ORIGINAL ARTICLE

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

Charles Bachy[1], John R Dolan[2], Purificación López-García[1], Philippe Deschamps[1] and David Moreira[1]

[1]*Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, Orsay, France and* [2]*Université Pierre et Marie Curie, CNRS UMR 7093, Laboratoire d'Océanographie de Villefranche, Marine Microbial Ecology, Station Zoologique, Villefranche-sur-Mer, France*

Deep-sequencing technologies are becoming nearly routine to describe microbial community composition in environmental samples. The 18S ribosomal DNA (rDNA) pyrosequencing has revealed a vast diversity of infrequent sequences, leading to the proposition of the existence of an extremely diverse microbial 'rare biosphere'. Although rare microbes no doubt exist, critical views suggest that many rare sequences may actually be artifacts. However, information about how diversity revealed by molecular methods relates to that revealed by classical morphology approaches is practically nonexistent. To address this issue, we used different approaches to assess the diversity of tintinnid ciliates, a species-rich group in which species can be easily distinguished morphologically. We studied two Mediterranean marine samples with different patterns of tintinnid diversity. We estimated tintinnid diversity in these samples employing morphological observations and both classical cloning and sequencing and pyrosequencing of two different markers, the 18S rDNA and the internal transcribed spacer (ITS) regions, applying a variety of computational approaches currently used to analyze pyrosequence reads. We found that both molecular approaches were efficient in detecting the tintinnid species observed by microscopy and revealed similar phylogenetic structures of the tintinnid community at the species level. However, depending on the method used to analyze the pyrosequencing results, we observed discrepancies with the morphology-based assessments up to several orders of magnitude. In several cases, the inferred number of operational taxonomic units (OTUs) largely exceeded the total number of tintinnid cells in the samples. Such inflation of the OTU numbers corresponded to 'rare biosphere' taxa, composed largely of artifacts. Our results suggest that a careful and rigorous analysis of pyrosequencing data sets, including data denoising and sequence clustering with well-adjusted parameters, is necessary to accurately describe microbial biodiversity using this molecular approach.

Introduction

During the past decade, studies of the diversity of unicellular eukaryotes, or protists, have been deeply affected by the advent of cultivation-independent molecular methods. Initially, this approach was based on the analysis of 18S ribosomal RNA (rRNA) genes using PCR amplification, cloning and Sanger sequencing. When these approaches were first applied to study marine plankton, they revealed a wide and unsuspected protist diversity (López-García *et al.*, 2001; Moon-van der Staay *et al.*, 2001; Not *et al.*, 2007a, b; Guillou *et al.*, 2008; Massana and Pedrós-Alió, 2008). More recently, the development of 454 pyrosequencing, which produces thousands of 18S ribosomal DNA (rDNA) amplicon sequence reads from a single sample, has revolutionized the study of natural microbial communities via a so-called 'deep-sequencing' approach (Margulies *et al.*, 2005; Sogin

et al., 2006). In particular, this has revealed the existence of large collections of lineages represented by only a few sequences, which have been interpreted as evidence of a rare microbial biosphere (Pedrós-Alió, 2007). Furthermore, very high estimates of eukaryotic species richness have been established in this way, notably in marine waters with several thousands of different operational taxonomic units (OTUs) detected in just a few liters (Brown et al., 2009; Edgcomb et al., 2011). The 18S rDNA amplicon pyrosequencing has become a widespread approach for protist diversity studies.

The sequence length obtained by 454 pyrosequencing, originally of ~100 bp, now exceeds 400 bp (Titanium chemistry, 454 Life Sciences, Basel, Switzerland), allowing a more precise taxonomic classification of reads. However, multiple sources of bias have been identified, including pyrosequencing errors (Huse et al., 2007; Kunin et al., 2010), poor alignment quality (Sun et al., 2009; Schloss, 2010), inappropriate clustering approaches (Huse et al., 2010) or inconsistent results from different targeted 18S rDNA regions (Stoeck et al., 2010). These biases can have profound effects on diversity estimation, most often leading to the overestimation of OTU richness. Algorithms to remove pyrosequencing errors have been developed to reduce such overestimations. This is the case of AmpliconNoise (Quince et al., 2011), based on flowgram clustering, or the single-linkage preclustering (SLP) approach (Huse et al., 2010), based on sequence similarity analysis. Nevertheless, only a few studies have compared the results of using different in silico analytical approaches when assessing species diversity in natural samples using pyrosequencing reads (see, for example, Quince et al., 2009; Huse et al., 2010; Bik et al., 2012).

In spite of a rich and long history of taxonomic description of microbial eukaryotes, accurate comparative studies between morphological and molecular approaches in natural environments remain rare. One exception is the recent study of the overall diversity of eukaryotes in a freshwater lake, which demonstrated the efficiency of 18S rDNA pyrosequencing in detecting rare species but revealed strong incongruencies among taxa with regard to their relative abundances (Medinger et al., 2010). In fact, to be reliable and informative, environmental 18S rDNA sequence data analysis requires knowledge about how genotypic and phenotypic diversity relate (Caron et al., 2009). Consequently, we need to know how diversity defined by classical morphology approaches relates to that based on molecular methods, either conventional (cloning/Sanger sequencing) or recent deep-sequencing technologies (for example, pyrosequencing). Employing these different approaches on the same samples, using morphology-based species assessments as a control, could be a robust approach to assess the magnitude of the microbial rare biosphere indicated by pyrosequencing of 18S rDNA fragments. A robust comparison of these distinct approaches requires their application on a biological model group that is abundant, diverse and easily distinguishable by morphology.

The species-rich order Tintinnida represents one of the most morphologically diverse groups of planktonic protists. Tintinnids are commonly found in marine surface waters of the globe, in concentrations generally ranging from 10^1 to 10^4 cells per liter (Pierce and Turner, 1993). By examining specimens in plankton settling chambers using an inverted microscope, they can be easily identified based on their species-specific secreted shell, the lorica (see, for example, Dolan et al., 2009). Several studies based on tintinnid single-cell identification, 18S rRNA gene and internal transcribed spacer (ITS) PCR amplification and sequencing have recently been carried out, providing a comprehensive phylogenetic framework that allows the identification of environmental sequences at the species level (Snoeyenbos-West et al., 2002, Strüder-Kypke and Lynn, 2008; Bachy et al., 2012; Saccà et al., 2012; Santoferrara et al., 2012). Thus, we applied a taxon-specific approach focused on marine tintinnid ciliates to assess their diversity in aliquots of the same samples using microscopic analysis (species identified based on classical morphological descriptions) and both cloning/Sanger sequencing and pyrosequencing analyses of 18S rDNA and ITS regions. Furthermore, we also examined different pyrosequencing data processing approaches to estimate species richness.

