Investigating Microbial Eukaryotic Diversity from a Global Census: Insights from a Comparison of Pyrotag and Full-Length Sequences of 18S rRNA Genes

Citation	Lie, A. A. Y., Liu, Z., Hu, S. K., Jones, A. C., Kim, D. Y., Countway, P. D., ... & 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 Genes. Applied and Environmental Microbiology, 80(14), 4363-4373. doi:10.1128/AEM.00057-14
DOI	10.1128/AEM.00057-14
Publisher	American Society for Microbiology
Version	Version of Record
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

AEM
Journals.ASM.org

Investigating Microbial Eukaryotic Diversity from a Global Census: Insights from a Comparison of Pyrotag and Full-Length Sequences of 18S rRNA Genes

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

Department of Biological Sciences, University of Southern California, Los Angeles, California, USA[a]; Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine, USA[b]; The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA[c]; Department of Geological Sciences, Brown University, Providence, Rhode Island, USA[d]; Environmental Research Institute, School of Science, University of Waikato, Hamilton, New Zealand[e]; College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA[f]; Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA[g]; College of Earth, Ocean, and Environment, University of Delaware, Lewes, Delaware, USA[h]

Next-generation DNA sequencing (NGS) approaches are rapidly surpassing Sanger sequencing for characterizing the diversity of natural microbial communities. Despite this rapid transition, few comparisons exist between Sanger sequences and the generally much shorter reads of NGS. Operational taxonomic units (OTUs) derived from full-length (Sanger sequencing) and pyrotag (454 sequencing of the V9 hypervariable region) sequences of 18S rRNA genes from 10 global samples were analyzed in order to compare the resulting protistan community structures and species richness. Pyrotag OTUs called at 98% sequence similarity yielded numbers of OTUs that were similar overall to those for full-length sequences when the latter were called at 97% similarity. Singleton OTUs strongly influenced estimates of species richness but not the higher-level taxonomic composition of the community. The pyrotag and full-length sequence data sets had slightly different taxonomic compositions of rhizarians, stramenopiles, cryptophytes, and haptophytes, but the two data sets had similarly high compositions of alveolates. Pyrotag-based OTUs were often derived from sequences that mapped to multiple full-length OTUs at 100% similarity. Thus, pyrotags sequenced from a single hypervariable region might not be appropriate for establishing protistan species-level OTUs. However, nonmetric multidimensional scaling plots constructed with the two data sets yielded similar clusters, indicating that beta diversity analysis results were similar for the Sanger and NGS sequences. Short pyrotag sequences can provide holistic assessments of protistan communities, although care must be taken in interpreting the results. The longer reads (>500 bp) that are now becoming available through NGS should provide powerful tools for assessing the diversity of microbial eukaryotic assemblages.

Protists are a group of extremely diverse organisms that dominate the living biomass and energy flow of planktonic microbial eukaryotic communities (1–3). They possess a wide range of distinct morphologies and physiologies and play important ecological roles in aquatic ecosystems as primary producers, consumers, predators, decomposers, parasites, and links to higher trophic levels (1, 4, 5). A major goal in protistan community ecology is to understand how diverse taxa within a community interact to influence overall ecosystem function, an objective that would benefit greatly from an ability to assess the entire protistan community using a single approach.

Characterizing the entire microbial eukaryotic community in an environment presents significant obstacles because of the very high diversity of natural protistan communities. Attaining this goal is nearly impossible using traditional, morphology-based approaches due to the existence of cryptic and morphologically nondescript species, the many taxonomic schemes employed for various protistan groups, and the different collection, fixation, and processing procedures on which they depend. The application of DNA sequencing and the movement toward a "molecular taxonomy" for protists, however, has begun to provide morphology- and culture-independent approaches to the investigation of protistan community diversity (4, 6, 7). Molecular characterization of protistan communities has focused largely on sequencing the small-subunit ribosomal gene (the 18S rRNA gene) rather than other genes (e.g., the cytochrome *c* oxidase I gene) more commonly used for barcoding other eukaryotes (8, 9). The application of such molecular approaches has lagged behind similar investigations of prokaryotic communities (10) but has already led to significant new discoveries in the field of protistan ecology, such as the presence of a large diversity of rare taxa in different environments (11, 12) and several previously undetected protistan lineages (13, 14).

The DNA sequence information used to characterize the species composition and diversity of natural protistan assemblages must be translated into taxonomic information that reflects species-level distinction if it is to be useful for protistan ecologists. Sequences are generally grouped into operational taxonomic units (OTUs) based on their similarity to each other or to reference sequences in databases. The use of DNA sequences for protistan ecology started with Sanger sequencing of relatively long segments (>500 bp) of the 18S rRNA gene or the full-length gene (≈1,800 bp) (see, e.g., references 15, 16, and 17). These sequence lengths

Received 7 January 2014 **Accepted** 2 May 2014
Published ahead of print 9 May 2014
Editor: C. R. Lovell
Address correspondence to Alle A. Y. Lie, alie@usc.edu.

doi:10.1128/AEM.00057-14

provided sufficient information for the development of OTU-calling programs that allowed species-level discrimination (18–20). More recently, however, next-generation sequencing (NGS) approaches have led to a shift toward the use of shorter but much more numerous sequences, generally from hypervariable regions of the 18S rRNA gene (see, e.g., references 21, 22, and 23). The relationship of these shorter sequences to longer reads and their appropriateness for use in diversity analysis have been questioned (24–28), but few direct comparisons have been made between the two types of sequences for microbial taxa, especially for microbial eukaryotes.

We obtained data sets of full-length and 454 pyrotag (V9 region) sequences of the 18S rRNA genes from a set of globally distributed oceanic samples in order to compare the information generated by these two independent sequencing approaches. We compared OTU calling for a data set of ~190,000 pyrotag sequences to OTUs established using ~6,500 full-length sequences. Our results revealed that pyrotag sequences called at 98% sequence similarity yielded a number of OTUs similar to that from full-length sequences called at 97% sequence similarity (approximately species level). Moreover, pyrotag sequences from a single pyrotag OTU aligned perfectly to full-length sequences from multiple full-length sequence OTUs, indicating that pyrotag sequences from one hypervariable region may not be suitable for differentiating all species present within an entire protistan community. Singletons (i.e., OTUs with only 1 sequence represented in the entire data set) strongly influenced the predictions of species richness for both types of sequence data but had little effect on the overall higher-level taxonomic composition. The protistan species richness observed from the pyrotag and full-length sequence data sets differed markedly, but the patterns of community similarities of samples as indicated by nonmetric multidimensional scaling (nMDS) plots were similar for the two data sets. Our study indicates that while the use of NGS for generating data sets has great potential for the investigation of microbial eukaryote diversity, some caution must be exercised when the data set is used for estimating species richness.

MATERIALS AND METHODS

Sample collection and extraction. Ten samples were collected from five different marine environments at various depths (5 to 2,500 m) between August 2000 and November 2005 as part of a Global Protistan Survey (Table 1). Water samples were collected using Niskin bottles mounted on a rosette equipped with conductivity, temperature, and depth (CTD) sensors.

Water samples of 2 to 20 liters were prefiltered through 200-μm and 80-μm mesh to remove most metazoa and were then filtered (<10 mm Hg) onto 45-mm GF/F (Whatman) filters. The filters were placed in 2 ml of 2× lysis buffer (100 mM Tris [pH 8], 40 mM EDTA [pH 8], 100 mM NaCl, 1% SDS) and were stored frozen (−20°C) on shipboard or flash frozen in liquid nitrogen for later DNA extraction. DNA was extracted according to the procedures detailed by Countway et al. (15).

Pyrotag sequencing and quality filtering. Pyrosequencing of the V9 hypervariable region of the 18S rRNA gene was performed according to the procedures described by Amaral-Zettler et al. (29). The primers used to sequence the V9 region were 1380F (5′-CCCTGCCHTTTGTACACAC-3′), 1389F (5′-TTGTACACACCGCCC-3′), and 1510R (5′-CCTTCYGCAGGTTCACCTAC-3′). The emulsion PCR was performed using standard Roche protocols, and sequencing was carried out on the Genome Sequencer FLX system (Roche, Basel, Switzerland) using the GS LR70 long-read sequencing kit (Roche).

