The Journal of
Eukaryotic Microbiology

Published by the International Society of Protistologists

Journal of Eukaryotic Microbiology ISSN 1066-5234

SHORT COMMUNICATION

Estimating Protistan Diversity Using High-Throughput Sequencing

Sarah K. Hu[a], Zhenfeng Liu[a], Alle A. Y. Lie[a], Peter D. Countway[b], Diane Y. Kim[a], Adriane C. Jones[c], Rebecca J. Gast[d], S. Craig Cary[e,f], Evelyn B. Sherr[g], Barry F. Sherr[g] & David A. Caron[a]

a Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA
b Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544, USA
c Mount St. Mary's College, Los Angeles, California 90049, USA
d Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
e Environmental Research Institute, School of Science, University of Waikato, Hamilton 3240, New Zealand
f College of Earth and Ocean Science, University of Delaware, Newark, Delaware 19716, USA
g College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA

Keywords
18S rRNA gene; biogeography; microbial eukaryotes; protists; V4 tag sequencing.

Correspondence
S. Hu, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-0371, USA
Telephone number: (213) 821-1800;
FAX number: (213) 740-8123;
e-mail: sarah.hu@usc.edu

Received: 7 November 2014; revised 17 January 2015; accepted January 26, 2015.

doi:10.1111/jeu.12217

ABSTRACT

Sequencing hypervariable regions from the 18S rRNA gene is commonly employed to characterize protistan biodiversity, yet there are concerns that short reads do not provide the same taxonomic resolution as full-length sequences. A total of 7,432 full-length sequences were used to perform an in silico analysis of how sequences of various lengths and target regions impact downstream ecological interpretations. Sequences that were longer than 400 nucleotides and included the V4 hypervariable region generated results similar to those derived from full-length 18S rRNA gene sequences. Present high-throughput sequencing capabilities are approaching protistan diversity estimation comparable to whole gene sequences.

HIGH-THROUGHPUT sequencing (HTS) of hypervariable regions within the small subunit 18S ribosomal RNA gene provides researchers the means to delve deeply into the species richness of natural protistan communities (Grattepanche et al. 2014; Mahé et al. 2014). Nevertheless, read lengths enabled by these methods are still relatively short (ca. 100–400 nt) and the taxonomic and phylogenetic resolution afforded by them can be limited. As a consequence, we presently have a relatively poor understanding of how 18S rRNA gene sequences of various lengths, and specific regions targeted for sequencing, impact downstream ecological interpretations (Hadziavdic et al. 2014; Hugerth et al. 2014; Stoeck et al. 2010).

Previous work comparing sequencing results from full-length 18S rRNA genes with hypervariable regions of the 18S gene has demonstrated inconsistencies in the conclusions gleaned from these datasets regarding species richness and diversity. For example, full-length 18S rRNA sequences of natural protistan communities and short sequences of the V4 or V9 regions led to different conclusions on the composition of abundant taxa (Wolf et al. 2014) and the level of taxonomic detail (Stoeck et al. 2010). Discrepancies between hypervariable region results are related, in part, to methodological problems of sequencing errors, PCR bias, and differences among applications used to call operational taxonomic units (OTUs). Inconsistencies in results may also be attributed to differing rates of evolution among hypervariable regions. These issues affect the number of OTUs measured and therefore the estimated species richness in natural samples (Decelle et al. 2014; Dunthorn et al. 2012; Hugerth et al. 2014; Kim et al. 2011; Wolf et al. 2014; Youssef et al. 2009).

One approach to eliminate PCR and sequencing artifacts from sequence diversity comparisons is to bioinformatically extract the shorter fragments of interest from longer sequences. Analysis of in silico extractions of short sequence reads from full-length or nearly full-length 18S rRNA sequences have recently been reported in the literature (Hadziavdic et al. 2014; Hugerth et al. 2014). These

studies concluded that hypervariable regions within the 18S rRNA gene did not yield similar estimates of protistan biodiversity to each other or to full-length sequences. The choice of the hypervariable region played a greater role in influencing interpretations of microbial biodiversity than the choice of sequencing technology and sequence length (Hadziavdic et al. 2014; Hugerth et al. 2014).

