environmental microbiology

Environmental Microbiology (2022) **24**(7), 2979–2993 doi:10.1111/1462-2920.16081

Intra-genomic rRNA gene variability of Nassellaria and Spumellaria (Rhizaria, Radiolaria) assessed by Sanger, MinION and Illumina sequencing

Miguel M. Sandin [1,2*] **Sarah Romac** [1] **and Fabrice Not** [1]

[1]*Sorbonne University, CNRS - UMR7144 - Ecology of Marine Plankton Group - Station Biologique de Roscoff, 29680, Roscoff, France.*

[2]*Department of Organismal Biology (Systematic Biology), Uppsala University, Norbyv. 18D, 75236, Uppsala, Sweden.*

Summary

Ribosomal RNA (rRNA) genes are known to be valuable markers for the barcoding of eukaryotic life and its phylogenetic classification at various taxonomic levels. The large-scale exploration of environmental microbial diversity through metabarcoding approaches has been focused mainly on the V4 and V9 regions of the 18S rRNA gene. The accurate interpretation of such environmental surveys is hampered by technical (e.g. PCR and sequencing errors) and biological biases (e.g. intra-genomic variability). Here we explored the intra-genomic diversity of Nassellaria and Spumellaria specimens (Radiolaria) by comparing Sanger sequencing with Illumina and Oxford Nanopore Technologies (MinION). Our analysis determined that intra-genomic variability of Nassellaria and Spumellaria is generally low, yet some Spumellaria specimens showed two different copies of the V4 with <97% similarity. Of the different sequencing methods, Illumina showed the highest number of contaminations (i.e. environmental DNA, cross-contamination, tag-jumping), revealed by its high sequencing depth; and MinION showed the highest sequencing rate error (∼14%). Yet the long reads produced by MinION (∼2900 bp) allowed accurate phylogenetic reconstruction studies. These results highlight the requirement for a careful interpretation of Illumina-based metabarcoding studies, in particular regarding low abundant amplicons, and open future perspectives towards full-length rDNA environmental metabarcoding surveys.

Received 12 April, 2022; revised 18 May, 2022; accepted 20 May, 2022. *For correspondence. E-mail: miguelmendezsandin@gmail.com.

Introduction

Ribosomal RNA (rRNA) genes are known to be valuable taxonomic markers for the barcoding of eukaryotic life and its phylogenetic classification at different levels; mainly due to its intra-genomic tandem repeated structure, the presence of conserved and variable regions and its occurrence in all eukaryotes (Pawlowski *et al*., 2012; del Campo *et al*., 2018). The 18S rRNA gene has been widely used in molecular environmental surveys, in particular the short hypervariable regions V4 and V9, thanks to the extensive occurrence in public databases and the availability of generalist primers flanking their sides (Amaral-zettler *et al*., 2009; Stoeck *et al*., 2010). The advent of high-throughput sequencing (HTS) techniques has allowed the massive sequencing of molecular environmental diversity supporting its exploration through a metabarcoding approach (de Vargas *et al*., 2015; Massana *et al*., 2015; Forster *et al*., 2016; Pernice *et al*., 2016). The large number of reads generated by HTS is normally classified into operational taxonomic units (OTUs) based on arbitrary similarity thresholds. OTUs are not only used to identify taxonomic entities but also to describe community structure (Blaxter *et al*., 2005). The increasing use of the HTS has led to the development of different clustering methods resulting in finer-scale OTUs that focus on single nucleotide differences (Mahé *et al*., 2015) or on the correction of sequencing errors based on the error rate entropy (so-called amplicon sequence variants or ASVs; Callahan *et al*., 2016).

