerable difference between the two biomarkers represented the supergroups 'Excavata' and 'Amoebozoa'. Within the supergroup 'Excavata', the V9 region showed 87,300 reads in total, whereas the V4 region biomarker revealed only 329 reads (Supplementary Table 1). In the supergroup 'Amoebozoa', the relative numbers of the V9 and V4 region reads were 125,437 reads and 57,730 reads, respectively (Supplementary Table 1). Hence, it is hypothesized that the taxonomic affinities or reads intensely fluctuate depending on the choice of the primer sets.

Interestingly, the inventory of the major sub-division groups in the V9 region dataset did not correspond to that in the V4 region dataset. The major sub-division groups with almost $<1\%$ of total reads (rare subgroups) showed a substantial bias between V9 and V4 sequencing dataset, except 'Opisthokonta'. The classes Karyorelictea, Prostomatea, and Nassophorea in Ciliophora and the family Perkinsida belonging to the highest rank group 'Alveolata' were not detected using the V4 sequencing data, whereas they were detected using the V9 sequencing data (Fig. 4 and Supplementary Table 1). Furthermore, the classes Echinamoebida, Eumycetozoa, and Euamoebida in the supergroup 'Amoebozoa' and Chloropicophyceae, Pyramimonadales, and Mamiellophyceae in the supergroup 'Archaeplastida' were not detected in the V4 sequencing data (Fig. 4 and Supplementary Table 1).

In most NGS studies, the V4 or V9 region sequences from NGS data have been placed into an alignment of almost full-length 18S rDNA sequences retrieved from the reference databases due to difficulty of alignment[27]. Very few attempts have been made for the phylogenetic inference using short V4 or V9 region sequences so far. In the present study, we examined phylogenetic inference amongst the supergroups or major sub-division groups reported by Adl *et al.*[26] based on OTUs (Fig. 5). The molecular phylogenetic trees based on the V4 and V9 region sequences indicated that the eukaryotic supergroups formed a paraphyletic or polyphyletic group in our datasets (Fig. 5), confirming that current datasets have not resolved the previous phylogenetic relationships (i.e. a monophyletic group) at the highest-rank taxonomic group level. Furthermore, the length of the phylogenetic branch in the V4 region dataset was more varied than that in the V9 region dataset (e.g. supergroup 'Amoebozoa', Fig. 5). Thus, it seems that the V4 region might contain a more hypervariable region compared to the V9 region. Conversely, at the lower taxonomic level, some major sub-division groups including 2–27 OTUs represented a robust clade with $>90\%$ bootstrapping supports, which was depending on the dataset (Supplementary Table 1). In maximum likelihood analyses, a total 7 major sub-division groups formed a clade with 93–100% bootstrapping supports in the V9 region dataset (i.e. Echinamoebida, Heterolobosea, Granofilosea, Centroplasthelida, Peronosporomycetes, Phragmoplastophyta, and Telonema; Supplementary Table 1). In the V4 region dataset, a total of 7 major sub-division groups (Spirotrichea, Peronosporomycetes, Centroplasthelida, Enoplea, Chaetonotida, Pezizomycotina, and Rigifilida; Supplementary Table 1) formed a robust clade with 100% bootstrapping support. Classes Peronosporomycetes and Centroplasthelida showed a strong clade in both the V4 and V9 region datasets (Supplementary Table 1). Although the sub-division groups in the V9 region dataset showed more diversity than the V4 region dataset, most of the major sub-division groups could not reveal a previous monophyletic relationship in both the V4 and V9 region sequences using an Illumina platform.

Figure 4. The relative fraction of major eukaryotic groups belonging to the original supergroups 'Amoebozoa', 'Excavata', 'Rhizaria', 'Archaeplastida', and 'Opithokonta' or the highest rank groups 'Alveolata (Chromalveolata)', 'Cryptista (Chromalveolata)', 'Stramenopiles (Chromalveolata)', 'Haptista (Chromalveolata)', 'Ancyromonadida', 'CruMs', and 'Telonemia' detected in the present V4 and V9 datasets based on read count (also see Supplementary Table 1). Asterisk (*) indicates the original supergroups in eukaryotes.