Materials and methods

Sample collection

Seawater samples were collected with Niskin bottles at the deep chlorophyll maximum depth from two sites in Mediterranean Sea: the Bay of Villefranche-sur-Mer (Ligurian Sea, sample VilleFr-43) at the site 'Point B + ' (43°41′00″N, 7°18′44″E; 50 m depth) on 18 November 2009; and the Ionian Sea (sample Ioni-7; 36°28′58″N, 15°38′58″E; 77 m depth) on 1 October 2010, during the Marmara 2010 cruise. A total of 20 l were collected in the coastal station of Villefranche and 168 l in the Ionian Sea offshore station. After sampling, plankton samples were immediately prefiltered through a 200-μm Nitex screen and then subdivided into two equal volumes. For the VilleFr-43 sample (Dynamic Aqua-Supply Ltd, Surrey, BC, Canada), one subsample was immediately preserved with Lugol iodine solution (2% final concentration) and then concentrated via serial gravity sedimentation (that is, 10 to 2 liters and then to 200 ml) to a final volume of 20 ml for taxonomic identification by morphological observation (see below). The second subsample was filtered onto a 5-μm pore-diameter polycarbonate filter (Millipore, Billerica, MA, USA) to concentrate all biomass. The filter was conserved in absolute ethanol at − 20 °C

following a protocol already tested for plankton samples (López-García et al., 2003; Bachy et al., 2011). For the Ioni-7 sample, the two subsamples were concentrated by gravity filtering through a 20-µm Nitex filter to a final volume of 4 ml. One subsample was fixed with Lugol solution (Sigma, St Louis, MO, USA) and the other subsample was fixed in ethanol (80% final concentration) for further processing. The cells fixed in ethanol were then filtered onto a 5-µm pore-diameter filter to concentrate the biomass and immediately processed for DNA extraction (see below).

Morphological observation
For each subsample fixed in Lugol solution, the entire concentrated volume was examined to exhaustion at ×200 magnification using an inverted microscope (Olympus IX51, Rungis, France) in 2 ml aliquots settled in sedimentation chambers. Taxonomic assignation of every single cell of tintinnid species was determined based on lorica morphology following standard taxonomic monographs (Kofoid and Campbell, 1929, 1939; Marshall, 1969). Empty loricae were included in the enumerations.

Generation of clone libraries and sequencing
Total DNA was purified from filters cut in small pieces using the PowerSoil DNA Isolation Kit (MoBio, Carlsbad, CA, USA). Nucleic acids were resuspended in 10 mM Tris-HCl, pH 8. Near-full-length 18S rDNA fragments were amplified by PCR using the tintinnid-specific forward primer 18S-Tin3F (5′-GCGGTATTTATTAGATAWCAGCC-3′; Bachy et al., 2012) and the eukaryotic-specific reverse primer 1498R (5′-CACCTACGCAAACCTT GTTA-3′; López-García et al., 2003). PCR was carried out under the following conditions: 35 cycles (denaturation at 94 °C for 15 s, annealing at 52 °C for 30 s, extension at 72 °C for 2 min) preceded by 3 min denaturation at 94 °C and followed by 15 min extension at 72 °C. We also targeted longer DNA fragments covering the near-full-length 18S rRNA gene, the ITSs 1 and 2, the 5.8S rDNA gene and a partial 28S rDNA fragment under similar PCR conditions as described above. These fragments were amplified with the tintinnid-specific forward primer 18S-Tin3F and the tintinnid-specific reverse primer 28S-TinR1 (5′-TGGTGCACTAGTATCAA AGT-3′). This primer set yielded an amplicon size of ∼2200 bp. Clone libraries were constructed using the Topo TA cloning system (Invitrogen, Carlsbad, CA, USA) following the instructions provided by the manufacturer. Positive inserts of expected size were selected from each library and Sanger-sequenced with the forward primer. We generated a total of 200 high-quality partial sequences (>700 bp) for each sample, covering the 18S rDNA regions also targeted by the primers used for pyrosequencing analysis. These clone sequences served for a preliminary phylogenetic analysis and the identification of

OTUs (defined here as clusters of sequences having ⩾99% identity). For each library, we completely sequenced at least one clone per OTU to obtain complete sequences representative of the whole taxonomic diversity found, preferentially the ones that also encompassed the ITS, 5.8S and 28S rDNA. A total of 116 ciliate sequences (among which 100 were tintinnid sequences) formed a reference data set that provided a phylogenetic framework for the attribution of environmental clones. From all the tintinnid OTUs detected by our stringent criterion (⩾99% sequence identity), 25 and 17 OTUs were found in the coastal and offshore samples, respectively. In addition, two OTUs represented by one sequence each were excluded as non-tintinnids from the VilleFr-43 sample as they belonged to strombiliid ciliate species. The new tintinnid sequences formed with the previous reference data set the '18S rDNA data set' (that is, 100 tintinnid reference sequences plus 42 new clone sequences attributed to tintinnids). The sequences for which the ITS region was also sequenced formed the 'ITS data set', comprising a smaller number of 43 sequences (22 tintinnid reference sequences and 21 new clone sequences attributed to tintinnids). The 42 new complete sequences were submitted to GenBank (accession numbers JX567350 – JX567503).

Clustering of clone sequences into OTUs
To affiliate clone sequences to different tintinnid taxa, they were compared by BLAST (Altschul et al., 1997) against the reference data set. A fasta file with all the high-quality partial 18S rDNA clone sequences was created and submitted to the module NDIST of AmpliconNoise (Quince et al., 2011) to generate a distance matrix of all pairwise sequence combinations by Needleman–Wunsch pairwise alignment (with default parameters). A first distance matrix was generated with an alignment of all available positions (∼840 bp), and a second one with only the region covering positions targeted in the 18S rDNA pyrosequence read analysis (see below). The resulting matrices were used as input for Mothur (Schloss et al., 2009) in order to group clone sequences by average linkage clustering at different sequence identity levels.