In this study, we compared diversity analyses derived from 7,432 Sanger-sequenced, full-length 18S rRNA genes obtained from natural microbial eukaryotic communities dominated by protists, with several bioinformatically extracted regions from the same full-length sequences. Full-length and nearly full-length sequences served to ground truth the ecological interpretations attained from in silico extracted sequence fragments between 107 and 1,200 nucleotides (nt). Short sequence fragments included the V7, V4, V1-V3, V1-V4, V4-V7, and V1-V7 regions (Table S1) determined by 18S rRNA gene specific primers previously used in protistan studies (Hadziavdic et al. 2014; Medlin et al. 1988; Stoeck et al. 2010; Weekers et al. 1994). OTUs clustered at 97%, 98%, and 99% sequence similarity from full-length and short fragment sequence datasets revealed that short reads (< 400 nt) predicted fewer OTUs and yielded lower values for two commonly employed diversity indices than the full-length sequences. Nevertheless, sequences ≥ 400 nt yielded ecological interpretations similar to full-length sequences, especially when the extracted sequence included a taxonomically informative region such as the V4 hypervariable region.

MATERIALS AND METHODS

Sample collection and processing

Samples from a global survey of natural microbial eukaryotic communities (Lie et al. 2014) were collected for DNA extraction and sequencing (Countway et al. 2010). A total of 12 seawater samples were obtained from depths ranging from 5 m to 2,500 m at five locations (Table S2): Arctic Ocean (AO), San Pedro Ocean time series station in the eastern North Pacific (ENP), East Pacific Rise (EPR) in the eastern Pacific, Gulf Stream (GS) in the western North Atlantic, and Ross Sea, Antarctica (RS). Water was collected using Niskin bottles mounted on a rosette, and 2–20 liters of water was prefiltered through 200-μm Nitex mesh to exclude most multicellular organisms (samples from ENP were prescreened with both 200 and 80-μm Nitex mesh). Microbial biomass was collected onto GF/F filters (Whatman™, International Ltd., Florham Park, NJ), rolled and placed into vials containing 2-ml of 2X lysis buffer (100-mM Tris pH 8, 40-mM EDTA pH 8, 100-mM NaCl, 1% SDS), and stored frozen (−20 °C) aboard the ship or flash frozen in liquid nitrogen for later DNA extraction.

Genomic DNA was extracted by thawing filters on a 70 °C heating block followed by three rounds of bead-beating on a vortex mixer (3 min) and heating at 70 °C (3 min) before extraction using phenol–chloroform as outlined by Countway et al. (2007). Extracted DNA was precipitated, dried, and then resuspended in sterile water. Genomic DNA was PCR-amplified to enrich for 18S rRNA genes using universal eukaryotic primers Euk-A and Euk-B (Table S1, Medlin et al. 1988). PCR amplifications were performed as described by Countway et al. (2010). The PCR thermal protocol consisted of a single cycle at 95 °C for 2 min, 35 cycles of 95 °C for 30 s, 50 °C for 30 s, 72 °C for 2.5 min, and a final elongation step at 72 °C for 7 min.

Cloning procedures were performed as described in Countway et al. (2010). 18S rDNA products were run on a 1.2% SeaKem agarose gel, bands were excised and PCR amplicons were cloned with the TOPO-TA kit (Invitrogen, Carlsbad, CA) using the TOP-10 electrocompetent cells (Countway et al. 2007, 2010). Initial plating was done to ensure successful transformation and to estimate cloning efficiency. Ten percent glycerol stocks were prepared for shipment to the Joint Genome Institute (JGI, http://www.jgi.doe.gov/) for sequencing.

Sequencing and data analysis

Clones from glycerol stocks were plated and discrete colonies were robotically picked at JGI after overnight growth. Sanger sequencing was conducted by JGI on an ABI 3730 capillary DNA sequencer (Applied Biosystems, Foster City, CA). A total of 1,152 clones were sequenced for each sample using T3 and V7 vector primers (Sambrook et al. 1989) and the 570-F internal primer (Weekers et al. 1994). Assembly was performed by JGI using the program Phrap v0.990319. The resulting full-length and nearly full-length 18S rRNA gene sequences had an average of 1,647 nt in length.

Sequences were screened for chimeras using a local implementation of the Pintail algorithm (Ashelford et al. 2005). Each sequence was searched with BLASTN v2.2.25+ (Altschul et al. 1990) against a database consisting of the eukaryotic SSU sequences from SILVA (Pruesse et al. 2007) to calculate a deviation of expected (DE) value (Ashelford et al. 2005). Sequences with a DE value greater than five were identified as possible chimeras and removed (6% of the sequences). Sequences were submitted to GenBank with accession numbers: KJ757035–KJ764638 (Lie et al. 2014).