Sequence reads that did not have (i) an exact match to the proximal primer (1380F or 1389F) and (ii) the presence of the distal primer (1510R) were removed, as were reads that had one or more ambiguous bases (N′s). Further quality filtering was done using the free software package mothur, version 1.30.0 (http://www.mothur.org/), by following the standard operating procedures for 454 sequences (30). These procedures included removing pyrotags with an average quality score of <35 using a 50-bp sliding window and those with >8 homopolymers present. Preclustering with a 1-bp mismatch allowance was performed to minimize the effects of sequencing or amplification errors on OTU clustering (31), and chimeras were removed using the UCHIME algorithm (32). Sequences are available at the Visualization and Analysis of Microbial Population Structure (VAMPS) website, hosted by the Marine Biological Laboratory (VAMPS.mbl.edu), and in the NCBI Sequence Read Archive under SRA number SRP001225.

TABLE 1 Date, location, coordinates, and depth of sample collection and the number of sequences obtained from each sample

Collection date (day mo yr)	Location	Coordinates	Sampling depth (m)	No. of sequences[a] obtained	
				PT	FL
24 Aug 2000	Gulf Stream, North Atlantic	34.73N, 73.95W	15	23,329	474
			105	8,182	513
29 Oct 2001	Eastern North Pacific	33.55N, 118.4W	5	9,392	603
12 Aug 2002	Arctic Ocean	73.42N, 157.40W	35	16,202	733
11 Aug 2002	Arctic Ocean	73.42N, 157.40W	500	18,078	869
8 Dec 2003	East Pacific Rise, North Pacific	9.84N, 104.35W	20	37,034	450
			1,500	4,736	879
			2,500	5,463	707
14 Nov 2005	Ross Sea, Southern Ocean	76.04S, 170.30E	20	37,034	657
			600	31,246	694

[a] PT, pyrotag; FL, full-length.

Full-length sequencing and quality filtering. Full-length 18S rRNA gene sequences were amplified by PCR using the universal eukaryote primers Euk-A (5′-AACCTGGTTGATCCTGCCAGT-3′) and Euk-B (5′-GATCCTTCTGCAGGTTCACCTAC-3′). The PCR and cloning procedures are based on those of Countway et al. (33) with slight modifications, including the use of AmpliTaq Gold for PCR amplification, and amplicons were cloned using a TOPO-TA kit (Invitrogen). Sequencing of the full-length 18S rRNA gene was conducted by the Joint Genome Institute (JGI, Walnut Creek, CA) using two vector primers (T3 and T7) and an internal primer (Euk-570F).

Sequences that did not have both the proximal (Euk-A) and distal (Euk-B) primers, or that could not be assembled due to failure of the internal sequencing primer, were removed from the data set. Sequences were then passed through the pintail algorithm to remove chimeras (34).

OTU calling. The mothur software package (version 1.30.0) was also used for calling OTUs for both the pyrotag and full-length sequence data sets (average neighbor method) to ensure that the OTU-calling procedures for the two sequence data sets were comparable. An aligned SILVA eukaryotic full-length 18S rRNA gene reference database (version 102; aligned and provided by the mothur software package) was used to aid in the alignment of full-length and pyrotag sequences using mothur prior to OTU calling. All full-length sequences were pooled, and OTU calling was performed at a sequence similarity of 97% based on the observation that OTUs called using full-length sequences at 97% by mothur yielded a number of OTUs most comparable to that called by the Microbial Eukaryote Species Assignment (MESA) program at 95% sequence similarity (20; S. K. Hu and D. A. Caron, unpublished data). Calling OTUs at 95% sequence similarity in MESA was designed to create OTUs of full-length or nearly full length 18S rRNA gene sequences at approximately species-level distinctions using a data set of well-curated and morphologically well described protistan species and their sequences (20).

Comparison of the numbers of OTUs formed. The pyrotag and full-length sequence data sets were first compared by examining the number of

OTUs formed by the two data sets. Pyrotag sequences were randomly subsampled to yield the same number of sequences in each sample as the full-length data set (i.e., to standardize to the full-length sequence data set). This subsampled pyrotag data set was used to call OTUs at different sequence similarities (95 to 100%) using mothur, and the resulting numbers of OTUs were compared to the numbers of OTUs called for the full-length sequences using mothur at 97% sequence similarity.

The analysis described above allowed the comparison of OTU formation between the pyrotag and full-length sequence data sets using the same number of sequences, in order to avoid the potential bias inherent in using a much larger pyrotag sequence data set. One of the strengths of NGS, however, is the substantially greater number of sequences that can be generated from a sample. Subsequent comparisons between the pyrotag and full-length sequence data sets, therefore, included all pyrotag sequences. Richness for the two data sets (all full-length or all pyrotag sequences) was examined through the construction of rarefaction curves for both data sets. Rarefaction curves of the observed numbers of OTUs were generated for OTUs called at 97% sequence similarity for both data sets and additionally at 98% and 99% similarity for the pyrotag data set. Rarefaction curves were constructed with and without the inclusion of singletons in the two data sets.

Taxonomy of pyrotag and full-length sequence OTUs. The best BLAST result for a representative sequence from each OTU using the SILVA database (version 111) was used to provide taxonomic information for pyrotag and full-length sequence OTUs. The representative sequence from each OTU was selected by mothur and was the sequence with the smallest distance to all other sequences within the OTU. If more than one sequence matched this condition, then the sequence with the smallest average distance to other sequences in the OTU was selected.

Mapping pyrotags to full-length sequences. All pyrotag sequences from the 20 pyrotag OTUs that had the most pyrotag sequences were mapped to the full-length sequences in order to investigate how an OTU generated with pyrotag sequences related to OTUs generated with full-length sequences. The pyrotag sequences were mapped to full-length sequences using the Burrows-Wheeler Aligner algorithm with no mismatch allowance (35), and the number of full-length OTUs matched perfectly by the pyrotag sequences in a single pyrotag OTU was tallied. In the event that a pyrotag sequence mapped perfectly to multiple full-length sequences (i.e., the full-length sequences had identical V9 regions), one of these matching full-length sequences was randomly selected as the perfect match to the pyrotag sequence (i.e., each pyrotag sequence matched with only one full-length sequence for subsequent analysis).

nMDS analysis of the pyrotag and full-length data sets. The results of nonmetric multidimensional scaling (nMDS) analysis using the pyrotag and full-length sequence data sets were compared in order to investigate potential differences in community similarities inferred from the two types of sequences. The number of sequences in each data set was standardized separately for the two data sets (450 for the full-length data set and 4,736 for the pyrotag data set) by randomly subsampling the sequences in each sample down to the number of sequences in the sample with the least number of sequences for each data set. This standardization procedure was performed to avoid the potential bias of having different numbers of sequences in each sample. In addition to the analyses using the full-length and pyrotag sequence data sets that included singletons, a third analysis was performed on the pyrotag data set with the singletons removed (yielding a data set of 4,472 pyrotags per sample). All subsampling of sequences was performed using mothur. The standardized data sets were square-root transformed to down-weight highly represented OTUs. The PRIMER software package (version 6) was used to calculate the Bray-Curtis community composition similarity values that were based on the OTU composition of each sample. A matrix of pairwise Bray-Curtis similarity values was constructed for each data set (i.e., full-length sequences, pyrotag sequences, and pyrotag sequences without singletons) and was then used to perform CLUSTER (group average mode with the SIMPROF test for significance) and nMDS analyses. The results of the SIMPROF significance test from the CLUSTER analysis were overlaid on the nMDS plots, which are visual representations of the similarities among communities.