Discussion

We studied the comparative analyses of V4 and V9 regions of 18S rDNA, which are widely used for next-generation sequencing, in the field surveys to address the following questions: Do V4 and V9 region sequencing data provide similar taxonomic profiles in the dominant or rare eukaryotes? How different are eukaryotes at the high-rank taxonomic group level? Overall, we strongly recommend that the V4 and V9 regions are used together to investigate the diversity and taxonomic assignment of eukaryotes in field surveys[13,28,29], and it is possible that some important eukaryotic groups may be unexplored or underestimated by the V4 or V9 region alone in the environmental samples, particularly for rare subgroups.

The two biomarkers V4 and V9 regions show a remarkable difference using the same Illumina platform within a total of approximately 24 L subsamples (6 L × 4 surveys in two years) of brackish water. On the bases of raw total bases and raw read counts, the numbers for the V4 region were relatively higher than those for the V9 region. However, eukaryotic OTUs for V9 region (i.e. 1,413 OTUs) are 20% more abundant than those for the V4 region

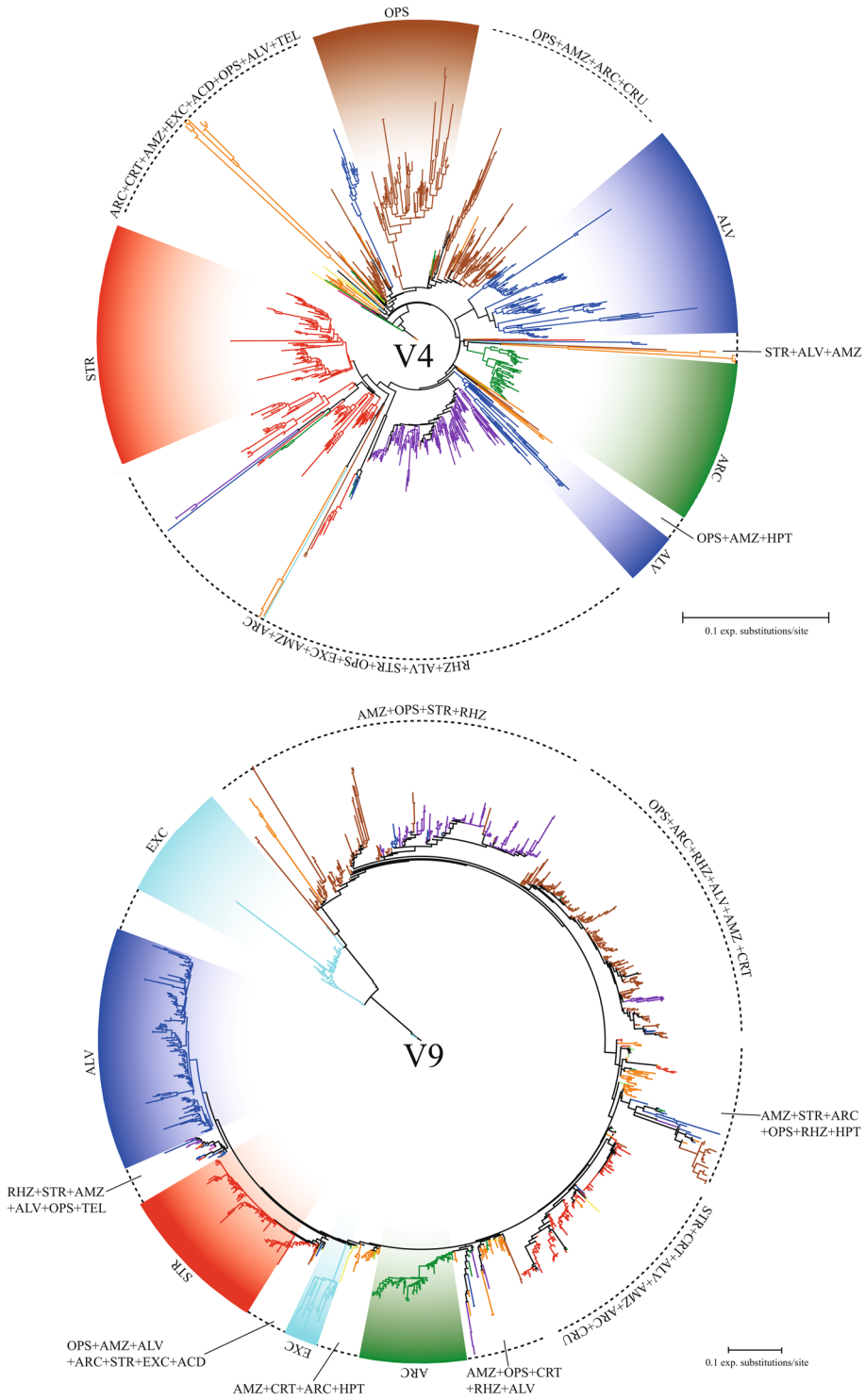

Figure 5. Unrooted maximum likelihood phylogenetic trees estimated from V4 (915 sequences total) and V9 (1,413 sequences total) region sequences of 18S rDNA using the Illumina MiSeq. Color of each node indicates a different supergroup or the highest rank group. OPS, Opisthokonta (brown); AMZ, Amoebozoa (orange); ALV, Alveolata (blue); STR, Stramenopiles (red); ARC, Archaeplastida (green); CRT, Cryptista (yellow); RHZ, Rhizaria (purple); EXC, Excavata (sky blue); HPT, Haptsita (light green); ACD, Ancyromonadida (gold); CRU, CRuMs (silver); and TEL, Telonemia (pink).

(i.e. 915 OTUs) at 97% identity threshold. The determination of the identity threshold is one of the critical steps affecting the number of examined OTUs[21,23]. At the higher identity threshold, the estimation of diversity in eukaryotes is more conservative[13,20]. Surprisingly, Tragin *et al.*[21] and Piredda *et al.*[8] demonstrated that the fraction of the

V9 region in total eukaryotic OTUs was 20% higher than that of the V4 region at 97% and 95% identity threshold, respectively, in the field samples (i.e. Atlantic Ocean, Pacific Ocean, and the Mediterranean Sea). Additionally, Maritz *et al.*[20] reported that the V9 region OTUs (i.e. 1,444 OTUs) were 27% higher than the V4 region OTUs (i.e. 824 OTUs) from sewage samples in New York City at 98% identity threshold. This result suggests that the number of examined OTUs for V9 region are often 20% more likely to be larger than that for the V4 region, regardless of the identity threshold with 95% or higher and sampling locations[21].