A global alignment using 7,432 of the sequences was performed using Geneious v6.1.6 (http://www.geneious.com, Kearse et al. 2012), followed by in silico isolation of the short sequence fragments. A virtual "bioinformatic PCR" was performed in Geneious using forward and reverse primers for V7, V4, V1-V3, V1-V4, V4-V7, and V1-V7 regions with an allowance of 3 mismatches per primer (indels not considered, Table S1). Targeted regions were chosen based on 18S primers previously employed to evaluate microbial eukaryotic biodiversity (Hadziavdic et al. 2014; Medlin et al. 1988; Stoeck et al. 2010; Weekers et al. 1994). Each region was then isolated in silico from all of the 7,432 sequences based on the location of the primer hit to a position on the aligned full-length 18S rRNA gene sequences. Sequence fragments used in downstream analysis included the forward and

reverse primers. These fragments and the original full-length sequences were then processed in mothur v1.33.1 (Schloss et al. 2009) to call OTUs and calculate diversity indices. Distances between sequences were calculated using the furthest neighbor algorithm at 97%, 98%, and 99% sequence similarity. Representative OTUs from each sequence dataset were blasted against the SILVA database to assign taxonomic identities based on an *e*-value threshold of $1e^{-}10$. In the case of more than one equal blast hit, the first one provided in the BLAST output was chosen for taxonomic assignment.

Choice of primers

Primers chosen for this study originated from previously published work that targeted natural protistan communities (Table S1). The effectiveness of these 18S rRNA gene primers has been evaluated for a number of protistan lineages (Dunthorn et al. 2012; Ki 2012), although there are taxa for which these primers are ineffective due to variances in the length of the 18S, such as foraminifera (Pawlowski and Lecroq 2010).

V1-V3 and V4 primers reflect commonly used primers, including Euk-A from Medlin et al. (1988), 570-R from Weekers et al. (1994) (which is approximately the end of the V3 region), and V4 primers from Stoeck et al. (2010). The V1-V3 and V4 regions examined in this study overlapped by 94 nt (including primers, Table S1).

V9 hypervariable region

The V9 region was not included in our analysis because complete Euk-B reverse primer sequences were not obtained for many of the full-length sequences. Only approximately 2,000 sequences had suitable V9 regions (intact Euk-B reverse primer), significantly less than the original 7,432 full-length and nearly full-length 18S rRNA gene sequences.

Diversity and community similarity of protistan assemblages

Inverse Simpson and Shannon indices were calculated to provide a measurement of diversity based on the abundance and evenness of OTUs generated from extracted fragments and full-length sequences (Schloss et al. 2009). Community composition was dominated by protists in each sample. Community composition was compared using the Bray–Curtis similarity parameter. OTU abundances for each dataset were square-root transformed before calculating Bray-Curtis similarity values. Bray-Curtis similarity matrices for each dataset were used to cluster samples for non-metric multidimensional scaling plots based on the group average (Primer-E v6, Clark and Warwick 2001).

RESULTS AND DISCUSSION

Assessing sequence diversity of rRNA genes is commonly used to investigate the structure and composition of natural protistan communities, but our understanding of the ecological information contained in sequence data is still evolving. We show here that the ecological inference of protistan species richness and diversity obtained from 18S rRNA gene sequences varied as a factor of read length and the inclusion of the V4 hypervariable region (Fig. 1). Previous work, specifically in regards to ciliates, has shown that in comparison to the shorter V9 hypervariable region, the V4 region was better for resolution of taxonomies (Dunthorn et al. 2012) and phylogenetic placement (Dunthorn et al. 2014). Use of the V4 region for protistan biodiversity studies has been shown to be superior to