Generation of amplicon libraries and pyrosequencing
Two sets of primers were designed for this study to specifically target variable regions of the tintinnid 18S rRNA gene and ITS region. The primers were adapted for 454 pyrosequencing following the manufacturer's recommendations: addition of the forward and reverse fusion primers (A/B) and also of a 10-bp molecular identifier (MID) tag inserted between the adapter A and the specific primer to distinguish the two samples. The primers targeted 18S rDNAs from all tintinnid families sequenced to date. The first set of primers amplified a 330–340-bp fragment of the 18S rRNA gene including the

hypervariable V4 region, whereas the second one amplified a 250–260-bp fragment of the ITS region including partial ITS1, complete 5.8S and partial ITS2. PCR reactions were carried out using 2 µl of environmental DNA in a volume of 20 µl Platinum HIFI reaction mix (Invitrogen) containing the tintinnid-specific 18S rDNA primers Tin454-18SFw (5′-GTAGTGACAAGAAATAAYGG-3′) and Tin454-18SRev (5′-CAACCCCTCARACCAACAG-3′). To amplify the ITS region, PCR reactions were carried out using the tintinnid-specific ITS primers Tin454-ITSFw (5′-CAATYTGTTGCAGRGCGMAAGC-3′) and Tin454-ITSRev (5′-AGCAATAGAAGGGCATCTA-3′). The PCR reactions were done under the following conditions: 25 cycles (denaturation at 94 °C for 15 s, annealing at 52 °C for 30 s, extension at 72 °C for 1 min) preceded by 3 min denaturation at 94 °C and followed by 15 min extension at 72 °C. We carried out only 25 PCR cycles in order to minimize the generation of recombinant PCR products and compensated for the lower yield by pooling the products of 10 independent PCRs. DNA was purified for each amplicon from each sample and mixed in 1:1:1:1 proportions before being 454-pyrosequenced from the A adapter on a 454 GS FLX sequencer using Titanium protocols (454 Life Sciences) and supplies (Eurofins MWG Operon, Ebersberg, Germany). Sequences are available at the Sequence Read Archive (ERP001739).

Low-quality and phylogenetic filtering of pyrosequencing reads
As a first step of filtering, only pyrosequencing reads having exact forward and reverse primer sequences and with no ambiguous nucleotides were kept. The remaining reads were then trimmed after eliminating the primer sequences. Then, they were compared by BLAST (Altschul *et al.*, 1997) against the '18S rDNA' and 'ITS' data sets (see above) to determine their preliminary taxonomic affiliation. Sequences with a similarity <98% with *bona fide* tintinnid sequences were extracted and their affiliation to the Tintinnida was checked by phylogenetic analysis. These reads were aligned using MAFFT (Katoh *et al.*, 2002) and an approximately maximum likelihood (ML) tree was constructed using FastTree (Price *et al.*, 2009) with choreotrich and oligotrich sequences as outgroup for the '18S rDNA' data set and with spirotrich sequences as outgroup for the 'ITS' data set. All 18S rDNA and ITS Ionian Sea reads did belong to tintinnid species, whereas 804 18S rDNA Villefranche reads appeared to belong to strombiliid species, a sister clade of tintinnids, and were therefore discarded. The rest of Villefranche sequences corresponded to tintinnid species. No chimera was detected with the Chimera Slayer algorithm (Haas *et al.*, 2011) implemented in Mothur (Schloss, 2010) using the '18S rDNA' and 'ITS' curated data sets as references. To avoid potential biased comparisons because of variable data set

sizes (Gihring *et al.*, 2012), we randomly selected within each sample the number of pyrosequence reads contained in the smaller data set (10 504 and 8734 reads in VilleFr-43 and Ioni-7, respectively). The trimmed, phylogenetically checked and size-equivalent pyrosequence read libraries served as starting point for subsequent analyses. With the corresponding read identifiers, we also prepared flowgram data sets for AmpliconNoise, which needs them as input files.

Clustering of pyrosequence reads into OTUs
We applied several ways of clustering the pyrosequencing reads. First, with the quality- and taxonomy-filtered pyrosequence read data sets (see above), we used MAFFT (Katoh *et al.*, 2002) with default parameters for multiple sequence alignment and distance matrix generation. We then used Mothur (Schloss *et al.*, 2009) to estimate OTUs richness using the average linkage clustering option. Another clustering process was to apply the NDIST module of AmpliconNoise (Quince *et al.*, 2011) with default parameters to generate the distance matrix of all Needleman–Wunsch pairwise sequence alignments. Then, the average linkage option in the AmpliconNoise FCluster module was used to cluster reads into OTUs. As an alternative, we also applied the clustering procedure implemented in Usearch (Edgar, 2010) with default parameters at different identity levels. The first steps of filtering (see above) were done to diminish the per nucleotide error rate of sequences (estimated at ∼0.25% for pyrosequencing; see Huse *et al.*, 2007). We then used different additional approaches to correct the remaining PCR and pyrosequencing errors. We first analyzed the filtered pyrosequence reads with the SLP algorithm as detailed in Huse *et al.* (2010), with a pairwise distance <0.005 used as threshold (equivalent to a single-nucleotide difference in the 18S rDNA and ITS regions). We then used Mothur to create OTUs using the average linkage option. Alternatively, we ran AmpliconNoise with parameters identified as efficient to minimize the number of spurious OTUs for pyrosequencing: −r 0.005 for FCluster, −s 60.0 and −c 0.005 for PyroNoiseM and −s 10.0 and −c 0.08 for SeqNoise (Quince *et al.*, 2011). The sequences passing this filter were clustered using pairwise alignments and average linkage clustering as described above. In addition, we ran CD-HIT-OTU, a recent clustering algorithm for 18S rDNA pyrosequence data, with default parameters (including −e 0.0025) on the OTU finder web server (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/cd-hit-otu/). It should be noted that before the last step of OTU clustering with this algorithm, the clusters of only 1 or 2 sequences are considered as noise and therefore removed. For OTUs inferred using all the different methods, rarefaction analyses were computed with Mothur (Schloss *et al.*, 2009).

248

Phylogenetic analyses

All representative clone sequences and 18S rDNA reads selected after filtering of pyrosequences by AmpliconNoise, together with their most similar sequences identified by BLAST (Altschul *et al.*, 1997) in the reference data set, were aligned using MAFFT. The multiple sequence alignment was then manually edited using the program ED from the MUST package (Philippe, 1993). An approximate ML tree was constructed with FastTree for all these sequences in order to choose one sequence (the longest one) as representative for each OTU for further phylogenetic analyses. The selected data set was then analyzed by ML using TREEFINDER (Jobb *et al.*, 2004) applying a general time reversible model of sequence evolution with a 4-rate categories Gamma law to accommodate for among-site rate variation. ML bootstrap proportions were inferred using 1000 replicates.

Results

We studied the tintinnid ciliate community in two deep chlorophyll maximum marine plankton samples from a coastal (Villefranche) and an offshore (Ionian Sea) location. Each sample was divided in two equal parts, the first for morphological identification of tintinnid cells and the second for both 18S rRNA gene library construction and Sanger sequencing and direct pyrosequencing of 18S rRNA gene and ITS amplicons using tintinnid-specific primers.