The sequencing error rates from ambiguous, low-quality, and chimera reads in the V4 region (2.7% of total read counts) are almost equivalent to those of the V9 region (3.2% of total read counts). Furthermore, prokaryotic OTUs (i.e. 207 OTUs of the total 1,122 OTUs) from the V4 region sequencing data are similar to those from the V9 region sequencing data (i.e. 219 OTUs of total 1,632 OTUs), implying that non-eukaryotic sequences are often detected in the V4 and V9 region datasets due to capability of the primer sets to amplify the small subunit rDNA of all three domains[30] (i.e. archaea, bacteria, and eukaryotes). However, low coverage and low identity reads with $<$85% in the V4 region dataset was much higher than those in the V9 region dataset, whereas the chimera sequences in the V4 region dataset were slightly lower than those in the V9 region dataset. This suggests that the artifactual sequences from the V4 region sequencing data are more abundant, except for the chimera sequences. Previous studies revealed that the longer sequence, such as those of the V4 region, was susceptible to sequencing bias[21,31,32]. Although our results confirm those of the previous studies that the numbers of the artifactual sequences vary depending on the V4 or V9 region, a proportion of the sequencing error rates in total read counts from the V4 and V9 regions are similar to each other (see above). Thus, it is likely that the sequencing error rates are considered to weakly impact the estimation of eukaryotic diversity, irrespective of the amplicon regions.

The Chao1 and Good's coverage indexes do not discriminate between V4 and V9 region datasets. Although the average values are not significantly different from each other (*t*-test, $p = 0.752$), the average Shannon diversity and inverse Simpson indexes for V4 region is slightly higher than those for the V9 region. Several previous studies reported that the Shannon diversity and inverse Simpson indexes for V4 region were lower than those for V9 region[20,21,33]. Conversely, Hirakata *et al.*[30] reported that the Shannon diversity and inverse Simpson indexes for V4 region were higher than those for the V9 region, consistent with our results. Probably, this difference may be due to the fraction of different OTUs, intra-individual polymorphism, and/or different sampling locations[18,21,30,34].