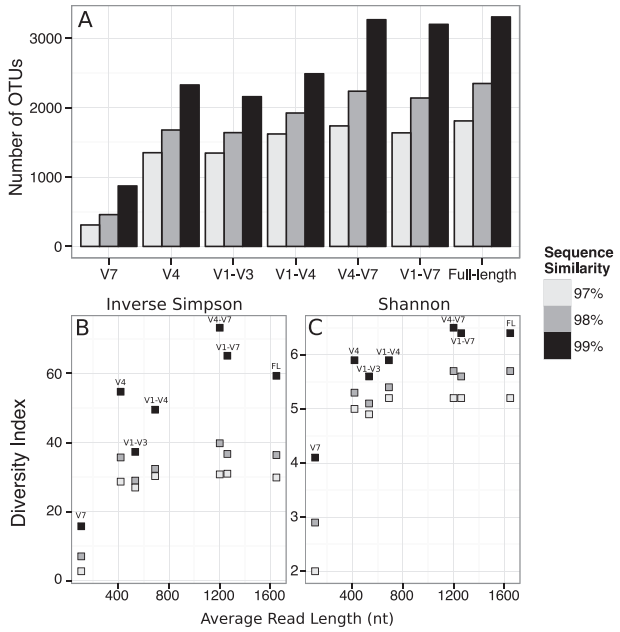

Figure 1 Total number of observed OTUs and diversity estimates for full-length and in silico extracted regions of the small subunit 18S ribosomal RNA gene. **A**. Number of operational taxonomic units (OTUs) obtained using full-length sequences or gene fragments of different lengths (Table S1). Sequences were clustered at 97%, 98% or 99% sequence similarity in mothur v. 1.33.1 (Schloss et al. 2009). Full-length, V1-V7, and V4-V7 (> 1,200 nt) sequence datasets yielded the largest number of OTUs (average of 1,680 OTUs at 97% sequence similarity) compared to the V4, V1-V4, and V1-V3 (400–700 nt) sequence results (average of 1,430 OTUs at 97% sequence similarity) and the short V7 (107 nt) sequence results (300 OTUs at 97% sequence similarity). OTUs from various sequence fragments were proportionally the same for 97%, 98%, or 99% sequence similarities. **B.** Inverse Simpson diversity index and **C.** Shannon diversity index calculated in mothur v. 1.33.1 (Schloss et al. 2009) using the OTU information presented in (A) as a function of sequence length (nt). Sequences longer than 1,200 nt (V4-V7, V1-V7) yielded inverse Simpson and Shannon values that were comparable to values for the full-length sequences. Values for the V4, V1-V3, and V1-V4 fragments were marginally lower than the three longer sequences for both indices, while values for the V7 region were markedly lower. Results from the various gene fragments in (B) and (C) are in the same order as (A), depicted left to right in order of average read length (Table S1): V7 (107 nt), V4 (418 nt), V1-V3 (533 nt), V1-V4 (690 nt), V4-V7 (1,200 nt), V1-V7 (1,260 nt), and full length (1,647 nt).

other 18S hypervariable regions, as the entire length (ca. 400 nt) is easily obtainable with current paired end HTS methods, namely Illumina MiSeq and HiSeq (Mahé et al. 2014).

The number of OTUs in this study generated from in silico extracted sequences was directly related to sequence length (Fig. 1A), while both read length and inclusion of the V4 region appeared to influence diversity indices (Fig. 1B,C). In particular, read lengths between 400 and 600 bases recovered between 300 and 900 fewer OTUs compared to full-length or nearly full-length 18S gene sequences (> 1,200 nt). Analysis of the shortest region (V7) yielded only a small fraction of the species richness obtained using longer reads. Results generated using different levels of sequence similarity to form OTUs yielded different absolute numbers of OTUs, as expected, but relative changes in OTUs between the various fragment lengths were consistent across the dataset (different shading in Fig. 1A).

Values for inverse Simpson and Shannon diversity indices reflected the trends in total number of OTUs generated. Diversity indices for the V7 region were several-fold lower than values obtained using longer reads, and read lengths greater than 400 nt were only moderately lower than values obtained using full-length 18S rRNA gene sequences (Fig. 1B,C). The V4-V7 and V1-V7 reads had diversity indices that were virtually indistinguishable from those based on full-length reads (Fig. 1B,C), in accordance with findings that the V4 region is taxonomically informative (Dunthorn et al. 2012; Hugerth et al. 2014; Mahé et al. 2014; Nickrent and Sargent 1991).

A nonmetric multidimensional scaling plot of Bray–Curtis dissimilarity values from the short V7 region was unable to resolve more than a few differences in protistan community structure among the 12 globally distributed natural samples analyzed in this study (Fig. S1 and Table S2). In contrast, data sets derived from DNA fragments with lengths $\geq$ 400 nt all yielded patterns that were relatively similar. There were only small inconsistencies among patterns obtained using longer fragments of the rRNA genes and the full-length sequences, a finding that paralleled results from the diversity indices (Fig. 1B,C, S1).