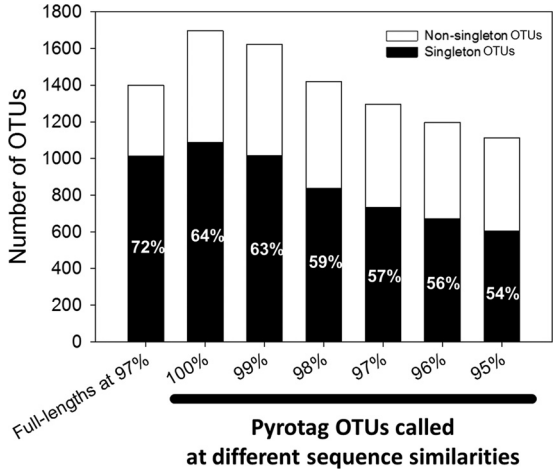

FIG 1 Numbers of pyrotag OTUs formed at different sequence similarities compared to numbers of full-length sequence OTUs formed at 97% sequence similarity (leftmost bar). The filled section of each bar represents the number of singleton OTUs, and the exact percentage is given inside each bar. The pyrotag data set for this analysis was standardized to the full-length sequence data set by randomly subsampling pyrotag sequences in each sample to match the number of full-length sequences in the sample (Table 1).

Nucleotide sequence accession numbers. The full-length 18S rRNA gene sequences analyzed in this study have been submitted to GenBank under accession numbers KJ757035 to KJ759741, KJ760393 to KJ762454, and KJ762829 to KJ764638.

RESULTS

The number of pyrotags obtained after applying quality-filtering procedures ranged from 4,736 to 37,034 per sample, resulting in a total of 190,696 pyrotag sequences for all 10 samples (Table 1). Pyrotag sequences were ~110 bp long (mean, 110 bp; 25% quartile, 110 bp; 75% quartile, 112 bp) after quality-filtering procedures (including the removal of primers). A total of 6,579 full-length sequences were available after quality filtering. The number of full-length sequences in each sample ranged from 450 to 879.

Numbers of OTUs from pyrotag and full-length sequence data sets. We compared the number of OTUs formed from pyrotag sequences called at different sequence similarities (95 to 100%) to the number of OTUs formed from full-length sequences called at 97% sequence similarity in order to investigate which sequence similarity used to call OTUs for the pyrotag data set resulted in a number of OTUs most comparable to that generated by the full-length sequence data set. The pyrotag data set for this analysis was standardized by randomly subsampling the number of pyrotags in each sample to match the number of full-length sequences in each sample (Table 1) in order to avoid potential biases from using a much larger pyrotag data set. OTU calling for all full-length sequences (6,579 sequences) at 97% similarity in mothur resulted in a total of 1,400 OTUs, while the same number of pyrotag sequences called at various sequence similarities, ranging from 95 to 100% sequence similarity, resulted in 1,113 to 1,695 OTUs (i.e., ≈80 to 120% of the number of OTUs called for the full-length sequences) (Fig. 1). OTU calling of the pyrotag sequences at 98% sequence similarity yielded a total of 1,419 OTUs, which was most

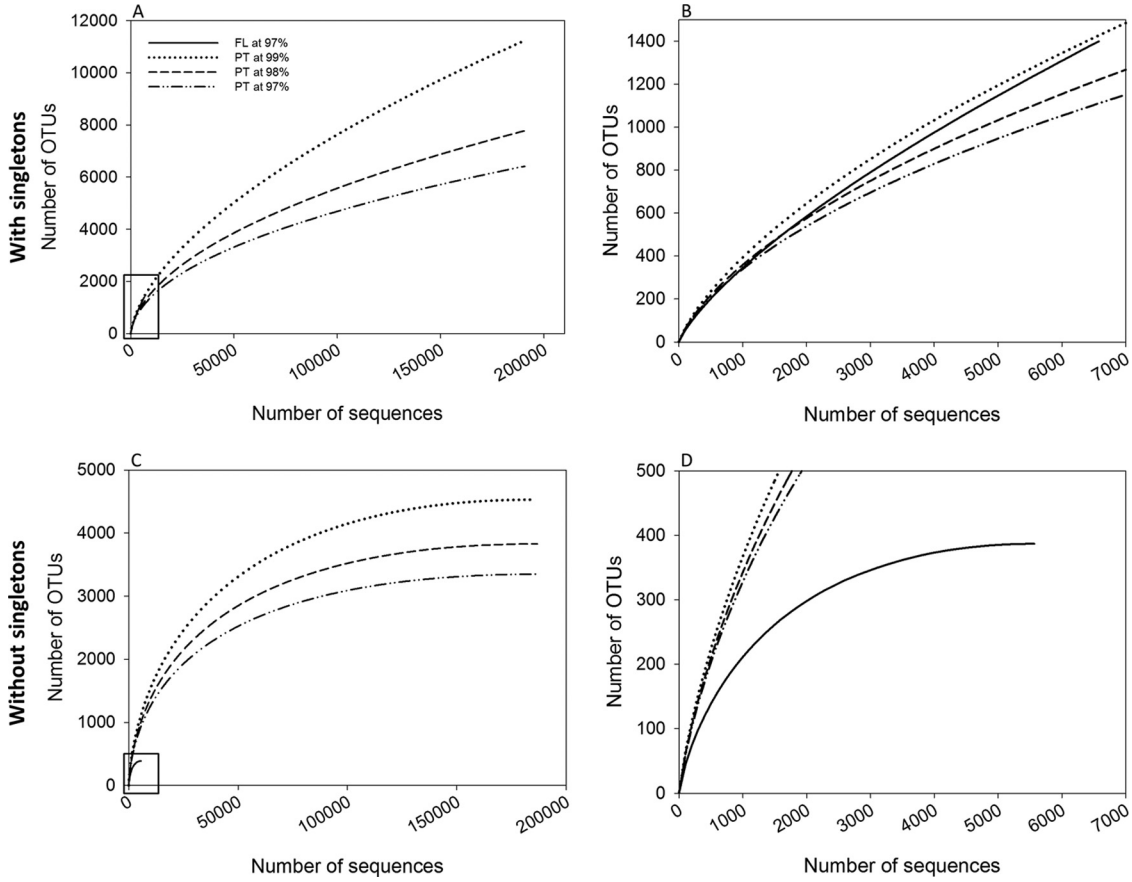

FIG 2 Rarefaction curves generated using the full-length (FL) and pyrotag (PT) sequence data sets. OTUs were called at 97%, 98%, and 99% sequence similarities for the pyrotags. (A and B) Curves obtained using all sequences and OTUs. Panel B is an enlargement of the boxed area in panel A. (C and D) Curves obtained after the removal of singleton OTUs (i.e., OTUs with only one sequence each). Panel D is an enlargement of the boxed area in panel C.

similar to the number of full-length OTUs called at 97%. Singleton OTUs (i.e., OTUs composed of only one sequence in the data set) constituted a large portion (>50%) of the OTUs in all treatments of both data sets. Calling OTUs using full-length sequences at 97% sequence similarity generated the highest proportion of singletons (72% of the total number of OTUs) compared to the number of singletons generated from the standardized pyrotag sequence data called at different levels of similarity (range, 54 to 64% of the total number of OTUs). The pyrotag singletons in this analysis included some "secondary" singletons that were generated as a consequence of the subsampling process. The use of the standardized data set yielded an average of 59% singleton OTUs over the range of sequence similarities used to call OTUs (Fig. 1), whereas calling of OTUs using the entire pyrotag data set yielded an average of 51% singleton OTUs.

OTU rarefaction curves from the pyrotag and full-length sequence data sets. Rarefaction curves were constructed using all sequences available for the pyrotag (190,696 sequences) and full-length (6,579 sequences) sequence data sets in order to examine sequence coverage (Fig. 2). Full-length OTUs were called at 97% similarity, while pyrotag OTUs were called at 97%, 98%, and 99% sequence similarities. Calling pyrotag sequences at 98% sequence similarity resulted in a number of OTUs that was most comparable to that from the full-length sequence data set in the previous standardized comparison (see the preceding section), and OTU definitions of 97% and 99% were included in order to investigate how differences of 1% in sequence similarity would affect the analysis results. The entire pyrotag data set generated a large number of OTUs (at 97% sequence similarity, 6,414 OTUs; at 98%, 7,775 OTUs; at 99%, 11,234 OTUs), while the full-length sequence data set generated 1,400 OTUs, as noted above. The rarefaction curve for the full-length sequence data set was similar to the pyrotag rarefaction curves for the portion of the curves where the data sets (i.e., number of sequences sampled) overlapped (Fig. 2B). The curve for the full-length sequences appeared to have a slightly different trajectory than those for the pyrotag sequences (Fig. 2B), but the curve for the full-length sequences was always bracketed by the curves generated by the pyrotag OTU data sets called at 97%, 98%, and 99% sequence similarities.