In the supergroups 'Amoebozoa', 'Excavata', and 'Archaeplastida' or highest rank groups 'Cryptista', 'Ancyromonadida', 'CRuMs', and 'Telonema', the most abundant sub-division groups in the V9 region dataset represent identical eukaryotes taxonomic profiles in the V4 region dataset. Other eukaryotic supergroups (i.e. 'Rhizaria' and 'Opisthokonta') or the highest rank groups (i.e. 'Alveolata', 'Stramenopiles', and 'Haptista') do not allow for the identical profiles in both the V4 and V9 region datasets. The second or third abundant sub-division groups in the V9 region taxonomic profiles displays the most abundant groups in the V4 region taxonomic profiles. Thus, according to the V4 or V9 region datasets, the order of the most abundant group in some eukaryotes taxonomic profiles can be altered, suggesting different affinity of the V4 or V9 region primer sets to the sub-division groups[8,32,33]. Further, it seems likely that a first- to third-ranked sub-division groups, which mostly occupied $>$50% of total cumulative read counts, are unchanged with respect to dominance in eukaryotes taxonomic profiles in both the V4 and V9 region datasets. Thus, it is unlikely that the taxonomic profiles in the predominant groups result from the primer bias[13].

The V4 region may often provide the missing information on the diversity of some important eukaryotic groups in comparison to the V9 region[8,9,21]. Our result indicates that the V4 region fails to detect some important eukaryotic groups (e.g. Heterolobosea) and underestimates the diversity of the major eukaryotic groups (e.g. Thecofilosea and Centroplastithelida). In addition, most protists supergroups (i.e. 'Chromalveolata', 'Amoebozoa', 'Excavata', and 'Rhizaria') show high numbers of the unclassified subgroups in the V4 region dataset. Salonen *et al.*[33] noted that the unclassified protistan subgroups in the V4 region dataset were much higher than those in the V9 region dataset, similar to our result. It is possible that the V4 region sequences may be deficient in the current database compared to the V9 region sequences[32]. Thus, prior to the completion of both the V4 and V9 region databases, the choice of primer set plays a crucial role in estimating the extant diversity of the target eukaryotes in field samples.

Due to the short NGS amplicons of the V4 and V9 regions, the molecular phylogenetic analyses shows that each supergroup represents a paraphyletic or polyphyletic group, indicating the lack of phylogenetic clustering at the high-rank taxonomic level. Adl *et al.*[26] reported that Echinamoebida, Heterolobosea, Granofilosea, Centroplasthelida, Peronosporomycetes, Phragmoplastophyta, Telonema, Spirotrichea, Enoplea, Chaetonotida, Rigifilida, and Pezizomycotina also tend to fall into a monophyletic group, consistent with our result from the Illumina MiSeq. Interestingly, the monophyletic clustering of the major sub-division groups from the V4 region sequences (i.e. 7 groups of total 70 major sub-division groups, 100% bootstrapping support) is somewhat reliable than the V9 region sequences (i.e. 7 groups of total 94 major sub-division groups, 93–100% bootstrapping supports). This result suggests that most of the major sub-division groups do not reveal a monophyletic clade, and numbers of monophyletic clades in V4 region dataset are similar to the V9 region dataset. Because all supergroups or most major sub-division groups failed to examine the monophyletic relationships across eukaryotes, a reliable approach should be developed for assessing the eukaryotic relationships on the Illumina MiSeq.

Since NGS technologies have been introduced, rare eukaryotic taxa can be recently accessed for their ecological roles in natural ecosystems[21]. Rare eukaryotic taxa play a crucial role as seeds for species succession or blooming and can be flexible to environmental changes[23,24]. In this study, we also detected rare taxa with $<$1% of total reads (i.e. rare subgroups) using either the V4 or V9 regions. However, patterns of rare taxa are varied depending on the V4 or V9 regions. The rare taxa in the supergroups 'Amoebozoa', 'Rhizaria', and 'Archaeplastida' or highest rank group 'Alveolata' remained unexplored using the V4 region but were detected using the V9 region. Thus, it seems that the V9 region reveals much better resolution for the detection of rare taxa in most protistan supergroups than the V4 region. Furthermore, Heterolobosea was revealed to be the second most abundant group

(38.6% of total read counts) in the supergroup 'Excavata' using the V9 region, but this group was not detected using the V4 region. Heterolobosea, including amoeba, amoeboflagellate, and flagellate forms, represents one of rare taxa in environmental surveys[17,30,35], but were commonly isolated from freshwater, marine, soil, and extreme environments[36–46]. Pawlowski *et al.*[17] reported that Heterolobosea comprises <1% of total reads using NGS technology with only the V9 region. Thus, the rare taxon Heterolobosea may be successfully detected using the V9 primer set, rather than the V4 primer set. However, our study is spatially limited for accessing the diversity of rare taxa in protists. Thus, further studies are needed at other locations or during other seasons.