Morphology and sequence data sets

We identified 673 tintinnid cells in the plankton subsamples subjected to microscopic examination (465 in Villefranche VilleFr-43 sample and 208 in the Ionian Sea Ioni-7 sample). From plankton subsamples devoted to molecular analyses, after DNA extraction and PCR amplification, we generated partial 18S rDNA sequences of 400 clones (200 from each locality) using Sanger sequencing and a total of 76 206 pyrosequence reads (termed here as 'pyroreads') for both samples and molecular markers. Quality and phylogenetic filtering of these pyroreads yielded 50 817 sequences (10 504 and 8734 18S rDNA reads and 17 378 and 14 201 ITS reads for Villefranche and the Ionian Sea, respectively). To allow direct comparison between the two fragments targeted by pyrosequencing (18S rDNA and ITS) within each sample, we limited the pyroread data sets to the same size by random selection of sequences within each set (that is, 10 504 18S rDNA and ITS reads for VilleFr-43, and 8734 18S rDNA and ITS reads for Ioni-7; see Table 1).

In the coastal sample VilleFr-43, 21 morphospecies were found, largely dominated by the family Tintinnidae, with *Salpingella faurei* and

Table 1 Number of tintinnid cells, clone sequences and pyroreads retained for comparison from the VilleFr-43 and Ioni-7 samples

	VilleFr-43	*Ioni-7*
Cells	465	208
Clone sequences	200	200
18S rDNA pyroreads	10 504	8734
ITS pyroreads	10 504	8734

Abbreviations: ITS, internal transcribed spacer; rDNA, ribosomal DNA.

Steenstrupiella steenstrupii representing 54% of the total number of cells (Figure 1). The corresponding 18S rDNA clone library was also dominated by Tintinnidae (37.3% of clones), followed by Undellidae (26%) and Codonellidae (18%). Similarly, pyroreads affiliating to Tintinnidae were the most abundant in both the 18S rDNA and ITS data sets, where they represented 64.6% and 43.2% of pyroreads, respectively. In the offshore sample Ioni-7, a more diverse community of 36 morphospecies was found, with the community more evenly distributed compared with the Villefranche assemblage. Individuals identified as belonging to the families Tintinnidae (23%), Undellidae (22%) and Eutintinnidae (18%) co-dominated the tintinnid community. In contrast, the proportions of tintinnid families in the clone library was different from the morphological observations, with a majority of 18S rDNA sequences attributed to Undellidae (47%), followed by Tintinnidae (22%) and Codonellidae and Codonellopsidae (12% each). Pyrosequencing of the 18S rDNA region yielded a majority of reads identified as Tintinnidae (51%) and Undellidae (32%), whereas the ITS region produced a majority of sequences affiliated to Undellidae (51%) and then Tintinnidae (16%; Figure 1). Overall, the distribution of 18S rDNA and ITS pyroreads related well with that of clone sequences and with the diversity of tintinnids based on microscopic observations at the family level. However, 18S rDNA and ITS pyroread data sets differed significantly from each other in both the coastal and off-shore samples (χ^2 test, *P*-value < 0.001), showing the difficulties in comparing diversity estimates when different markers are used.

The microscopy-based determination of species diversity showed that both the Villefranche and Ionian Sea samples contained diverse tintinnid communities. The differences observed between the two samples were probably because of the different geographical origins and environmental conditions and sampling periods (collected at > 1000 km of distance and at a 1-year interval). In particular, despite the fact that the number of cells in the Ioni-7 sample was half of that found in the VilleFr-43 sample, the Ionian sample harbored a more species-rich community (Supplementary Table S1). Furthermore, the coastal VilleFr-43 sample was
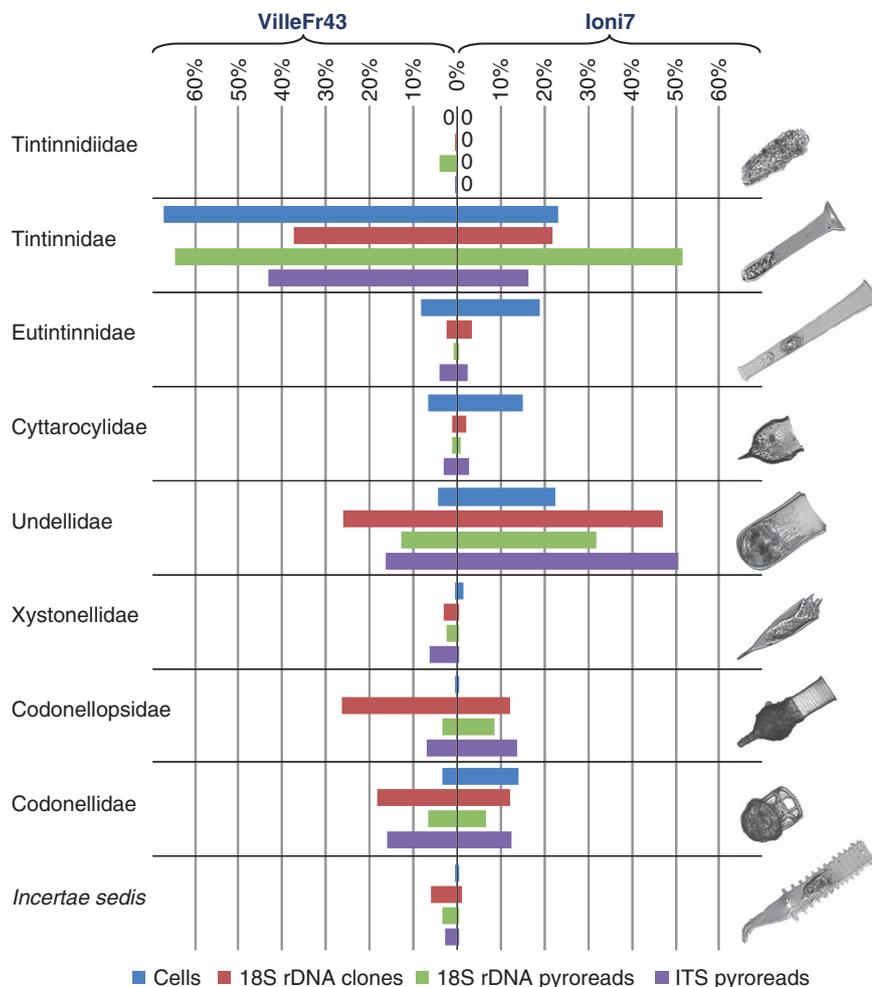
Figure 1 Frequencies of the different tintinnid families found in the VilleFr-43 and Ioni-7 samples by morphology (cells), 18S rDNA library sequencing (clones) and pyrosequencing of 18S rDNA and ITS regions. Sequences were assigned to families by BLAST searches against a curated tintinnid sequence database. Micrographs on the right show representative morphologies encountered in each family. Complete absence of a family in a sample is indicated by 0.

dominated by a few forms, with >50% of the tintinnid community composed of two species, whereas the Ioni-7 sample showed a more even species distribution, with the five most abundant species just reaching 42% of the total community. These differences in the tintinnid communities in terms of cell abundance and species compositions were in agreement with previous studies (see, for example, Dolan *et al.*, 2002, 2006). Thus, the observed differences between the two samples provided us with two different species distribution models to study the performance of different molecular approaches to assess OTU composition.