HTS produce a vast amount of reads carrying errors that are difficult to distinguish from real biological variation, which is considered as a main factor inflating diversity (Kunin *et al*., 2010). Intra-genomic rDNA polymorphism and its different copy numbers among taxa can also affect diversity assessments (Gong *et al*., 2013; Gong and Marchetti, 2019). Other less common causes, yet important, have also been reported as factors inflating

diversity estimates, such as lateral gene transfer (Yabuki *et al*., 2014) or presence of pseudogenes (Thornhill *et al*., 2007). Considering the very high sequencing depth allowed by the current HTS technologies it is likely that most of this intra-genomic diversity could be sequenced, potentially leading to an overestimation of the environmental diversity. Several studies have argued that, quantitatively at the community level, the number of molecular clusters largely exceeds that of morphological counts (Medinger *et al*., 2010; Bachy *et al*., 2013; Santoferrara *et al*., 2016), leading to hypothesize that using HTS approaches scientists are actually measuring species intra-genomic variability (Caron and Hu, 2018). Current HTS technologies used for environmental surveys can sequence short fragments of DNA of about 400 base pairs (bp) only, such as the hypervariable regions (V4 and V9 most commonly used in protist) of the 18S rRNA gene. Comparing such short hypervariable regions to, far from exhaustive reference sequences databases may also contribute to inflating environmental diversity by misidentification of environmental clusters or lack of intra-genomic rDNA variability representation (Pitsch *et al*., 2019), among other causes.

New sequencing technologies have been developed with the ability of high-throughput sequencing longer nucleotide fragments in real-time, such as Oxford Nanopore Technologies (ONT) or Pacific Bioscience (PacBio). These sequencing methods have already showed their useful capabilities, for example PacBio has developed a circular consensus sequencing resulting in near-zero error long reads, improving aspects such as genome assembly (Wenger *et al*., 2019) or even phylogenetic analysis of environmental diversity (Jamy *et al*., 2019). However, its limited high-throughput sequencing ability and its relatively high cost (Goodwin *et al*., 2016) may affect the sequencing depth at a metabarcoding community level. On the other side, ONT provides a large quantity of reads, inexpensively and highly portable with the MinION device (Levy and Myers, 2016). Despite that the error rate of the ONT reads is improving since it was first released (for MinION, from ~60% in 2014 to <15% error rate; Rang *et al*., 2018) it is still a major concern, reaching up to 3%–6% of errors in the best-case scenarios (Tyler *et al*., 2018).

In this study, we aim at assessing both (i) the intra-genomic variability within protists and (ii) whether long-read sequencing provides a reliable source of diversity estimate suitable for environmental molecular surveys. To test these hypotheses we compare three different sequencing methods, Sanger, ONT (MinION) and Illumina. We focused our efforts on two groups of Radiolaria, the Nassellaria and Spumellaria (Polycystines). Some species of Radiolaria harbour endosymbiotic micro-algae, constituting a community system which can be considered as a so-called holobiont. The Radiolaria host is an important group of protists in eukaryotic plankton communities, contributing for a major fraction of the total reads in environmental molecular surveys (de Vargas *et al*., 2015; Pernice *et al*., 2016) but the number of morphologically described species (Suzuki and Not, 2015) does not match with the molecular barcodes. Nassellaria and Spumellaria environmental diversity lags far behind other radiolarian, despite possessing the largest morphological diversity described. Recent studies have dwelled on exploring their extant morpho-molecular diversity (Sandin *et al*., 2019; Sandin *et al*., 2021) showing their uncharted diversity among Radiolaria. The ecological importance of these groups and their observed low molecular diversity in environmental surveys stresses the need for understanding such differences between molecular and morphological diversity.

Experimental procedures

Single-cell sampling, isolation and DNA extraction

Plankton samples were collected in the Bay of Villefranche-sur-Mer (France) and in the West Mediterranean Sea (MOOSE-GE 2017 expedition) by plankton nets tows (from 20 to 64 μm mesh size). More information on sampling methodology for specific samples can be found in the RENKAN database (http://abims.sb-roscoff.fr/renkan/). Specimens were individually handpicked with Pasteur pipettes from the plankton community and maintained in 0.2 μm filtered seawater for several hours. They were transferred three to four times into new 0.2 μm filtered seawater to allow self-cleaning from debris, particles attached to the cell or prey(s) digestion. By doing so, it is expected to keep only essential entities from the holobiont (the radiolarian host + associated symbionts and bacteria).

Amplification and sequencing

A schematic representation of the study design and the different amplification and sequencing steps can be found in Supplementary material Fig. S1 and a fully detailed explanation of the experimental procedures can be found in Supplementary Information.