In conclusion, the simultaneous application of V4 and V9 biomarkers in 18S rDNA is certainly advantageous for evaluating the diversity and phylogenetic relationship of the dominant or rare eukaryotic community in brackish water in comparison to the V4 or V9 region alone. Notably, the two biomarkers should complement each other for analysis of metabarcoding in the eukaryotic community. Thus, we strongly recommend the two biomarkers be widely used together in field surveys.

Methods

Sample collection. Surface water samples were collected from the Mulgol pond on Dokdo (island), Korea (37°14′22″N, 131°52′08″E) that is located in the East Sea/Sea of Japan, in August 2016, September 2017, and April and June 2018 (Fig. 1). Surface water samples were carefully taken with a sterile 1 L polycarbonate bottle to exclude any floating debris. Dokdo was designated as a Natural Monument and is maintained as a natural conservation district[47]. To maintain its natural ecosystem, only limited people are allowed to enter. Water in the Mulgol pond was historically used as drinking water for residents on Dokdo, but now it is no longer used for this purpose. This pond is reconstructed and now covered with a metal lid to protect from pollutants by a Korean government[48]. Dimensions of Mulgol are 1.4 m (width) × 1.2 m (length) × 1.7 m (depth). Because this pond has been covered with a metal lid, it has a limitation of light resources[45,48] (Fig. 1). The light intensity was measured by light meter TES-1332A (TES Electrical Electronic Corp., Taipei, Taiwan). Additionally, because of ambient seawater and wastes from abundant seabirds, this pond contains a high concentration of nitrogen and a little dissolved salt[45,48]. The concentration of chlorophyll-*a* was measured by a standard protocol as described by Parsons *et al.*[49], and water temperature and salinity were measured using a digital salinity/temperature meter (EUTECH Salt 6+, Thermo Fisher Scientific, Republic of Korea).

Environmental DNA Extraction. A total of 6 L of each water samples from the Mulgol were collected and filtered through 0.45 µm pore-sized Durapore membrane filters (Merck Millipore, Billerica, MA, USA) using a vacuum pump (model DOA-P704-AC, GAST, Benton Harbor, MI, USA) during the field periods. Probably, dissolved extracellular DNA passed through 0.45 µm pore-sized membrane filters[50]. These filters were stored in a 50 mL conical tube at −20 °C and taken to the laboratory for further experiments.

For environmental DNA extraction, 20% (w/v) lysozyme (final concentration, Sigma-Aldrich, St. Louis, MO, USA) were directly added to each conical tube after the filters were cut into several pieces. The tubes were then incubated at 37 °C for 30 min. Further, 0.5 mg mL^{-1} proteinase K (final concentration, Sigma-Aldrich) and 1% sodium dodecyl sulfate (final concentration, Bioneer, Daejeon, Korea) were added to the conical tube, and the tubes were incubated at 55 °C for 2 hrs. Nucleic acids were further purified using DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instruction. The concentration of extracted DNA was measured with Quantus fluorometer (Promega, Madison, WI, USA). The final concentrations of extracted DNA ranged from 0.68 to 14.19 ng μL^{-1}.

Illumina sequencing. The V4 and V9 regions of 18S rDNA were used for the Illumina sequencing. Primers V4 forward (5′-CCAGCAGCCGCGGTAATTCC-3′) and V4 reverse (5′-ACTTTCGTTCTTGATTAA-3′) were used to target the V4 variable region[13] of the 18S rDNA, and primers V9 forward (5′-CCCTGCCHTTTGTACACAC-3′) and V9 reverse (5′-CCTTCYGCAGGTTCACCTAC-3′) were used to target the V9 variable region[28]. Amplification conditions for V4 regions comprised initial denaturing step at 95 °C for 5 min, followed by 10 cycles of 94 °C for 30 s, 57 °C for 45 s, 72 °C for 1 min, and then, 15 cycles of 94 °C for 30 s, 47 °C for 45 s, 72 °C for 1 min, and ending at 72 °C for 10 min[29]. Reaction conditions for the V9 region comprised an initial denaturing step at 94 °C for 3 min. followed by 30 cycles of 94 °C for 30 s, 57 °C for 60 s, 72 °C for 90 s, and ending at 72 °C for 10 min[28]. Library size was confirmed by running on Agilent Technologies 2100 Bioanalyzer using a DNA 1000 chip (Aligent, Santa Clara, CA, USA). Library was quantified using a qPCR as described in the Illumina qPCR quantification protocol guide. A paired-end read was performed with the Illumina platform (i.e. Illumina MiSeq, Macrogen, Republic of Korea).