The removal of singleton OTUs altered the form of the rarefaction curves dramatically. The effects of removing singletons were investigated because there have been suggestions that some singleton OTUs may be artifactual (31, 36, 37) (for more detail, see Discussion). The number of OTUs observed decreased substantially for both data sets with the removal of singletons, as expected (Fig. 2C and D). The number of OTUs observed for the pyrotag data set was reduced to 40 to 52% of the total number of OTUs when singletons were removed (from 11,234 to 4,530 OTUs at 99% sequence similarity; from 7,775 to 3,828 OTUs at 98%; from 6,414 to 3,349 OTUs at 97%). The number of OTUs observed in

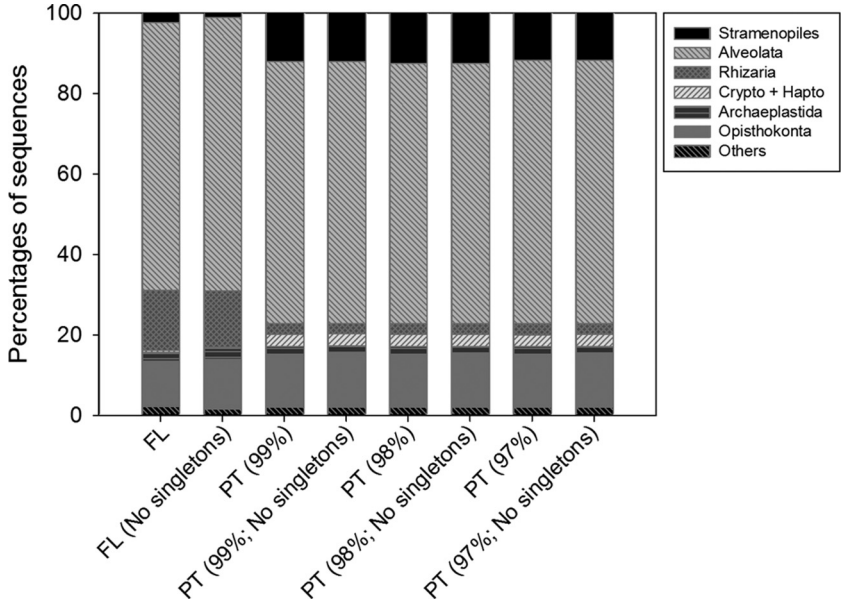

FIG 3 Comparison of the higher-level taxonomic compositions (percentages of total sequences) of the pyrotag (PT) and full-length (FL) sequence data sets. Taxonomic information was derived from best-match results obtained by BLAST searching of representative sequences from each OTU (selected by mothur) against the SILVA database. Full-length sequence OTUs were called at 97% sequence similarity, while pyrotag OTUs were called at 97%, 98%, and 99% sequence similarities. Results with and without singleton OTUs are shown.

the full-length sequence data set without singletons was reduced to 28% of the total number of OTUs (from 1,400 to 387 OTUs). The rarefaction curve for the full-length data set without singletons had a much more shallow slope and a lower maximum than the pyrotag rarefaction curves for the same number of sequences sampled (Fig. 2D).

Taxonomic compositions of pyrotag and full-length sequence data sets. The higher-level taxonomic compositions (supergroup level) of the entire pyrotag and full-length sequence data sets were tallied in order to (i) compare the taxonomic compositions obtained by calling pyrotag OTUs at different levels of similarity, (ii) compare the taxonomic compositions of the pyrotag and full-length sequence data sets, and (iii) examine the effect of the removal of singletons on the taxonomic information inferred using both data sets. This analysis focused on pyrotag OTUs called at 97%, 98%, and 99% sequence similarities for the reasons given in the preceding section.

Calling OTUs using the pyrotag data set at three different levels of sequence similarity had very little effect on the higher-level taxonomic composition of the sequences (Fig. 3). Similarly, the presence or removal of singleton OTUs did not substantially influence the taxonomic composition derived from either the full-length or the pyrotag sequence data set. However, the full-length sequence data set yielded a larger proportion of sequences identified as rhizarians than the pyrotag sequence data set (14 to 15% of full-length sequences and ≈3% of pyrotag sequences) and smaller proportions of sequences identified as stramenopiles (1 to 3% of full-length sequences and 12 to 13% of pyrotag sequences) and as cryptophytes and haptophytes (≈0% of full-length sequences and ≈3% of pyrotag sequences). Other groups (e.g., alveolates) constituted similar proportions of the two data sets.

Mapping pyrotags to full-length sequences. All pyrotag sequences from a single pyrotag OTU were mapped to full-length sequences at 100% sequence similarity in order to investigate how individual pyrotag OTUs related to full-length OTUs. Full-length OTUs for the analysis were called at 97% sequence similarity. This analysis focused on 20 pyrotag OTUs that had the most pyrotag sequences. Pyrotag OTUs were called at 95%, 97%, and 99% sequence similarities (Table 2).

Pyrotag sequences from nearly all of the 20 pyrotag OTUs with the most sequences mapped to full-length sequences from multiple full-length OTUs. The results were similar among the three sequence similarities examined for the pyrotags, but lower sequence similarity tended to hit more full-length OTUs. In only three cases did the pyrotags from a well-populated pyrotag OTU match a single full-length OTU (the 17th, 18th, and 19th most populated OTUs when pyrotag sequences were called at 99%, 97%, and 95% sequence similarities, respectively). The pyrotag OTUs with the most sequences formed at 95%, 97%, or 99% sequence similarity consisted of high numbers of pyrotags (34,787 to 49,709), and pyrotags from these three OTUs mapped perfectly to full-length sequences that were distributed among 127, 109, or 75 full-length OTUs, respectively (Table 2). While sequences from a single pyrotag OTU aligned perfectly to sequences from multiple full-length OTUs, the full-length OTUs were generally from the same higher-level (e.g., class or order) taxonomic group (e.g., Dinoflagellata) (Table 3).

nMDS analysis using pyrotag and full-length sequences. nMDS plots were constructed using the pyrotag and full-length sequence data sets to (i) compare community similarities among samples using pyrotag and full-length sequences (Fig. 4A and B) and (ii) examine the effect of the removal of singletons on the clustering of communities for the pyrotag data set (Fig. 4B and C). OTUs for the pyrotag data set were called at 98% sequence similarity, while OTUs for full-length sequences were called at 97%, for the reasons provided above. Each data set was standardized to

TABLE 2 Numbers of full-length OTUs containing full-length sequences that were mapped perfectly by pyrotags in each of the 20 pyrotag OTUs with the most pyrotag sequences called at 95%, 97%, and 99% sequence similarities[a]

PT OTU[b]	PT OTU called at 99%			PT OTU called at 97%			PT OTU called at 95%		
	No. of PT sequences in the OTU	No. of PT sequences with perfect alignment to FL sequences	No. of FL OTUs	No. of PT sequences in the OTU	No. of PT sequences with perfect alignment to FL sequences	No. of FL OTUs	No. of PT sequences in the OTU	No. of PT sequences with perfect alignment to FL sequences	No. of FL OTUs
1	34,787	31,239	75	44,094	37,497	109	49,709	40,628	127
2	6,180	5,661	29	7,809	6,622	21	8,281	6,306	30
3	6,046	5,053	26	6,503	5,876	30	7,811	6,622	21
4	5,532	4,727	39	6,416	5,090	26	6,524	5,876	30
5	3,962	3,544	12	4,290	3,782	44	5,451	4,664	47
6	3,949	3,222	7	4,139	3,251	7	4,177	3,255	7
7	3,758	3,452	44	3,014	1,965	6	3,274	2,786	4
8	3,699	3,075	11	2,509	2,082	6	3,138	1,965	6
9	2,420	2,109	3	2,453	2,110	3	2,615	2,146	7
10	2,240	1,865	6	2,298	0	0	2,345	1,784	7
11	2,188	0	0	2,201	1,784	7	2,314	0	0
12	2,041	1,687	6	2,096	1,687	6	2,116	1,687	6
13	1,890	1,618	4	1,928	1,618	4	1,947	1,618	4
14	1,831	1,657	10	1,866	1,657	10	1,866	1,657	10
15	1,806	1,510	3	1,812	1,111	10	1,325	1,090	3
16	1,583	1,328	6	1,320	1,090	3	1,218	899	9
17	1,024	937	1	1,067	859	8	1,090	943	6
18	943	846	8	1,042	937	1	1,069	825	6
19	919	835	8	925	835	8	1,047	937	1
20	910	852	15	925	852	15	932	835	8