Four holobionts were selected to amplify the full length of the rRNA gene (Supplementary material Fig. S2), two of them belonging to Spumellaria: Mge17-81 (*Rhizosphaera trigonacantha*, Rhizosphaeroidea) and Mge17-82 (*Spongosphaera streptacantha*, Spongosphaeroidea); and two to Nassellaria: Vil325 (*Eucyrtidium cienkowskii*, Eucyrtidioidea) and Vil496 (*Eucyrtidium acuminatum*, Eucyrtidioidea). Each holobiont was PCR-amplified in three

different technical replicates covering from the beginning of the 18S rRNA gene until the end of the 2D region of the 28S rRNA gene. Final PCR product was divided into two for sequencing by Sanger (after cloning) and ONT using the MinION device (Jain *et al*., 2015; Laver *et al*., 2015) during its last 28 h (20 h → 48 h), after a prior independent experiment (data not analysed herein). Eight holobionts were selected to amplify the V4 region (~380 bp) of the 18S rRNA gene (Supplementary material Fig. S2), four belonging to Spumellaria: Mge17-81 (common to Sanger and MinION sequencing), Mge17-82 (common to Sanger and MinION sequencing), Vil480 (*Tetrapyle octacantha*, Pylonioidea), Vil497 (*Arachnospongus varians*, Liosphaeroidea); and four to Nassellaria: Mge17-9 (*Extotoxon undulatum*, Artostrobioidea), Mge17-124 (*Carpocanium obliqua*, Carpocaniidae), Vil490 (*Pterocorys* cf. *zanclea*, Pterocorythoidea), Vil496 (common to Sanger and MinION sequencing). Holobiont Vil325 (from Sanger and MinION sequencing) had no DNA left for further experiments and therefore was not possible to include in this section. After PCR amplification of three technical replicates, about 500 ng of pooled amplicons were sent to Fasteris (https://www.fasteris.com, Switzerland) for Illumina sequencing on a MiSeq nano V2 2 × 250.

Sequencing results from this study along with associated metadata for its analysis and native formats have been deposited in figshare with the following DOI: https://doi.org/10.6084/m9.figshare.16922764.v1. Raw sequencing results have also been deposited in SRA under the accession number PRJNA816840.

All scripts used in this study along with the tools for the replication of the analysis are available on github (github.com/MiguelMSandin/IntraGenomic-variability).

Sequences/amplicons analyses

Raw sequences/amplicons were measured in bp length and compared against reference sequences by local alignment (BLAST; using NCBI as reference sequences) for Sanger and MinION sequencing results and global alignment (vsearch; Rognes *et al*., 2016; using PR2 v4.14.0 database as reference sequence, Guillou *et al*., 2013) for Illumina results. Thereafter depending on the sequencing nature different pipelines were followed: Sanger reads belonging to the same replicate were concatenated, MinION reads were de-multiplexed with cutadapt (Martin, 2011) and Illumina reads were clustered into ASVs using DADA2 and the pipeline described in Callahan *et al*. (2016). Resulting ASVs were post-clustered based on co-occurrence, similarity and abundance with the 'LULU' algorithm (Frøslev *et al*., 2017). The V4 region of sequences coming from Sanger and MinION sequencing was extracted with cutadapt (Martin, 2011), using the same primer set used to amplify and sequence with Illumina (see Supplementary Information for further details), and clustered using swarm (Mahé *et al*., 2015).

In order to compare different sequencing methods and discriminate errors from intra-genomic variability, we perform an entropy analysis. Sequences taxonomically assigned to Nassellaria and Spumellaria were aligned independently using MAFFT v7.395 (Katoh and Standley, 2013) and for every position of each alignment, Shannon entropy was calculated using a custom script (see 'alignmentEntropy.py' in github.com/MiguelMSandin/IntraGenomic-variability). The entropy of the full V4 region was measured in an independent analysis to compare the three different sequencing technologies against reference sequences from PR2 v4.14.0.