Bioinformatic analysis and phylogenetic analysis. Paired-end reads were merged from Illumina sequencing with Fast Length Adjustment of SHort reads 1.2.11 program (FLASH)[51]. Sequences were trimmed and filtered and clustered by using CD-HIT-OTU software (v.0.0.1 for Illumina rRNA data)[52]. Through this process, short reads were filtered and long sequences were trimmed. Using CD-HIT-DUP, filtered sequences were clustered with 97% identity threshold and assigned to operational taxonomic units (OTUs). Also, chimeras were filtered out and extra-long tails were trimmed. Taxonomic composition for each sequence from phylum to species was generated using QIIME UCLUST[53]. Reference data were used with 18S rDNA data in the National Center for Biotechnology Information. The alpha diversity (i.e. Chao1, Shannon diversity, inverse Simpson, and Good's coverage) analyses were conducted by QIIME pipeline[54]. Totally, 1,413 and 915 sequences obtained from the V9 and V4 region datasets, respectively, were aligned using MAFFT program version 7, and then were subsequently edited by eye[55]. The V4 and V9 region datasets retained only the 316 and 113 unambiguously aligned sites present in all partial sequences, respectively. Maximum likelihood trees were estimated using IQ tree on the version of IQ-TREE 1.6.12[56–58]. TIM3 + F + I + G4 and TIM2e + I + G4 models were selected through best-fit model test

option (-m TEST) for V4 and V9 regions, respectively[56–58]. Ultrafast bootstrapping with 1,000 replications (-bb 1,000) was performed to estimate the branch support. The detailed command for this analysis was as follows: iqtree -s V4.fasta -m TEST -bb 1000 -nt AUTO, iqtree -s V9.fasta -m TEST -bb 1000 -nt AUTO. Analysis of *t*-test was performed using SPSS for Windows (version 25, SPSS Inc.).

According to the eukaryotic classification and nomenclature reported by Adl *et al.*[26,59,60], the original supergroups in this study were conventionally assigned as 'Amoebozoa', 'Archaeplastida', 'Chromalveolata', 'Excavata', 'Opisthokonta', and 'Rhizaria'. Furthermore, 'Chromalveolata' was divided into 'Alveolata', 'Cryptista', 'Haptista', and 'Stramenopiles' as the highest rank group. 'Ancyromonadida', 'Telonemia' and 'CRuMs (i.e. Collodictyonidae, Rigifilida, and Mantamonas)' were regarded as the other highest rank groups.

Data availability

The V4 (SRR10539012-SRR10539015) and V9 (SRR10539016-SRR10539019) sequence data were deposited in the NCBI Sequence Read Archive (SRA) under accession numbers between SRR10539012 and SRR10539019, and corresponding sample descriptions are accessible through BioProject PRJNA592034.

Received: 11 November 2019; Accepted: 1 April 2020;
Published online: 16 April 2020