Pernice et al. (2013) recognized the usefulness of establishing a reference database of well-curated, relatively long rRNA gene sequences for vetting the many shorter sequences presently produced by HTS. Using a similar approach, we used in silico extraction of shorter reads from the same dataset of full-length sequences and examined the consistency of the ecological interpretations derived from DNA fragments of various lengths and target regions. Our study systematically confirms that longer ($\geq$ 400 nt) sequences enabled by HTS technologies provide consistent ecological information on protistan richness and diversity, relative to full-length sequences of 18S rRNA genes. This result is comforting in that it supports the application of present-day HTS in protistan ecology (Mahé et al. 2014), although it does not necessarily capture the total eukaryotic diversity present in natural ecosystems. This study did not address the more fundamental limitations of using a single gene for investigating the enormous breadth of microbial eukaryote diversity, such as inconsistencies between the morphospecies and genetic species concepts, or the potential variable rates of mutations of rRNA genes among protistan lineages (Caron 2013). Given this inherent limitation, this study used full-length 18S rRNA environmental sequences to ground truth the use of smaller gene fragments. DNA sequence information will continue to augment our understanding of microbial diversity, and future efforts to incorporate sequences from multiple genes into these approaches will improve our ability to capture true protistan diversity in the environment (Caron 2013; Stoeck et al. 2008).

ACKNOWLEDGMENTS

This study was supported by grants from the National Science Foundation (OCE-0550829, MCB-0703159, MCB-0084231, OCE-1136818) and the Gordon and Betty Moore Foundation (1112). Sequencing was conducted by the U.S. Department of Energy Joint Genome Institute and supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Sequence assembly and processing at JGI was performed by Ed Kirton, Jim Bristow, and Sussannah Tringe. William C. Nelson assisted with early bioinformatic analysis of the full-length sequence data. The authors acknowledge Ramon Terrado for helpful comments on the manuscript.

LITERATURE CITED

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, 71:7724–7736.

Caron, D. A. 2013. Towards a molecular taxonomy for protists: benefits, risks and applications in plankton ecology. *J. Eukaryot. Microbiol.*, 60:407–413.

Clark, K. R. & Warwick, R. M. 2001. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation. Primer-E, Plymouth, UK.

Countway, P. D., Gast, R. J., Dennett, M. R., Savai, P., Rose, J. M. & Caron, D. A. 2007. Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ. Microbiol.*, 9:1219–1232.

Countway, P. D., Vigil, P. D., Schnetzer, A., Moorthi, S. D. & Caron, D. A. 2010. Seasonal analysis of protistan community structure and diversity at the USC Microbial Observatory (San Pedro Channel, North Pacific Ocean). *Limnol. Oceanogr.*, 55:2381–2396.

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

Dunthorn, M., Klier, J., Bunge, J. & Stoeck, T. 2012. Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA

for assessment of ciliate environmental diversity. *J. Eukaryot. Microbiol.*, 59:185–187.
Dunthorn, M., Otto, J., Berger, S. A., Stamatakis, A., Mahé, F., Romac, S., deVargas, C. & Audic, S., BioMarks Consortium, Stock, A., Kauff, F. & Stoeck, T. 2014. Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol. Biol. Evol.*, 31:993–1009.
Grattepanche, J.-D., Santoferrara, L. F., McManus, G. B. & Katz, L. A. 2014. Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends Microbiol.*, 22:432–437.
Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M. & Troedsson, C. 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS ONE*, 9:e87624.
Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P. & Andersson, A. F. 2014. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS ONE*, 9:e95567.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28:1647–1649.
Ki, J.-S. 2012. Hypervariable regions (V1–V9) of the dinoflagellate 18S rRNA using a large dataset for marker considerations. *J. Appl. Phycol.*, 24:1035–1043.
Kim, M., Morrison, M. & Yu, Z. 2011. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J. Microbiol. Methods*, 84:81–87.
Lie, A. A. Y., Liu, Z., Hu, S. K., Jones, A. C., Kim, D. Y., Countway, P. D., Amaral-Zettler, L., Cary, S. C., Sherr, E. B., Sherr, B. F., Gast, R. J. & Caron, D. A. 2014. Investigating microbial eukaryotic diversity from a global census: insights from a comparison of pyrotag and full-length sequences of 18S rRNA gene sequences. *Appl. Environ. Microbiol.*, 80:4363–4373.
Mahé, F., Mayor, J., Bunge, J., Chi, J., Siemensmeyer, T., Stoeck, T., Wahl, B., Paprotka, T., Filker, S. & Dunthorn, M. 2014. Comparing high-throughput platforms for sequencing the V4 Region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. *J. Eukaryot. Microbiol.* doi:10.1111/jeu.12187.
Medlin, L., Elwood, H. J., Stickel, S. & Sogin, M. L. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, 71:491–499.
Nickrent, D. L. & Sargent, M. L. 1991. An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic Acids Res.*, 19:227–235.
Pawlowski, J. & Lecroq, B. 2010. Short rDNA barcodes for species identification in foraminifera. *J. Eukaryot. Microbiol.*, 57:197–205.
Pernice, M. C., Logares, R., Guillou, L. & Massana, R. 2013. General patterns of diversity in major marine microeukaryote lineages. *PLoS ONE*, 8:e57170.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. & Glöckner, 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.
Sambrook, J., Fritsch, E. F. & Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
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., Sahl, J. W., Stres, B., Thallinger, G. G., van Horn, D. J. & Weber, C. F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75:7537–7541.
Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D., Breiner, H. W. & Richards, T. A. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.*, 19 (Suppl. 1):21–31.
Stoeck, T., Jost, S. & Boenigk, J. 2008. Multigene phylogenies of clonal *Spumella*-like strains, a cryptic heterotrophic nanoflagellate, isolated from different geographical regions. *Int. J. Syst. Evol. Microbiol.*, 58:716–724.
Weekers, P. H. H., Gast, R. J., Fuerst, P. A. & Byers, T. J. 1994. Sequence variations in small-subunit ribosomal RNAs of *Hartmannella vermiformis* and their phylogenetic implications. *Mol. Biol. Evol.*, 11:684–690.
Wolf, C., Silva Kilias, E. & Metfies, K. 2014. Evaluating the potential of 18S rDNA clone libraries to complement pyrosequencing data of marine protists with near full-length sequence information. *Mar. Biol. Res.*, 10:771–780.
Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A. & Elshahed, M. S. 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.*, 75:5227–5236.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Comparison of nonmetric Multidimensional Scaling (nmMDS) plots of full-length and in silico extracted regions. nmMDS plots of the Bray–Curtis dissimilarity parameter based on OTUs called at 97% sequence similarity. nmMDS plots and Bray–Curtis dissimilarity values for full-length sequences and extracted regions were generated in Primer-E v6 (Clark and Warwick 2001). Similarity overlays represent 10% (solid black lines), 20% (dashed black lines), and 40% community similarity (solid gray lines). Each sample is labeled by a location and depth identifier (e.g. AO500: Arctic Ocean [AO], 500 m), Arctic Ocean AO, the San Pedro Ocean time series station in the eastern North Pacific (ENP), East Pacific Rise (EPR) in the eastern Pacific, Gulf Stream (GS) in the western North Atlantic, and Ross Sea, Antarctica (RS), see Table S2. The pattern of the nmMDS for the V7 region (**A**) was not consistent (lacked the resolving power) with nmMDS plots for the V4 (**B**), V1-V3 (**C**), V1-V4 (**D**), V4-V7 (**E**), V1-V7 (**F**), and full-length sequences (**G**). In contrast, nmMDS plots of the datasets consisting of sequences ≥ 400 nt yielded patterns that were generally consistent with one another (B–F). Only minor differences (e.g. the grouping of the Ross Sea samples, and some North Pacific and North Atlantic samples) were apparent among the patterns generated by longer sequences (B–F).

Table S1. Summary of mean nucleotide (nt) length of each sequence fragment used, along with forward and reverse primer and reference position on the 18S rRNA

gene used for in silico extraction. Position reference refers to the nucleotide position in the aligned full-length 18S rRNA gene sequence database. Fragment lengths reported include the primer sequences. The V1-V3 and V4 regions overlap along the 18S rRNA gene. Primers chosen represent commonly used primers for sequencing environmental microbial eukaryotic communities and therefore serve to provide realistic examples of short fragments that would be used in HTS.

Table S2. Summary of sampling dates, coordinates, depth, and number of quality checked sequences recovered from each sample.