[a] PT, pyrotag; FL, full-length. Pyrotags were aligned to full-length sequences with no mismatch allowance, and the total numbers of OTUs (called at 97% sequence similarity) formed by these full-length sequences are presented.
[b] The 20 most populated PT OTUs (i.e., those with the most pyrotag sequences) are listed in order from the most to the least populated.

the number of sequences in the sample with the lowest number of sequences (for the full-length data set, 450 sequences/sample; for the pyrotag data set, 4,736 sequences/sample; for the pyrotag data set without singletons, 4,472 sequences/sample). The Bray-Curtis similarity values estimated for the full-length data set were generally lower (average, 5%) than those estimated for the pyrotag data sets (average with singletons, 14%; average without singletons, 13%), but the overall patterns and the clustering of some communities in the nMDS plots were similar for the three data sets (Fig. 4). For example, communities from the Ross Sea (RS) consistently clustered together, and communities from deep water at the East Pacific Rise (EPR) (sampling depths, 1,500 m and 2,500 m) consistently clustered together as well in both the pyrotag and full-length sequence data sets. One difference between the pyrotag and full-length sequence data sets concerned the community from the Arctic Ocean (AO) collected at 35 m. The two Arctic samples (35 and 500 m) were not significantly different (P, >0.05 by the SIMPROF test) in the full-length data set (Fig. 4A) but were significantly different in the pyrotag data sets. There were also some differences in the clustering of communities collected from the Gulf Stream (GS), EPR, and Eastern North Pacific (ENP) among the three data sets as indicated by the results of the SIMPROF significance test (i.e., subclusters on the nMDS plots), although these samples occupied the same positions on the plots in all three analyses.

The removal of singletons from the pyrotag data set had little effect on the resulting nMDS plot except that the community collected from the ENP at 5 m was statistically similar to the community collected from the GS at 15 m in the nMDS plot constructed with the pyrotag data set after the removal of singletons (Fig. 4C). On the other hand, the community collected from the ENP at 5 m was not statistically different from the community collected from the EPR at 20 m in the nMDS plot constructed with the entire pyrotag data set (Fig. 4B).

DISCUSSION

The use of DNA sequence information to characterize natural microbial communities has ushered in a new era in which next-generation sequencing (NGS; e.g., Illumina or 454 pyrosequencing) is rapidly replacing Sanger sequencing. NGS produces vast amounts of sequence information at a relatively low cost, but early versions of these approaches have been hampered by possible sequencing errors and the relatively short lengths (100 to 200 bp) of the sequences (25). The lengths of NGS sequences have been increasing as the technology advances, but there are presently few studies that have examined the relationship between short sequences (<200 bp) and much longer sequences (1,500 to 1,800 bp), as well as how differences might affect ecological interpretations of sequence-based investigations of microbial community structure and diversity (26–28). This study directly compared full-length 18S rRNA gene sequences obtained by Sanger sequencing with short reads of the hypervariable V9 region of the same gene obtained by pyrosequencing within a set of 10 globally distributed samples.

Comparison of different sequence similarities for calling pyrotag OTUs. Pyrotag OTUs that were formed at a higher sequence

TABLE 3 Most common taxonomy of the full-length sequences that were mapped perfectly by pyrotag sequences from each of the 5 pyrotag OTUs with the most pyrotag sequences called at 95%, 97%, and 99% sequence similarities[a]

Sequence similarity (%) at which PT OTUs were called	PT OTU[b]	No. of PT sequences with matches to FL sequences/ total no. of PT sequences in the OTU (% with matches to FL sequences)	Most common taxonomy of matching FL sequences that were mapped by PT sequences	% of PT sequences with matches to FL sequences that belonged in the most common taxonomy
99	1	31,239/34,787 (90)	Alveolata: Dinoflagellata	99
	2	5,661/6,180 (92)	Alveolata: Syndiniales	99
	3	5,053/6,046 (84)	Alveolata: Dinoflagellata	88
	4	4,727/5,532 (85)	Alveolata: Dinoflagellata	84
	5	3,544/3,962 (89)	Opisthokonta: Metazoa	99
97	1	37,497/44,094 (85)	Alveolata: Dinoflagellata	97
	2	6,622/7,809 (85)	Opisthokonta: Metazoa	99
	3	5,876/6,503 (90)	Alveolata: Syndiniales	99
	4	5,090/6,416 (79)	Alveolata: Dinoflagellata	88
	5	3,782/4,290 (88)	Alveolata: Ciliophora	96
95	1	40,628/49,709 (82)	Alveolata: Dinoflagellata	96
	2	6,306/8,281 (76)	Alveolata: Dinoflagellata	87
	3	6,622/7,811 (85)	Opisthokonta: Metazoa	99
	4	5,876/6,524 (90)	Alveolata: Syndiniales	99
	5	4,664/5,451 (86)	Alveolata: Ciliophora	96

[a] FL, full-length; PT, pyrotag. Taxonomies of full-length sequences were best-match results from BLAST searching of a representative full-length sequence (selected by mothur) from the full-length OTU against the SILVA database. Full-length OTUs were called at 97% sequence similarity.
[b] The 5 most populated PT OTUs (i.e., those with the most pyrotag sequences) are listed in order from the most to the least populated.

similarity (98%) yielded a similar number (~1,400 OTUs) as full-length OTUs called at 97% sequence similarity (Fig. 1). Our results indicate that the utility of the V9 hypervariable region of the 18S rRNA gene to adequately distinguish eukaryotic taxa may be less than the taxonomic resolution afforded by the entire gene. Dunthorn et al. (38) concluded that the genetic distances among species of ciliates that were estimated using the full-length 18S rRNA gene were more similar to the genetic distances estimated with the V4 region than to those estimated with the V9 region, similarly suggesting that the V9 region of the 18S rRNA gene may not be the ideal choice for distinguishing between ciliate species. Youssef et al. (24) also reported that some hypervariable regions of the 16S rRNA gene underestimated bacterial richness in comparison to full-length sequences called at the same level of sequence similarity.

The choice of a sequence similarity for calling OTUs heavily influenced the overall species richness observed for the entire pyrotag data set (Fig. 2A). A 2% difference in sequence similarity for OTU calling resulted in an almost 2-fold difference in the number of OTUs formed. The numbers of OTUs formed from the pyrotag data set called at 97% and 99% sequence similarities were 6,414 and 11,234, respectively, when all sequences were included for OTU formation (Fig. 2A). The removal of singletons greatly reduced the overall number of OTUs observed, as well as the difference in the number of OTUs observed between the two sequence similarities (3,349 OTUs for 97% sequence similarity and 4,530 OTUs for 99%). Stoeck et al. (23) also observed that a small difference in the sequence similarity used for calling OTUs resulted in a large difference in the number of OTUs formed. Caution is therefore needed in assessing the ecological information contained in sequence data sets, because the sequence similarity level used for calling OTUs can significantly affect the species richness. There is currently no consensus on a single sequence similarity for distinguishing between species, even for a closely related and well-defined taxonomic group such as the ciliates (39), let alone for all the different groups of protists.