Sequences assigned to Polycystines from both Sanger and Minion sequencing have been clustered into consensus sequences for phylogenetic analysis. Consensus sequences coming from Sanger were produced with a custom script (see 'alignmentConsensus.py' in github.com/MiguelMSandin/IntraGenomic-variability) and MinION consensus sequences according to Wurzbacher *et al*. (2019) with the script *consension* (https://microbiology.se/software/consension/). Resulting consensus sequences from MinION were mapped against the raw fastq file using minimap2 (Li, 2018) and lastly polished with racon v1.4.22 (Vaser *et al*., 2017). Final consensus sequences were aligned against reference sequences (extracted from Decelle *et al*., 2012; Biard *et al*., 2015; Sandin *et al*., 2019 and Sandin *et al*., 2021; Supplementary material Table S2) using MAFFT v7.395 (Katoh and Standley, 2013) and automatically trimmed using trimal (Capella-Gutiérrez *et al*., 2009) with a 30% gap threshold. The final data set contains 525 taxa and 2434 positions and phylogenetic analysis was done by RAxML (Stamatakis, 2014) and GTR + Gamma and GTR + CAT evolutionary models with 1000 rapid bootstraps. Final trees were visualized and edited with FigTree version 1.4.3 (Rambaut, 2016).

Results

Quality of sequencing results

In total four Radiolaria holobionts were sequenced by both Sanger and MinION after three independent PCR amplifications or replicates for each (see Supplementary material Fig. S1 for a schematic representation of the experimental procedure and Supplementary material Fig. S2 for a table with detailed information on each holobiont). Prior to Sanger sequencing, 24 cloning reactions per replicate were performed to further amplify the targeted rRNA gene. In total 834 Sanger sequences (representing 737 817 bases) were successfully retrieved for

864 cloned amplicons (Supplementary material Table S1). These sequences had an average length of 884.67 (±198.03) bp, with a median of 1005 bp and an average BLAST similarity identity of 98.13% (±2.59%) with a reference sequence (Fig. 1, Sanger).

The MinION flow cell was used during its last 28 h (20 h → 48 h; after a previous independent experiment) resulting in 864 total reads (representing a total of 1 645 774 bases). The low throughput obtained in our study contrast with the high number of sequences obtained during the first 20 h (0 h → 20 h) of the same flow cell in the previous independent study, where a total of 225 573 sequences were obtained (representing 296 072 423 total bases; data not analysed herein). Based on these observations, it turns that, the last 28 h of the flow cell sequenced 0.38% of the total sequences obtained by the flow cell, representing 0.56% of the total bases. These 864 sequences were compared against NCBI database by BLAST tool and 81 had no match at all. Among the remaining 783 sequences, 185 were assigned to bacteria, 593 to eukaryotes and five matching simultaneously different domains (i.e. the same sequence matching virus, uncultured prokaryote and uncultured eukaryote). These 783 sequences had an average length of 2043.48 bp (±1143.42) bp, with a median of 2746 bp and an average identity of 86.01% (±2.91%) with a reference sequence. De-multiplexing MinION sequences resulted in 55 sequences among the different replicates and specimens (Supplementary material Table S1). De-multiplexed sequences had an average length of 2641.82 bp (±681.80) bp, a median length of 2921 bp and an average identity of 86.48% (±2.33%) similarity with a reference sequence (Fig. 1, MinION).

Regarding Illumina sequencing, a total of 1 019 196 sequences (representing 254 799 000 bases) were sequenced in both the R1 and the R2 files from eight Radiolaria holobionts, of which only three were also sequenced by Sanger and MinION (due to original DNA limitations; Supplementary material Fig. S2). On average 34 671 (±19,609) reads were obtained for each replicate. These reads were merged resulting in 30 524 (±17,359) amplicons on average per replicate and 66 913 total unique amplicons (Supplementary material Table S1). They had an average length of 375.3 (±15.89) bp, a