Species richness estimation
To estimate species richness from sequence data, we defined tintinnid OTUs applying a sequence similarity threshold of ≥99%, which was based on the 18S rDNA nucleotide differences observed between distinct tintinnid morphospecies (Bachy *et al.*, 2012;

and Supplementary Figure S1). A lower sequence similarity threshold (that is, <99%) would entail the mixing of not only different species but even different genera, such as *Dictyocysta* and *Codonella*, *Amphorella* and *Amphorellopsis*, or *Tintinnopsis* and *Climacocylis*. Moreover, considering both full-length 18S rDNA sequences or exclusively the short ~340-bp fragment targeted by our pyrosequencing analysis, the use of ≥99% sequence identity provided similar species-level resolution. In fact, applying this threshold on the Sanger sequences from the clone libraries, we retrieved comparable OTU numbers for the complete 18S rDNA sequences and for the fragments corresponding to the region chosen for pyrosequencing (25 and 22 OTUs in VilleFr-43 and 17 and 22 OTUs in Ioni-7, respectively). The rarefaction curves corresponding to these clone libraries were close to saturation (Supplementary Figure S2). In the case of VilleFr-43, not only the number of OTUs but also the species abundance distributions (with 32–40% of rare

OTUs, namely, those representing ⩽1% of the total number of sequences in a data set) were quite close to the ones described by morphological identification. In contrast, the OTUs identified in Ioni-7 represented only ∼50% of the 36 species identified by morphological examination. However, the percentage of rare OTUs in this sample ranged from 35% to 52% and was similar to the 42% observed by morphological identification.

A variety of approaches have been developed to assess species richness based on pyroreads (see review in Zinger *et al.*, 2012). We applied several of these methods in order to compare the results among them and with the classical 18S rDNA clone libraries/Sanger sequencing and traditional morphological analysis (Table 2 and Supplementary Figure S3). As explained above, we first defined an adequate sequence similarity threshold (99%) to identify OTUs in the 18S rDNA pyroread data set. According to pairwise sequence comparisons of reference sequences belonging to identified species, and sequences from clone libraries, the 18S rDNA and ITS regions targeted by our pyrosequencing analysis showed similar divergence rates, comparable to those of the complete 18S rDNA sequences (Supplementary Figures S1 and S4). Therefore, we also used the 99% sequence identity threshold to identify OTUs with both the 18S rDNA and ITS pyroreads. Nevertheless, we have shown in previous work that phylogenetically distinct lineages may

Table 2 Number of tintinnid species and total and rare OTUs identified using morphological and molecular methods

	VilleFr-43		Ioni-7	
	Total	*Rare*	*Total*	*Rare*
Cell identification				
Morphotypes	21	9 (42.9%)	36	15 (41.7%)
Clone sequences				
Complete sequences	25	10 (40.0%)	17	6 (35.3%)
Partial 'pyroread' region	22	7 (31.8%)	22	11 (50.0%)
18S rDNA pyroreads				
MSA	1494	1474 (98.7%)	1089	1073 (98.5%)
PW	1159	1144 (98.7%)	854	840 (98.4%)
Usearch	518	498 (96.1%)	400	375 (93.8%)
SLP-PW	1848	1833 (99.2%)	1383	1372 (99.2%)
CD-HIT-OTU	27	15 (55.6%)	23	13 (56.5%)
AmpliconNoise	22	8 (36.4%)	12	4 (33.3%)
ITS pyroreads				
MSA	1036	1022 (98.6%)	663	653 (98.5%)
PW	832	818 (98.3%)	519	510 (98.3%)
Usearch	544	525 (96.5%)	259	251 (96.9%)
SLP-PW	166	156 (94.0%)	616	608 (98.7%)
CD-HIT-OTU	23	15 (65.2%)	18	12 (66.7%)
AmpliconNoise	19	11 (57.9%)	7	1 (14.3%)

Abbreviations: ITS, internal transcribed spacer; MSA, multiple sequence alignment; OTU, operational taxonomic unit; PW, pairwise alignment; rDNA, ribosomal DNA; SLP-PW, single-linkage preclustering and pairwise alignment. For rare OTUs, total numbers and relative proportions (%) are indicated.

contain identical ITS region sequences (Bachy *et al.*, 2012); and hence it should be noted that the ITS pyroreads were not as accurate as the 18S rDNA ones to discriminate different species (see Supplementary Figure S1). This may reflect a greater selective constraint at the ITS than that experienced by the targeted 18S rDNA variable region V4, which is characterized by particularly rapid rates of evolution within eukaryotes. With the 99% sequence identity threshold established, we tested several sequence alignment, clustering and denoising methods on the 18S rDNA and ITS pyroread data sets (see Materials and methods). Surprisingly, depending on the analytical method and molecular marker used, 18S rDNA or ITS, the number of OTUs varied from only 7 up to 1848. Accordingly, the proportion of rare OTUs represented by few reads showed extreme variations, between 14% and 99% of the total OTUs. Table 2 shows that the number of OTUs detected using the common method of multiple sequence alignment or the more stringent pairwise alignment and average-linkage exceeded the number of expected species (based on unambiguous morphological taxon identification) by one or two orders of magnitude. The greedy algorithm for clustering of representative sequences implemented in Usearch (Edgar, 2010) appeared to be more efficient, but the number of OTUs still exceeded the number of expected species by one order of magnitude. We can reject these diversity estimates with some confidence, not only because they overestimated the number of species observed based on microscope examination but also because they exceeded, sometimes by twofold or more, the total number of tintinnid cells observed in twin volumes of samples.

To avoid this kind of artificial inflation, we applied several approaches recently developed to remove PCR and pyrosequencing errors, including sequence preclustering and the more complex pipelines implemented in CD-HIT-OTU (Wu *et al.*, 2011) and AmpliconNoise (Quince *et al.*, 2011). The agglomerative approach applying SLP before the classical clustering step also overestimated species richness, especially for the 18S rDNA pyroreads (Table 2). This could be because of the fact that the preclustering step at 98% sequence identity usually used for SLP (Huse *et al.*, 2010) was too large compared with the threshold that we empirically determined for tintinnids. Nevertheless, applying a preclustering step with a 99.5% similarity threshold was not sufficient to denoise the data efficiently (between 166 and 1848 OTUs were defined depending on the marker, Table 2). Compared with the other approaches that we tested, the denoising and clustering procedures implemented in CD-HIT-OTU and AmpliconNoise were the only ones able to produce estimates coherent with those obtained from classical clone library/Sanger sequencing and, especially, with the morphological identification. These methods inferred a total number of OTUs of

~25 for the 18S rDNA region and ~20 for the ITS region in VilleFr-43 and 12–23 for the 18S rDNA region and 7–18 for the ITS region in Ioni-7. Whereas the other methods produced estimates of rare OTUs ranging from 94% to 99.2% of all OTUs, they accounted for only 36.4–55.6% and 33.3–56.5% of all OTUs inferred by AmpliconNoise and CD-HIT-OTU in VilleFr-43 and Ioni-7, respectively (Figure 1). As AmpliconNoise produced reasonable species estimates by taking into account all sequences (that is, it does not eliminate the clusters with only 1 or 2 sequences as CH-HIT-OTU does), we kept the OTUs identified from 18S rDNA pyroreads by this method for subsequent analyses.