Effect of singletons on observed richness. Singletons contributed significantly to the number of OTUs formed in all treatments of the data set (Fig. 1) and heavily influenced the observed species richness (Fig. 2). Singletons would also heavily influence the estimation of total species richness using richness estimators (40–42). Questions regarding the potentially artifactual nature of pyrotag singletons have been raised previously, and a few studies employing collections of known taxa as artificial communities have demonstrated that some singletons were sequences that were not a part of the original community (31, 36, 37). Dickie (41) noted that the number of sequencing artifacts for pyrosequencing increased as the number of pyrotags increased. Therefore, some studies have removed singletons to avoid overestimating species richness, despite the risk of removing real taxa present at very low abundances (see, e.g., references 22, 43, 44, and 45). While the removal of singletons resulted in asymptotes of the rarefaction curves in this study (Fig. 2), the removal of singletons did not always lead to asymptotes (22, 44).

Full-length sequences are less likely to contain a significant percentage of sequencing artifacts that would lead to the formation of artifactual singleton OTUs than are very short sequences (e.g., sequences with ~100 bp in this study). Nonetheless, singleton OTUs still contributed significantly to the full-length sequence data set (72%) (Fig. 1). This result would argue that a substantial "rare biosphere" existed within the data set generated from these globally distributed samples, regardless of the nature of the pyrotag singletons. We also compared rarefaction curves generated from the pyrotag data set with singletons removed to the

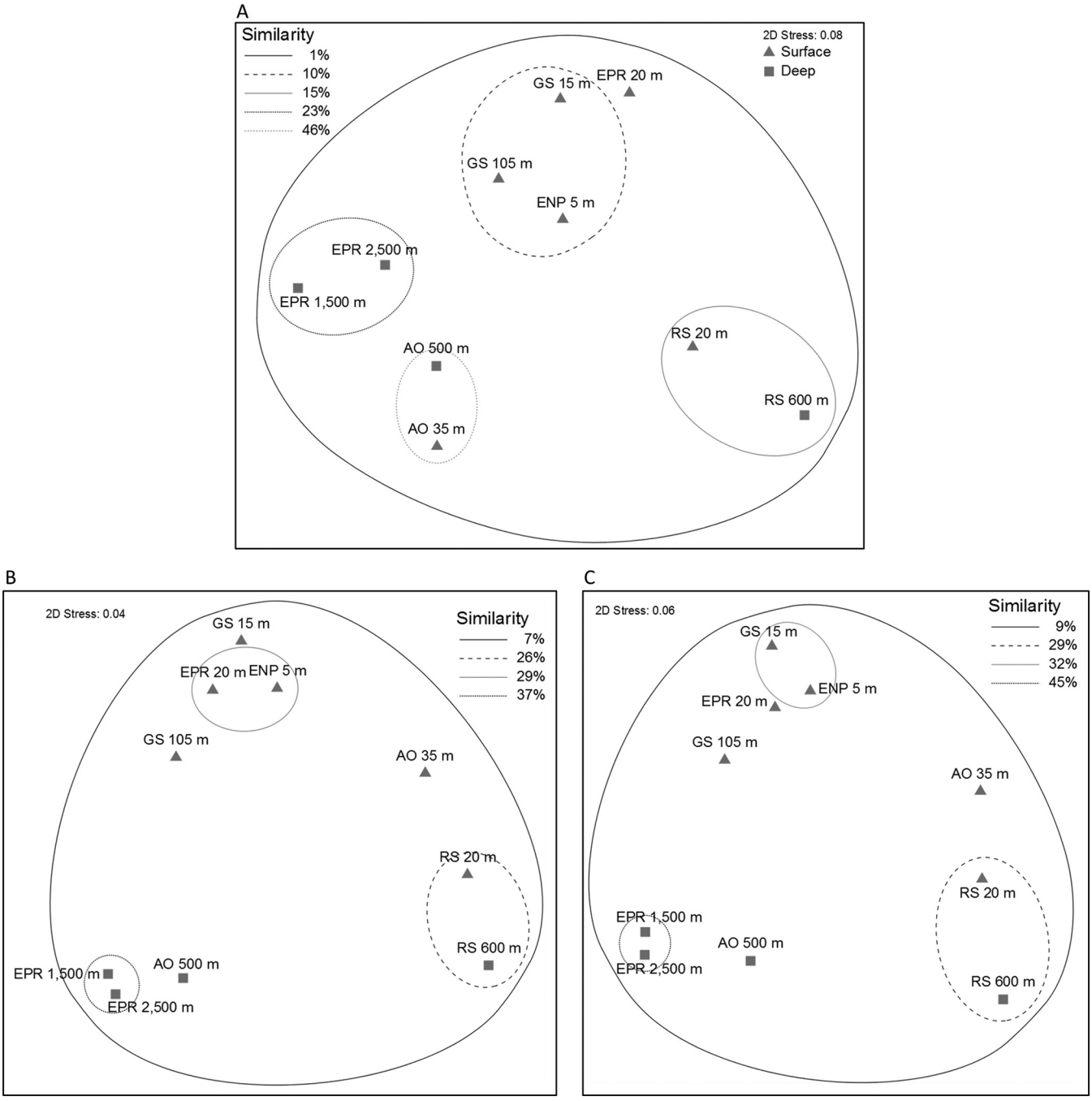

FIG 4 Nonmetric multidimensional scaling analysis using pairwise Bray-Curtis similarity values estimated from the pyrotag and full-length sequence data sets. (A) The number of full-length sequences was standardized to 450 sequences per sample. (B) The pyrotag data set including singletons was standardized to 4,736 sequences per sample. (C) The pyrotag data set without singletons was standardized to 4,472 sequences per sample. OTUs formed from full-length sequences were called at 97%, while OTUs formed from pyrotag sequences were called at 98%. Group average similarity values of clusters with significant differences, indicated by the CLUSTER analysis (P, <0.05 by the SIMPROF test), were overlaid on the MDS plot. Communities within the same subcluster were not significantly different from each other (P, >0.05).

curve generated from the full-length sequence data set with singletons included (data not shown). The curve for the full-length sequence data set had a steeper slope and indicated a higher observed richness than the pyrotag curves for a given number of sequences sampled. Taken together, these results imply that the pyrotag data set, with singletons removed, would result in a lower number of species observed (assuming that most full-length singletons represented valid OTUs). We speculate that the true observed richness of the samples (as constrained by the current methodologies for filtering, DNA extraction, and sequencing) most likely lies between the full-length curve with singletons and the pyrotag curves without singletons.

Taxonomic composition. Singletons had a large effect on observed OTU richness, but they had a negligible effect on the overall taxonomic composition for both data sets (Fig. 3). This result implies either that the singletons were real and had proportionally the same taxonomic composition as the entire data set or that they were artifactual sequences generated at proportions similar to those of the populated OTUs from which they were generated. Regardless of whether singletons are artifactual sequences or sequences from rare organisms, they are not biased toward specific or unique groups.

OTUs obtained using the pyrotag data set at different sequence similarities also did not lead to substantive changes in higher-level taxonomic composition (Fig. 3). Therefore, the coalescing of pyrotag OTUs as the imposed sequence similarity was relaxed resulted in OTUs generally converging to the same high-level taxonomic groups. On the other hand, the taxonomic composition differed between the full-length and pyrotag data sets (Fig. 3). The full-length sequence data set yielded a higher proportion of rhizarians and smaller proportions of stramenopiles, cryptophytes, and haptophytes. This difference in taxonomic composition was not due to differences between the samples or DNA extraction procedures, because both the full-length data set and the pyrotag data set were sequenced from the same DNA extracts. These differences are more likely due to differences in the primers used for the two types of sequencing approaches (46). Other studies comparing full-length and pyrotag sequences have also demonstrated differences in the relative proportions of certain groups within the community, with certain groups or species detectable only by either Sanger sequencing or pyrosequencing (26, 27). In the present study, the full-length sequence data set yielded a very small proportion of cryptophytes and haptophytes (0.38% from the data set with singletons; 0.25% from the data set without singletons), possibly due, in part, to biases against haptophytes in the primers used for the full-length sequences (11, 47).