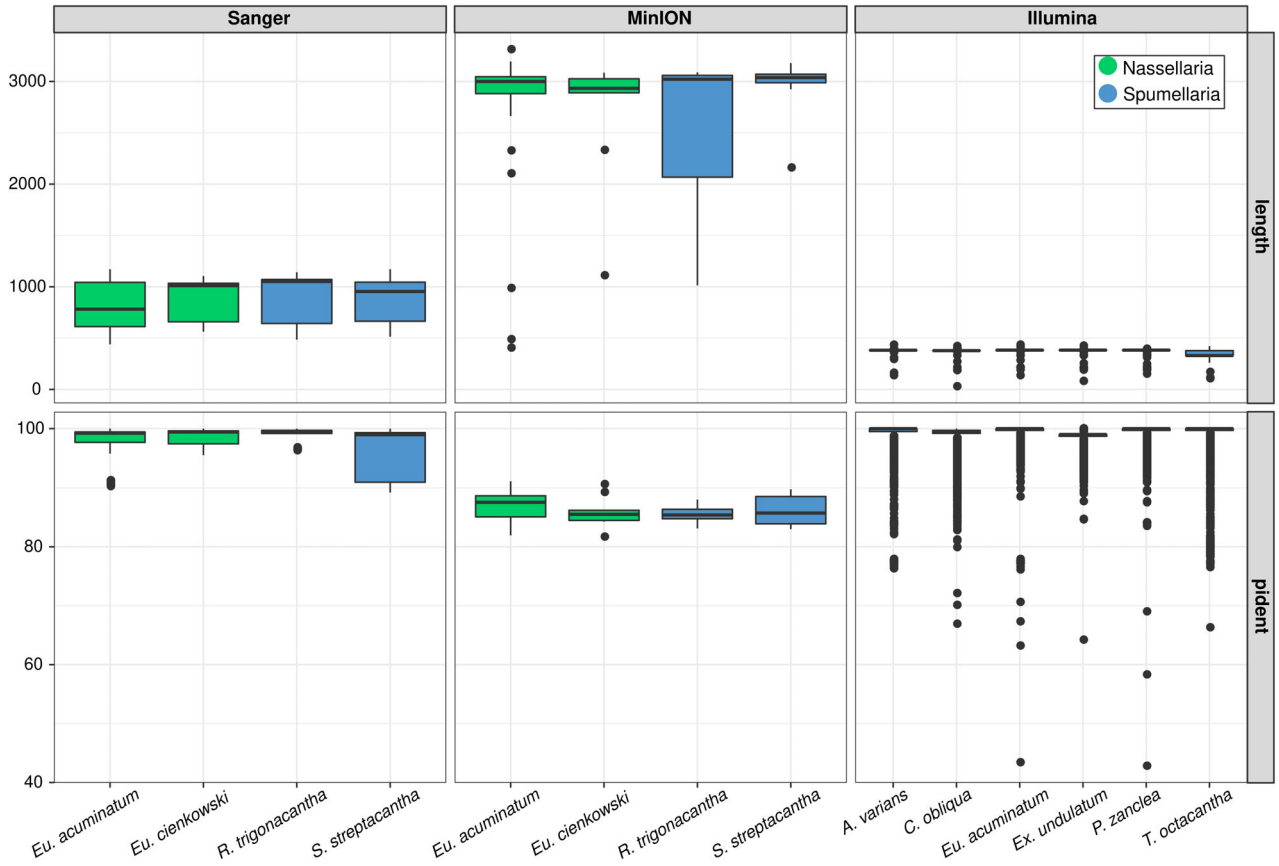

Fig. 1. Boxplot summarizing sequencing results of Sanger, Oxford Nanopore Technologies (MinION) and Illumina sequencing; top row shows the sequence length (in bp) and bottom row the percentage identity of the first match against a reference sequence per cell. Note that identity for Sanger and MinION was performed by BLAST and for Illumina by global alignment (vsearch) due to the shorter length of the reads.

median of 382 bp and an average identity of 99.28% (±2.26%) with a reference sequence (Fig. 1, Illumina).

Taxonomic assignation and diversity of reads

Sanger sequenced amplicons of the different parts of the ribosomal genes belonging to the same cloned replicate were concatenated resulting in 287 sequences with an average length of 2570.7 (±296.97) bp. This covers the first ∼1000 bp of the 18S rRNA gene (primer SA), the last ∼600–800 bp of the 18S and the Internal Transcribed Spacer 1 (ITS1) (primer S69f) and the regions D1 and D2 from the 28S rRNA gene (primer D1R). These sequences were assigned predominantly to Nassellaria and Spumellaria, representing 85.4% of the sequences (Fig. 2, Sanger). The rest of the sequences belong to Chrysophyceae, Ciliates, Diatoms, Fungi, Dinoflagellates and Alveolates. The 55 de-multiplexed sequences extracted from MinION sequencing showed a similar taxonomic assignation as for Sanger sequencing, yet less diverse: Nassellaria and Spumellaria represented 72.7% of the total assignations and Diatoms, Chrysophyceae, Dinoflagellates, Fungi and Alveolates the rest of the groups (Fig. 2, MinION).