Phylogenetic analyses
We built a phylogenetic tree based on sequences representative of all OTUs identified in clone libraries and pyrosequencing data together with their closest described relatives available in public databases (Figure 2). The resulting tree was strongly supported at the genus level, giving us the opportunity to compare at this fine taxonomic rank the abundance patterns generated by both molecular approaches and the morphological analysis. From a qualitative point of view, both the cloning and pyrosequencing approaches detected the majority of families and genera found using microscopic observations, although a slightly larger diversity was detected by cloning/Sanger sequencing than by pyrosequencing. Surprisingly, cloning and pyrosequencing approaches seemed thus to be almost equally efficient for detecting the most frequent species observed in our samples (*Salpingella*-, *Undella*-, *Codonella*- and *Dictyocysta*-related species) and even those present in low abundance (for example, those of families Xystonellidae or Codonellopsidae). Moreover, whereas no *Tintinnidium* spp. was found in the coastal sample (VilleFr-43) examined microscopically, 18S rDNA sequences affiliated to this genus were detected (4% of pyroreads and 0.5% of clones). In contrast, *Rhabdonella* sp. and *Steenstrupiella* sp. specimens observed in the sample Ioni-7 were not detected with any of the molecular approaches. The comparison of the two molecular approaches revealed that >50% of all OTUs were detected by both methods (16 of the 27 OTUs in the coastal sample VilleFr-43 and 11 of the 17 OTUs in the offshore sample Ioni-7; see Figure 2). The dissimilarities between the two methods concerned relatively rare OTUs, such as those affiliated to *Tintinnopsis* sp. (OTU 24, 5 clones in VilleFr-43) or *Codonella aspera* (OTU 36, 3 clones in VilleFr-43), undetected by pyrosequencing, or three OTUs within the *Eutintinnus* clade (OTUs 14, 15 and 18, with 37, 22 and 17 18S rDNA pyrosequence reads, respectively) undetected in clone libraries. Part of the diversity found only by pyrosequencing corresponded to OTUs represented by singletons or doubletons (OTUs 7, 19 and 21).

From a semiquantitative point of view, the species found to be the most abundant by microscopic examination were also the ones preferentially detected by the molecular approaches. Thus, in the VilleFr-43 sample, OTUs 6, 8 and 27, which were dominant in the clone and pyroread libraries, were related to *Salpingella* sp., *Steenstrupiella steenstrupii* and *Undella* sp., respectively, and were found to dominate the microscopically determined tintinnid community (Figure 2). In the Ioni-7 sample, the *Undella*-related OTU 27 was dominant in the clone and pyroread data sets and was also the dominant genus based on microscopic analysis (22.6% of cells). However, there were some discrepancies comparing data sets. *Eutintinnus* spp. specimens, which were abundant based on microscopic observations, especially in Ioni-7, were detected only in low abundances by the molecular approaches. In contrast, Codonellopsidae appeared overrepresented in the molecular results because only a few *Codonellopsis* sp. and *Stenosemella* sp. specimens were present based on microscopic examination (Figure 1).

Discussion

Pyrosequencing versus classical clone libraries/Sanger sequencing
Pyrosequencing and the classical cloning method revealed a relatively similar phylogenetic structure of the tintinnid community at the species level. These observations resembled other molecular comparative studies of whole eukaryotic communities targeted with eukaryotic universal primers (Edgcomb *et al.*, 2011) or of particular groups, such as fungi, targeted with specific primers (Tedersoo *et al.*, 2010). In the case of our tintinnid communities, the presence and proportion of the dominant lineages were similar for both molecular methods and the OTUs unique to each method corresponded to rare lineages. These discrepancies in low abundant OTUs might be induced by the relative low quantity of DNA coming from rare species in the extracted DNA samples, and hence randomly amplified or not amplified during PCR. Some discrepancies could also be because of the very different length of clone sequences and pyroreads, and consequently a different amount of phylogenetic information, leading to a different number of potentially detectable OTUs. The fact that differences in amplicon length could notably influence estimates of richness has been shown with amplicon libraries constructed for bacteria (Huber *et al.*, 2009; Engelbrektson *et al.*, 2010).