We also compared the taxonomic compositions of the two data sets at a higher taxonomic resolution (i.e., genus level) for sequences identified as ciliates (data not shown). The results were similar to those at the supergroup level, in which the presence/absence of singletons and the sequence similarity threshold used to call pyrotag OTUs had little effect on the genus composition. Certain genera were detected only by one of the two sequencing methods, but unsurprisingly, the pyrotag data set detected a much larger number of genera (43 to 58 genera for pyrotag sequences called at 97 to 99% sequence similarity, with and without the presence of singletons) than the full-length sequence data set (25 genera for full-length sequences called at 97% sequence similarity with singletons; 8 genera without singletons). Despite the discrepancies in the numbers of genera identified by the two data sets, both data sets had the most sequences identified as members of the genus *Strombidium*. Similarly, Santoferrara et al. (28) compared the investigation of ciliate diversity using Sanger sequencing, pyrosequencing, and microscopy and found that the dominant taxa obtained by the three methods were generally comparable.

Relating pyrotags from one pyrotag OTU to full-length OTUs. Pyrotags from a single pyrotag OTU that had many sequences generally mapped at 100% sequence similarity with full-length sequences associated with multiple full-length OTUs (Table 2). This result could be explained by an inability of the short pyrotags to resolve many OTUs or by the possibility that the OTUs formed from full-length sequences were overestimating diversity. We speculate that the former situation is the more likely explanation. The majority of full-length sequences with identical V9 regions were found to be placed in the same full-length OTU (called at 97% sequence similarity) in the present study, but there are rare cases in which full-length sequences with exactly the same V9 region can be <90% similar across the entire gene (data not shown). Similarly, Sogin et al. (48) randomly compared the pairwise distances of 5×10^6 full-length 16S rRNA (bacterial) gene sequences and found that >90% of full-length sequences with identical V6 hypervariable regions were at least 95% similar across the entire gene. This suggested that ~10% of full-length sequences with identical V6 regions can be >95% different across the gene and that a single hypervariable region might not contain sufficient information to differentiate many OTUs. If this speculation is true, our results indicate that estimates of community richness obtained from pyrotag-based OTUs should be viewed with caution.

Despite this caveat, full-length sequences from multiple OTUs that were matched perfectly by pyrotags from a single pyrotag OTU were generally from the same higher taxonomic group (e.g., Dinoflagellata [Table 3]). This was the case despite the fact that pyrotags from a single pyrotag OTU matched as many as 127 full-length OTUs (Table 2). Thus, investigations of the higher taxonomic composition (Fig. 3 presents an example) should be less affected by the level of resolution afforded by the pyrotags. Huse et al. (49) performed an *in silico* analysis that mapped V3 and V6 hypervariable regions to the full-length 16S rRNA gene and obtained similar results in that 99% of the short sequence tags matched full-length sequences from the same class, order, or phylum level.

The number of full-length OTUs matched by pyrotags in a single pyrotag OTU decreased gradually as the number of pyrotags in each pyrotag OTU decreased (Table 2). This finding may simply reflect the fact that less-populated pyrotag OTUs had fewer sequences to match to full-length OTUs. Our analysis yielded only 3 pyrotag OTUs (one for each sequence similarity threshold) among the 20 most populated pyrotag OTUs that contained pyrotags aligning perfectly to only a single full-length OTU each. There were also a few pyrotag OTUs (one for each sequence similarity threshold) that did not match to any full-length sequences at 100% sequence similarity. The latter situation is not surprising given the limited number of full-length sequences in the data set and our requirement for perfect sequence similarity between a pyrotag and a full-length sequence.

Community analysis using pyrotag or full-length sequence data sets. Despite differences in the absolute numbers of OTUs obtained from the pyrotag and full-length sequence data sets, community analysis using Bray-Curtis similarity values and nMDS plots indicated that the conclusions derived from the two data sets were generally similar (Fig. 4). Liu et al. (50) demonstrated that short sequences (100 bp) were able to provide similar resolution for the comparison of communities as full-length 16S rRNA genes using UniFrac (a analysis of the distances between two environments based on the evolutionary history inferred from phylogenetic trees) in an *in silico* study of short sequences extracted from near-full-length 16S rRNA gene sequences. In addition, the removal of singletons in our data set had little effect on the results of the nMDS plots (Fig. 4), indicating that this beta diversity analysis may be robust toward the different sequencing approaches (Sanger versus NGS), as well as different processing of

sequences (with or without singletons). Egge et al. (51) also showed that different quality filtering procedures for pyrotag sequences of the V4 hypervariable region of the 18S rRNA genes of haptophytes had little effect on the clustering of communities on nMDS plots.

Conclusions. Next-generation sequencing approaches offer tremendous potential for conducting studies of microbial ecology, but only if sequence data can be translated into generally accepted, taxonomically meaningful data. To date, the accuracy of these approaches (i.e., their ability to provide accurate taxonomic characterization of a microbial assemblage) has not been adequately investigated (50, 51). Community analysis using pyrotags of approximately 100 bp of the V9 region of the 18S rRNA gene in the present study was shown to yield overall higher-level taxonomic information similar to that yielded by an analysis of full-length 18S rRNA gene sequences from the same samples and also provided similar patterns of community similarity among samples (e.g., using the Bray-Curtis similarity index) (Fig. 4). However, OTUs called using short pyrotag sequences generally matched to multiple OTUs called using full-length sequences, confusing the meaning of OTUs generated using these two approaches. We conclude that the V9 region of the 18S rRNA gene is insufficient for distinguishing between low-level taxonomic groups (i.e., species) and may be more appropriate for identifying higher taxonomic levels, such as phyla or classes. Our results indicate that at present, caution is warranted when one is attempting to use NGS sequence data sets of the V9 hypervariable region alone to characterize community richness across the whole domain of microbial eukaryotes. Microbial ecologists are moving toward the goal of adopting a molecular taxonomy for protists that incorporates molecular data into holistic assessments of microbial diversity and community composition (7), but vetting these approaches and the resulting conclusions continues to be essential.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation (OCE-0550829, MCB-0703159, MCB-0084231, OCE-1136818) and the Gordon and Betty Moore Foundation. Pyrosequencing was provided by the International Census of Marine Microbes (ICoMM) with financial support from a W. M. Keck Foundation award to the Marine Biological Laboratory at Woods Hole. The full-length sequencing work conducted by the U.S. Department of Energy Joint Genome Institute (JGI) is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

We thank the staff from JGI, including Jim Bristow for sponsoring the sequencing, Susannah Tringe and Ed Kirton for processing the sequences, and Ed Kirton for assembling the sequences. William C. Nelson assisted with early bioinformatics analysis of the full-length sequence data, especially the implementation of the chimera-checking algorithm in our informatics pipeline.