References

1. Caporaso, J. G. *et al*. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**, 1621–1624 (2012).
2. Sogin, M. L. *et al*. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* **103**, 12115–12120 (2006).
3. Degnan, P. H. & Ochman, H. Illumina-based analysis of microbial community diversity. *ISME J* **6**, 183–194 (2012).
4. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30**, 418–426 (2014).
5. Bartram, A. K., Lynch, M. D., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl Environ Microbiol* **77**, 3846–3852 (2011).
6. Logares, R. *et al*. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* **16**, 2659–2671 (2014).
7. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**, e30087 (2012).
8. Piredda, R. *et al*. Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean long term ecological research site. *FEMS Microbiol Ecol* **93**, fiw200 (2016).
9. Bradley, I. M., Pinto, A. J. & Guest, J. S. Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl Environ Microbiol* **82**, 5878–5891 (2016).
10. Harder, C. B. *et al*. Local diversity of heathland Cercozoa explored by in-depth sequencing. *ISME J* **10**, 2488–2497 (2016).
11. Taib, N., Mangot, J. F., Domaizon, I., Bronner, G. & Debroas, D. Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. *PLoS One* **8**, e58950 (2013).
12. Hadziavdic, K. *et al*. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One* **9**, e87624 (2014).
13. Stoeck, T. *et al*. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**, 21–31 (2010).
14. Hu, S. K. *et al*. Estimating protistan diversity using high-throughput sequencing. *J Eukaryot Microbiol* **62**, 688–693 (2015).
15. Aguilar, M. *et al*. Next-generation sequencing assessment of eukaryotic diversity in oil sands tailings ponds sediments and surface water. *J Eukaryot Microbiol* **63**, 732–743 (2016).
16. Ferrera, I. *et al*. Evaluation of alternative high-throughput sequencing methodologies for the monitoring of marine picoplanktonic biodiversity based on rRNA gene amplicons. *Front Mar Sci* **3**, 147 (2016).
17. Pawlowski, J. *et al*. Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One* **6**, e18169 (2011).
18. Decelle, J., Romac, S., Sasaki, E., Not, F. & Mahe, F. Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLoS One* **9**, e104297 (2014).
19. Dunthorn, M., Klier, J., Bunge, J. & Stoeck, T. Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Eukaryot Microbiol* **59**, 185–187 (2012).
20. Maritz, J. M. *et al*. An 18S rRNA workflow for characterizing protists in sewage, with a focus on zoonotic trichomonads. *Microb Ecol* **74**, 923–936 (2017).
21. Tragin, M., Zingone, A. & Vaulot, D. Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. *Environ Microbiol* **20**, 506–520 (2018).
22. Kim, E. *et al*. Oligotrophic lagoons of the South Pacific Ocean are home to a surprising number of novel eukaryotic microorganisms. *Environ Microbiol* **18**, 4549–4563 (2016).
23. Caron, D. A. & Countway, P. D. Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat Microb Ecol* **57**, 227–238 (2009).
24. Logares, R. *et al*. Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* **24**, 813–821 (2014).
25. Massana, R. *et al*. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17**, 4035–4049 (2015).
26. Adl, S. M. *et al*. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukaryot Microbiol* **66**, 4–119 (2019).
27. Dunthorn, M. *et al*. Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol Biol Evol* **31**, 993–1009 (2014).
28. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4**, e6372 (2009).
29. Stoeck, T., Hayward, B., Taylor, G. T., Varela, R. & Epstein, S. S. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* **157**, 31–43 (2006).
30. Hirakata, Y. *et al*. Temporal variation of eukaryotic community structures in UASB reactor treating domestic sewage as revealed by 18S rRNA gene sequencing. *Sci Rep* **9**, 1–11 (2019).
31. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
32. Monier, A., Worden, A. Z. & Richards, T. A. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ Microbiol Rep* **8**, 461–469 (2016).