A total of 97.7% of clone sequences and 81.5% of pyroreads (18S rDNA and ITS together) could be reliably assigned to known tintinnid taxa at the genus level (⩾98% sequence identity) and, furthermore, a majority of them (85% of clones and 69.5% of pyroreads) could be related with strong
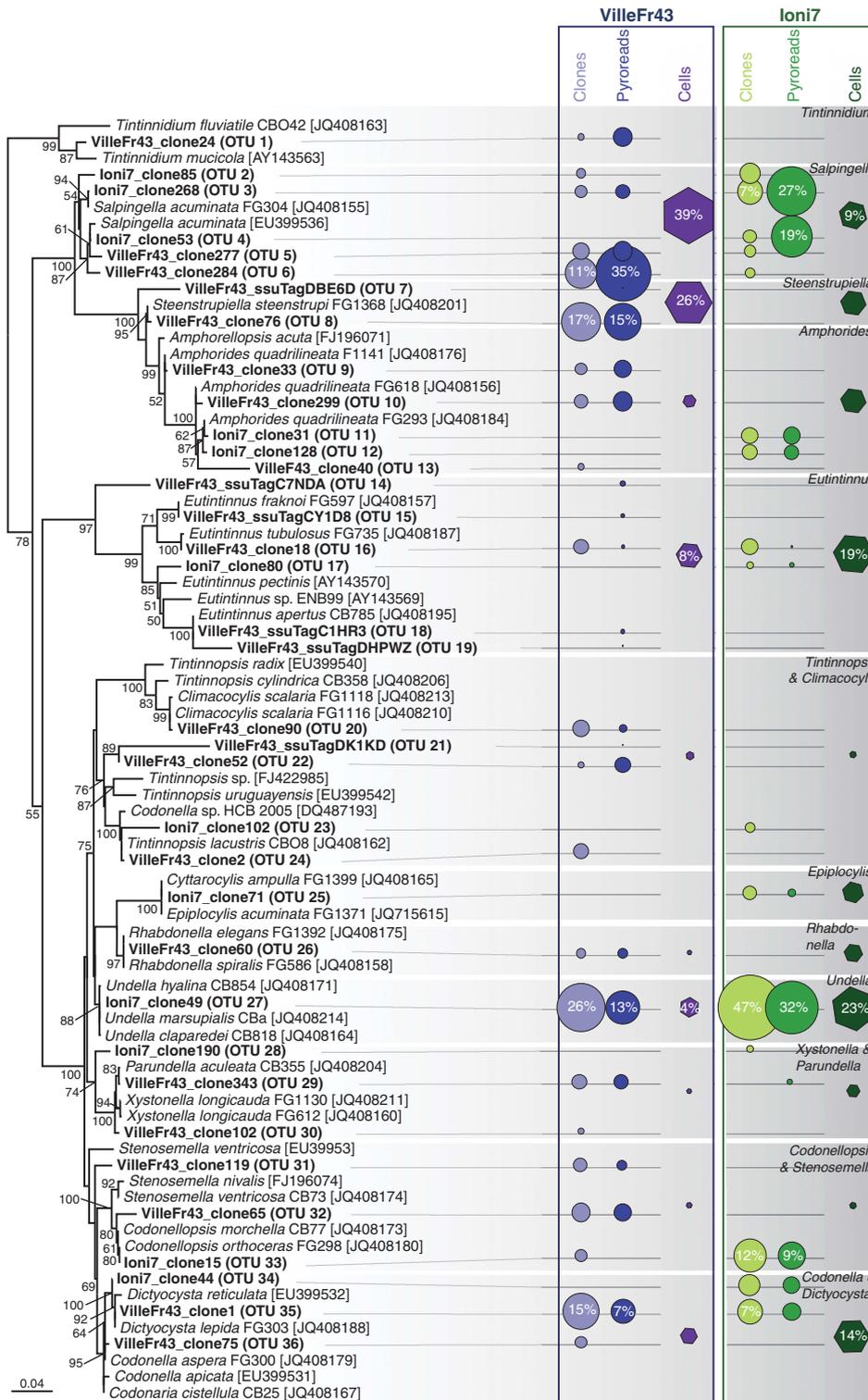
Figure 2 Maximum likelihood phylogenetic tree of 18S rDNA sequences of detected OTUs and their closest tintinnid relatives, based on 1200 aligned positions. Sequences from this study are in bold. Relative proportions of the different OTUs in each sample estimated from clone libraries and 18S rDNA pyroreads are indicated by circles of proportional size on the right. Relative proportions of the different tintinnid species estimated from cell counts under the microscope are indicated by hexagons of proportional size on the right. The numbers at nodes are bootstrap values (values <50% are omitted). Accession numbers are provided in brackets. The scale bar represents the number of substitutions for a unit branch length.

confidence to a particular species ($\geqslant 99\%$ sequence identity). In fact, no novel distinct clades without described representatives were detected despite our

deep sequence coverage, especially by the pyrosequencing approach. The lack of novel clades could be explained by recent efforts of species genetic

characterization in the Mediterranean area (Bachy et al., 2012). However, for certain genera such as *Salpingella*, some lineages (OTUs 2, 5 and 6) were only represented by environmental sequences, which argues for the need of further taxonomic work to genotype more tintinnid species.

Molecular versus morphological identification
In contrast with previous studies addressing the accuracy of pyrosequencing and/or clone libraries to describe microbial diversity in natural samples, we targeted a particular protist group with conspicuous morphological traits that permitted a precise description of the community using morphological inspection in parallel with molecular techniques. This provided us not only with a maximum diversity limit (no more OTUs than cells could exist in the samples), but also with a control reference for the expected number of OTUs and their relative abundances. In fact, the accurate characterization of microbial communities requires not only taxonomic identification but also measurement of relative abundances. OTU frequencies (cloning or pyrosequencing derived) are often used in molecular studies of microbial communities to infer the relative ecological relationships among taxa (for example, dominance or evenness; Hughes et al., 2001).

In our study, the frequency distribution of clone and pyroread sequences did not fit perfectly the frequency distribution of tintinnid morphospecies encountered under the microscope. For example, *Tintinnidium* was found in sequence libraries but not observed under the microscope. This could be because of its amorphous lorica that can be mistaken for detrital material, leading to overlooking it in microscopic examinations if it was present in low numbers. In addition, differences in rRNA gene copy number among taxa probably exist, which is often cited as the major reason of incongruence between environmental sequence data and morphology (Medinger et al., 2010). Among tintinnids, the number of rRNA gene repeats probably varies, in particular because species may contain different number of nuclei (Agatha and Strüder-Kypke, 2007). In addition, in the case of the less abundant species, our use of two different subsamples for each locality may explain some differences between morphological- and molecular-based estimations, as very rare species could have been unequally distributed when the initial water samples were divided into two subsamples.

Influence of marker choice and computational analyses on diversity estimates
Identifying and counting each tintinnid cell in our samples allowed us to unambiguously compare their species richness and distribution between different analyses. As previously observed by Doherty et al. (2007) for planktonic ciliates, we found that the number of morphologically recognizable tintinnid species was roughly similar to the number of OTUs detected in clone libraries. However, pyrosequencing data and the multiple approaches existing to analyze them yielded much more contradictory results. For our samples, the number of OTUs increased by a factor up to 100 depending on the *in silico* approach applied. The multiple sequence alignment and average clustering grossly overestimated the number of OTUs. Similarly, methods such as pairwise alignment, greedy algorithm for clustering of representative sequences and SLP inferred a number of OTUs that did not approach the numbers estimated by morphology until we went down to a 95% identity threshold for OTU definition (Supplementary Figure S3). However, a 95% threshold is clearly not realistic as we know from described tintinnid morphospecies that members of two different species may have 99% identical sequences (Supplementary Figure S1 and Santoferrara et al., 2012). Only the denoising algorithms based on sequence clustering approaches implemented in AmpliconNoise (Quince et al., 2011) and CD-HIT-OTU (Wu et al., 2011) were able to correctly estimate OTU diversities at 99% sequence identity to yield values comparable to those observed by morphological study and clone library analysis.

Although the denoising methods have been demonstrated to minimize the number of spurious OTUs (Quince et al., 2011), they tended to underestimate to some extent the species richness in the Ionian sample (Ioni-7). Compared with the 36 morphospecies found in this sample, the estimates obtained from the clone libraries (17–22 OTUs), 18S rDNA pyroreads (12–23 OTUs) and ITS pyroreads (7–18 OTUs) were low. The distinction as different species of what might be just morphological varieties could explain, at least partially, why we found such reduced diversity in molecular approaches. For instance, *Undella* spp. encompassed six morphotypes in the sample (Supplementary Table S1) whereas it is known that there is very low genetic divergence in this genus for the markers we studied (Bachy et al., 2012). Moreover, the relatively small power of the ITS region selected for pyrosequencing to discriminate closely related species could explain the even lower diversity found in the ITS pyroread data set.