REFERENCES

1. **Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A.** 2012. Marine protistan diversity. Annu. Rev. Mar. Sci. **4:**467–493. http://dx.doi.org/10.1146/annurev-marine-120709-142802.
2. **Landry MR, Calbet A.** 2004. Microzooplankton production in the oceans. ICES J. Mar. Sci. **61:**501–507. http://dx.doi.org/10.1016/j.icesjms.2004.03.011.
3. **Sherr EB, Sherr BF.** 1994. Bacterivory and herbivory: key roles of phagotrophic protists in pelagic food webs. Microb. Ecol. **28:**223–235. http://dx.doi.org/10.1007/BF00166812.
4. **Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK.** 2012. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol. Evol. **27:**233–243. http://dx.doi.org/10.1016/j.tree.2011.11.010.
5. **Calbet A, Saiz E.** 2005. The ciliate-copepod link in marine ecosystems. Aquat. Microb. Ecol. **38:**157–167. http://dx.doi.org/10.3354/ame038157.
6. **Gilbert JA, Dupont CL.** 2011. Microbial metagenomics: beyond the genome. Annu. Rev. Mar. Sci. **3:**347–371. http://dx.doi.org/10.1146/annurev-marine-120709-142811.
7. **Caron DA.** 2013. Towards a molecular taxonomy for protists: benefits, risks, and applications in plankton ecology. J. Eukaryot. Microbiol. **60:**407–413. http://dx.doi.org/10.1111/jeu.12044.
8. **Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R.** 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. **41:**D597–D604. http://dx.doi.org/10.1093/nar/gks1160.
9. **Hebert PDN, Cywinska A, Ball SL, deWaard JR.** 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. **270:**313–321. http://dx.doi.org/10.1098/rspb.2002.2218.
10. **Zinger L, Gobet A, Pommier T.** 2012. Two decades of describing the unseen majority of aquatic microbial diversity. Mol. Ecol. **21:**1878–1896. http://dx.doi.org/10.1111/j.1365-294X.2011.05362.x.
11. **Epstein S, López-García P.** 2009. "Missing" protists: a molecular prospective, p 27–42. *In* Foissner W, Hawksworth D (ed), Protist diversity and geographical distribution, vol 8. Springer, Dordrecht, Netherlands.
12. **Dawson S, Hagen K.** 2009. Mapping the protistan 'rare biosphere.' J. Biol. **8:**105. http://dx.doi.org/10.1186/jbiol201.
13. **Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, Scanlan DJ, Worden AZ.** 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environ. Microbiol. **10:**3349–3365. http://dx.doi.org/10.1111/j.1462-2920.2008.01731.x.
14. **Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A, Valentin K, Pedrós-Alió C.** 2004. Phylogenetic and ecological analysis of novel marine stramenopiles. Appl. Environ. Microbiol. **70:**3528–3534. http://dx.doi.org/10.1128/AEM.70.6.3528-3534.2004.
15. **Countway PD, Gast RJ, Savai P, Caron DA.** 2005. Protistan diversity estimates based on 18S rDNA from seawater incubations in the western North Atlantic. J. Eukaryot. Microbiol. **52:**95–106. http://dx.doi.org/10.1111/j.1550-7408.2005.05202006.x.
16. **López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D.** 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature **409:**603–607. http://dx.doi.org/10.1038/35054537.
17. **Moon-van der Staay SY, De Wachter R, Vaulot D.** 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409:**607–610. http://dx.doi.org/10.1038/35054541.
18. **Schloss PD, Handelsman J.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. **71:**1501–1506. http://dx.doi.org/10.1128/AEM.71.3.1501-1506.2005.
19. **Nebel ME, Wild S, Holzhauser M, Hüttenberger L, Reitzig R, Sperber M, Stoeck T.** 2011. JAGUC—a software package for environmental diversity analyses. J. Bioinform. Comput. Biol. **9:**749–773. http://dx.doi.org/10.1142/S0219720011005781.
20. **Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD, Dennett MR, Moran DM, Jones AC.** 2009. Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. Appl. Environ. Microbiol. **75:**5797–5808. http://dx.doi.org/10.1128/AEM.00298-09.
21. **Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK.** 2010. Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. ISME J. **4:**1053–1059. http://dx.doi.org/10.1038/ismej.2010.26.
22. **Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J, Schlötterer C.** 2010. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol. Ecol. **19:**2908–2915. http://dx.doi.org/10.1111/j.1365-294X.2010.04669.x.
23. **Stoeck T, Bass D, Nebel M, Christen R, Meredith D.** 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly com-

plex eukaryotic community in marine anoxic water. Mol. Ecol. **19:**21–31. http://dx.doi.org/10.1111/j.1365-294X.2009.04480.x.
24. **Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS.** 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. http://dx.doi.org/10.1128/AEM.00592-09.
25. **Huse S, Huber J, Morrison H, Sogin M, Welch D.** 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. **8:**R143. http://dx.doi.org/10.1186/gb-2007-8-7-r143.
26. **Bachy C, Dolan JR, López-García P, Deschamps P, Moreira D.** 2013. Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. ISME J. **7:**244–255. http://dx.doi.org/10.1038/ismej.2012.106.
27. **Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C, Holder M, Taylor GT, Suarez P, Varela R, Epstein S.** 2011. Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISME J. **5:**1344–1356. http://dx.doi.org/10.1038/ismej.2011.6.
28. **Santoferrara LF, Grattepanche J-D, Katz LA, McManus GB.** 2014. Pyrosequencing for assessing diversity of eukaryotic microbes: analysis of data on marine planktonic ciliates and comparison with traditional methods. Environ. Microbiol. http://dx.doi.org/10.1111/1462-2920.12380.
29. **Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM.** 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One **4:**e6372. http://dx.doi.org/10.1371/journal.pone.0006372.
30. **Schloss PD, Gevers D, Westcott SL.** 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One **6:**e27310. http://dx.doi.org/10.1371/journal.pone.0027310.
31. **Huse SM, Welch DM, Morrison HG, Sogin ML.** 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. **12:**1889–1898. http://dx.doi.org/10.1111/j.1462-2920.2010.02193.x.
32. **Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R.** 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27:**2194–2200. http://dx.doi.org/10.1093/bioinformatics/btr381.
33. **Countway PD, Vigil PD, Schnetzer A, Moorthi SD, Caron DA.** 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. http://dx.doi.org/10.4319/lo.2010.55.6.2381.
34. **Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ.** 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. http://dx.doi.org/10.1128/AEM.71.12.7724-7736.2005.
35. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.
36. **Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T.** 2011. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environ. Microbiol. **13:**340–349. http://dx.doi.org/10.1111/j.1462-2920.2010.02332.x.
37. **Kunin V, Engelbrektson A, Ochman H, Hugenholtz P.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118–123. http://dx.doi.org/10.1111/j.1462-2920.2009.02051.x.
38. **Dunthorn M, Klier J, Bunge J, Stoeck T.** 2012. Comparing the hypervariable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. J. Eukaryot. Microbiol. **59:**185–187. http://dx.doi.org/10.1111/j.1550-7408.2011.00602.x.
39. **Nebel M, Pfabel C, Stock A, Dunthorn M, Stoeck T.** 2011. Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. Environ. Microbiol. Rep. **3:**154–158. http://dx.doi.org/10.1111/j.1758-2229.2010.00200.x.
40. **Chao A.** 1984. Nonparametric estimation of the number of classes in a population. Scand. J. Stat. **11:**265–270.
41. **Dickie IA.** 2010. Insidious effects of sequencing errors on perceived diversity in molecular surveys. New Phytol. **188:**916–918. http://dx.doi.org/10.1111/j.1469-8137.2010.03473.x.
42. **Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS.** 2013. Robust estimation of microbial diversity in theory and in practice. ISME J. **7:**1092–1101. http://dx.doi.org/10.1038/ismej.2013.10.
43. **Monchy S, Grattepanche J-D, Breton E, Meloni D, Sanciu G, Chabé M, Delhaes L, Viscogliosi E, Sime-Ngando T, Christaki U.** 2012. Microplanktonic community structure in a coastal system relative to a *Phaeocystis* bloom inferred from morphological and tag pyrosequencing methods. PLoS One **7:**e39924. http://dx.doi.org/10.1371/journal.pone.0039924.
44. **Wolf C, Frickenhaus S, Kilias ES, Peeken I, Metfies K.** 2013. Regional variability in eukaryotic protist communities in the Amundsen Sea. Antarctic Sci. **11:**1–11. http://dx.doi.org/10.1017/S0954102013000229.
45. **Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J.** 2010. Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol. Ecol. **19**(Suppl 1):S32–S40. http://dx.doi.org/10.1111/j.1365-294X.2009.04478.x.
46. **Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB.** 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ. Microbiol. **11:**1292–1302. http://dx.doi.org/10.1111/j.1462-2920.2008.01857.x.
47. **Liu H, Probert I, Uitz J, Claustre H, Aris-Brosou S, Frada M, Not F, de Vargas C.** 2009. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proc. Natl. Acad. Sci. U. S. A. **106:**12803–12808. http://dx.doi.org/10.1073/pnas.0905841106.
48. **Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. **103:**12115–12120. http://dx.doi.org/10.1073/pnas.0605127103.
49. **Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML.** 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. **4:**e1000255. http://dx.doi.org/10.1371/journal.pgen.1000255.
50. **Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**e120. http://dx.doi.org/10.1093/nar/gkm541.
51. **Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B.** 2013. 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. PLoS One **8:**e74371. http://dx.doi.org/10.1371/journal.pone.0074371.