33. Salonen, I. S., Chronopoulou, P. M., Leskinen, E. & Koho, K. A. Metabarcoding successfully tracks temporal changes in eukaryotic communities in coastal sediments. *FEMS Microbiol Ecol* **95**, fiy226 (2018).
34. Giner, C. R. *et al*. Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes. *Appl Environ Microbiol* **82**, 4757–4766 (2016).
35. López-García, P., Vereshchaka, A. & Moreira, D. Eukaryotic diversity associated with carbonates and fluid–seawater interface in Lost City hydrothermal field. *Environ Microbiol* **9**, 546–554 (2007).
36. Hanousková, P., Táborský, P. & Čepička, I. *Dactylomonas* gen. nov., a novel lineage of heterolobosean flagellates with unique ultrastructure, closely related to the amoeba *Selenaion koniopes* Park, De Jonckheere & Simpson, 2012. *J Eukaryot Microbiol* **66**, 120–139 (2019).
37. Harding, T. *et al*. Amoeba stages in the deepest branching heteroloboseans, including
38. Page, F. C. Taxonomic criteria for limax amoebae, with descriptions of 3 new species of *Hartmannella* and 3 of *Vahlkampfia*. *J Protozool* **14**, 499–521 (1967).
39. Page, F. C. *A new key to freshwater and soil gymnamoebae with instructions for culture* (ed. Page, F. C.) 31–47 (Freshwater Biological Association 1988).
40. Park, J. S., Simpson, A. G., Lee, W. J. & Cho, B. C. Ultrastructure and phylogenetic placement within Heterolobosea of the previously unclassified, extremely halophilic heterotrophic flagellate *Pleurostomum flabellatum* (Ruinen 1938). *Protist* **158**, 397–413 (2007).
41. Park, J. S., Simpson, A. G., Brown, S. & Cho, B. C. Ultrastructure and molecular phylogeny of two heterolobosean amoebae, *Euplaesiobystra hypersalinica* gen. et sp. nov. and *Tulamoeba peronaphora* gen. et sp. nov., isolated from an extremely hypersaline habitat. *Protist* **160**, 265–283 (2009).
42. Park, J. S., De Jonckheere, J. F. & Simpson, A. G. Characterization of *Selenaion koniopes* n. gen., n. sp., an amoeba that represents a new major lineage within Heterolobosea, isolated from the Wieliczka Salt Mine. *J Eukaryot Microbiol* **59**, 601–613 (2012).
43. Pánek, T., & Čepička, I. Diversity of heterolobosea In *Genetic diversity in microorganisms* (ed. Caliskan, M.) 3–26 (Intech 2012).
44. Pánek, T., Simpson, A. G., Hampl, V. & Čepička, I. *Creneis carolina* gen. et sp. nov. (Heterolobosea), a novel marine anaerobic protist with strikingly derived morphology and life cycle. *Protist* **165**, 542–567 (2014).
45. Park, J. S. A new heterolobosean amoeboflagellate, *Tetramitus dokdoensis* n. sp., isolated from a freshwater pond on Dokdo Island in the East Sea, Korea. *J Eukaryot Microbiol* **64**, 771–778 (2017).
46. Tikhonenkov, D. V. *et al*. Ecological and evolutionary patterns in the enigmatic protist genus *Percolomonas* (Heterolobosea; Discoba) from diverse habitats. *PLoS One* **14**, e0216188 (2019).
47. Jung, S. Y. *et al*. The study of distribution characteristics of vascular and naturalized plants in Dokdo, South Korea. *J Asia Pac Biodivers* **7**, e197–e205 (2014).
48. Park, J. S. First record of potentially pathogenic amoeba *Vermamoeba vermiformis* (Lobosea: Gymnamoebia) isolated from a freshwater of Dokdo Island in the East Sea, Korea. *Anim Syst Evol Divers* **32**, 1–8 (2016).
49. Parsons, T. R., Maita, Y., & Lalli, C. M. Determination of chlorophylls and total carotenoids: spectrophotometric method. In *A manual of chemical and biological methods for seawater analysis* 101–112 (Pergamon Press 1984).
50. Sørensen, N., Daugbjerg, N. & Richardson, K. Choice of pore size can introduce artefacts when filtering picoeukaryotes for molecular biodiversity studies. *Microb Ecol* **65**, 964–968 (2013).
51. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
52. Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* **13**, 656–668 (2012).
53. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
54. Caporaso, J. G. *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335–336 (2010).
55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
56. Flouri, T. *et al*. The phylogenetic likelihood library. *Syst biol* **64**, 356–362 (2014).
57. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188–1195 (2013).
58. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2014).
59. Adl, S. M. *et al*. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* **52**, 399–451 (2005).
60. Adl, S. M. *et al*. The revised classification of eukaryotes. *J Eukaryot Microbiol* **59**, 429–514 (2012).

Acknowledgements

We thank the Research Institute for Dok-do and Ulleung-do Island (PI: Prof. Jae-Hong Pak) at Kyungpook National University for collecting samples from Dokdo (island) in the East Sea, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant, funded by the Korea government (MSIT) (No. 2016R1A6A1A05011910, 2017K2A9A1A06049946, and 2019R1A2C2002379) to J.S.P.

Author contributions

J.S.P. designed and supervised the present study. J.C. performed the sequencing preparation and analyses. J.C. prepared the figures and tables, and all authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-63561-z.

Correspondence and requests for materials should be addressed to J.S.P.

Reprints and permissions information is available at www.nature.com/reprints.

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