PCR and sequencing errors may inflate diversity estimates by creating false taxa, and this can also be the case if inadequate methodological approaches are applied to analyze pyrosequencing data (Quince et al., 2009; Reeder and Knight, 2009; Kunin et al., 2010). In fact, the use of a morphological reference allowed us to show that certain methods led to massive overestimation of OTU numbers, in particular for rare OTUs, not only because of sequence errors but also by other methodological problems. Notably, misaligned sequences in multiple sequence

alignments can be considered as different OTUs. The size and diversity of the 'rare biosphere' may be somewhat less than what has been so far inferred by the first analyses of microbial diversity based on pyrosequencing data. The increasing availability of rigorous bioinformatic tools to treat these massive sequence data sets should allow carrying out realistic biodiversity estimates. This and the continued effort to obtain sequence data from described species to serve as reference will greatly improve the description of microbial communities by massive sequencing approaches.

Conflict of Interest

The authors declare no conflict of interest.

References

Agatha S, Strüder-Kypke MC. (2007). Phylogeny of the order Choreotrichida (Ciliophora, Spirotricha, Oligotrichea) as inferred from morphology, ultrastructure, ontogenesis, and SSrRNA gene sequences. *Eur J Protistol* **43**: 37–63.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bachy C, López-García P, Vereshchaka A, Moreira D. (2011). Diversity and vertical distribution of microbial eukaryotes in the snow, sea ice and seawater near the North Pole at the end of the polar night. *Front Microbiol* **2**: 106.

Bachy C, Gómez F, López-García P, Dolan JR, Moreira D. (2012). Molecular phylogeny of tintinnid ciliates (Tintinnida, Ciliophora). *Protist* **6**: 873–887.

Bik HM, Sung WAY, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A *et al.* (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol Ecol* **21**: 1048–1059.

Brown MV, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM *et al.* (2009). Microbial community structure in the North Pacific ocean. *ISME J* **3**: 1374–1386.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD *et al.* (2009). Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**: 5797–5808.

Doherty M, Costas BA, McManus GB, Katz LA. (2007). Culture-independent assessment of planktonic ciliate diversity in coastal northwest Atlantic waters. *Aquat Microb Ecol* **48**: 141–154.

Dolan JR, Claustre H, Carlotti F, Plounevez S, Moutin T. (2002). Microzooplankton diversity: relationships of tintinnid ciliates with resources, competitors and predators from the Atlantic Coast of Morocco to the Eastern Mediterranean. *Deep Sea Res* part I **49**: 1217–1232.

Dolan JR, Lemée R, Gasparini S, Mousseau L, Heyndrickx C. (2006). Probing diversity in the plankton: using patterns in tintinnids (planktonic marine ciliates) to identify patterns. *Hydrobiologia* **555**: 143–157.

Dolan JR, Ritchie ME, Tunin-Ley A, Pizay M-D. (2009). Dynamics of core and occasional species in the marine plankton: tintinnid ciliates in the north-west Mediterranean Sea. *J Biogeogr* **36**: 887–895.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C *et al.* (2011). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J* **5**: 1344–1356.

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H *et al.* (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**: 642–647.

Gihring TM, Green SJ, Schadt CW. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14**: 285–290.

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al.* (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* **10**: 3349–3365.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**: 494–504.

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.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

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.

Jobb G, von Haeseler A, Strimmer K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**: 18.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment

based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

Kofoid CA, Campbell AS. (1929). *A Conspectus of the Marine and Freshwater Ciliata Belonging to the Suborder Tintinnoinea, with Descriptions of New Species Principally from the Agassiz Expedition to the Eastern Tropical Pacific* vol 34. Univ Calif Publs Zool 1904–1905.

Kofoid CA, Campbell AS. (1939). The Ciliata: the Tintinnoinea. *Bull Mus Comp Zool Harvard* **84**: 1–473.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

López-García P, Rodriguez-Valera F, Pedros-Alio C, Moreira D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603–607.

López-García P, Philippe H, Gail F, Moreira D. (2003). Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. *Proc Natl Acad Sci USA* **100**: 697–702.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Marshall SM. (1969). Protozoa, order Tintinnia. *Conseil International pour l'exploration de la Mer, Fiches d'Identification de Zooplancton: fiches:* Copenhagen, Denmark, pp 117–127.

Massana R, Pedrós-Alió C. (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213–218.

Medinger R, Nolte V, Pandey RV, Jost S, Ottenwalder B, Schlotterer C et al. (2010). Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* **19**: 32–40.

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.

Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ. (2007a). Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. *Environ Microbiol* **9**: 1233–1252.

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Tobe K et al. (2007b). Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315**: 253–255.

Pedrós-Alió C. (2007). Dipping into the rare biosphere. *Science* **315**: 192–193.

Philippe H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res* **21**: 5264–5272.

Pierce RW, Turner JT. (1993). Global biogeography of marine tintinnids. *Mar Ecol Prog Ser* **94**: 11–26.

Price MN, Dehal PS, Arkin AP. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–1650.

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM et al. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.

Reeder J, Knight R. (2009). The 'rare biosphere': a reality check. *Nat Methods* **6**: 636–637.

Saccà A, Strüder-Kypke MC, Lynn DH. (2012). Redescription of *Rhizodomus tagatzi* (Ciliophora: Spirotrichea: Tintinnida), based on morphology and small subunit ribosomal rna gene sequence. *J Eukaryot Microbiol* **59**: 218–231.

Santoferrara LF, McManus GB, Alder VA. (2012). Utility of genetic markers and morphology for species discrimination within the Order Tintinnida (Ciliophora, Spirotrichea). *Protist;* e-pub ahead of print 18 January 2012; doi:10.1016/j.protis.2011.12.002.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schloss PD. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* **6**: e1000844.

Snoeyenbos-West OLO, Salcedo T, McManus GB, Katz LA. (2002). Insights into the diversity of choreotrich and oligotrich ciliates (Class: Spirotrichea) based on genealogical analyses of multiple loci. *Int J Syst Evol Microbiol* **52**: 1901–1913.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR et al. (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

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

Strüder-Kypke MC, Lynn DH. (2008). Morphological versus molecular data - Phylogeny of tintinnid ciliates (Ciliophora, Choreotrichia) inferred from small subunit rRNA gene sequences. *Denisia* **23**: 417–424.

Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W et al. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**: e76.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I et al. (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol* **188**: 291–301.

Wu S, Zhu Z, Fu L, Niu B, Li W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**: 444.

Zinger L, Gobet A, Pommier T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–1896.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)