Amplicons of Illumina were clustered using an error-based method (dada2) resulting in 153 ASVs, comprising 570 041 reads. These ASVs had an average length of 379.2 (±8.99) bp, a median length of 378 bp and an average similarity identity of 99.70% (±5.97%). The most diverse groups are dinoflagellates (28 ASVs), Spumellaria (23 ASVs), Nassellaria (19 ASVs) and Fungi (14 ASVs mostly within Agaricomycotina and Ustilaginomycotina) (Fig. 2, MinION). Despite the big diversity within each holobiont, the number of reads is highly dominated by 1, 2 or 3 ASVs belonging to the host of the holobiont or the symbiotic algae (Supplementary material Fig. S3). Due to the great taxonomic diversity found within holobionts by Illumina sequencing, ASVs were only considered if they were present in at least three replicates (triplicates in the PCR), with a total abundance equal to or higher than the median value of the abundance (that is 102 reads) and assigned until at least the taxonomic rank 'Order'. We have chosen stringent thresholds in an attempt to remove all artefacts and/or contaminations, while being aware that part of the diversity within the holobionts might also be removed. After processing, the number of ASVs decreased to 36 ASVs but the total reads did not change drastically with 544 127 reads, representing up to 95.45% of the total reads. Main taxonomic affiliations were similar to the one described before filtering the ASVs. Furthermore, the relative proportion of unexpected ASVs took more importance; such as an ASV affiliated to Collodaria present in four holobionts, an ASV affiliated to Acantharea present in three holobionts and one to two ASVs affiliated to Craniata (assigned to the genuses *Homo* and *Capra* with 99 and 69 minimum bootstrap support respectively) present in every holobiont (Supplementary material Fig. S4). When exploring in detail the ASVs affiliated to Polycystines and their abundance and distribution among replicates after applying the stringent thresholds, it is possible to find up to three different and highly abundant ASVs within the same specimen (e.g. *Spongosphaera streptacantha* or *Arachnospngus varians*; Fig. 3). These ASVs appearing in the same three replicates within a unique specimen had a high similarity among one another, and they were grouped together into a single ASV after post-clustering based on co-occurrence and similarity with 'LULU' (Supplementary material Fig. S5). Some of the most abundant ASVs were found in the three replicates of their corresponding specimens and in a fourth and fifth replicate; as for 'ASV-2', 'ASV-5' and 'ASV-7', ASVs from the Nassellaria specimens *Pterocorys zanclea*, *Extotoxon undulatum* and *Eucyrtidium acuminatum* respectively. 'ASV-2' was present in the three replicates of *Eu*. *acuminatum* and in one replicate of *Carpocanium obliqua* and *Rhizosphaera trigonacantha*, a Spumellaria specimen, although at very low relative abundances. In *E*. *undulatum* there were two ASVs, of which one ('ASV-5') was present in the three replicates of the holobiont and in a fourth and fifth replicate from *Rhizosphaera trigonacantha* and *Spongosphaera streptacantha*, the second ASV ('ASV-47') was only present in one replicate. While ASV-47 is also found in *C*. *obliqua*, *T*. *octacantha* and *A*. *varians*, it never appears within the three replicates of the same holobiont. Its taxonomic affiliation is to Collodaria, with a 100% similarity to a reference sequence and 100 BS support for the taxonomic assignation.

Intra-genomic variability

The hyper-variable region V4 of the 18S rRNA gene was extracted from the sequences affiliated to Radiolaria from both Sanger and MinION sequencing, in order to compare with Illumina sequencing and assess if the same genetic diversity is detectable within the holobiont. In total 52, 56, 59 and 68 sequences were obtained from Sanger sequencing of *Eucyrtidium acuminatum*, *Eu*. *cienkowski*, *Spongosphaera streptacantha* and *Rhizosphaera trigonacantha* respectively. The same unique sequence was found to be the most abundant in both *Eu*. *acuminatum* and *Eu*. *cienkowski* (54 and 47 reads respectively) being exactly identical despite being obtained from specimens morphologically identified as different species. In addition, up to five other unique sequences were found with a very low abundance (2, 2, 1, 1 and 1 reads, green dots on Fig. 4 under unique

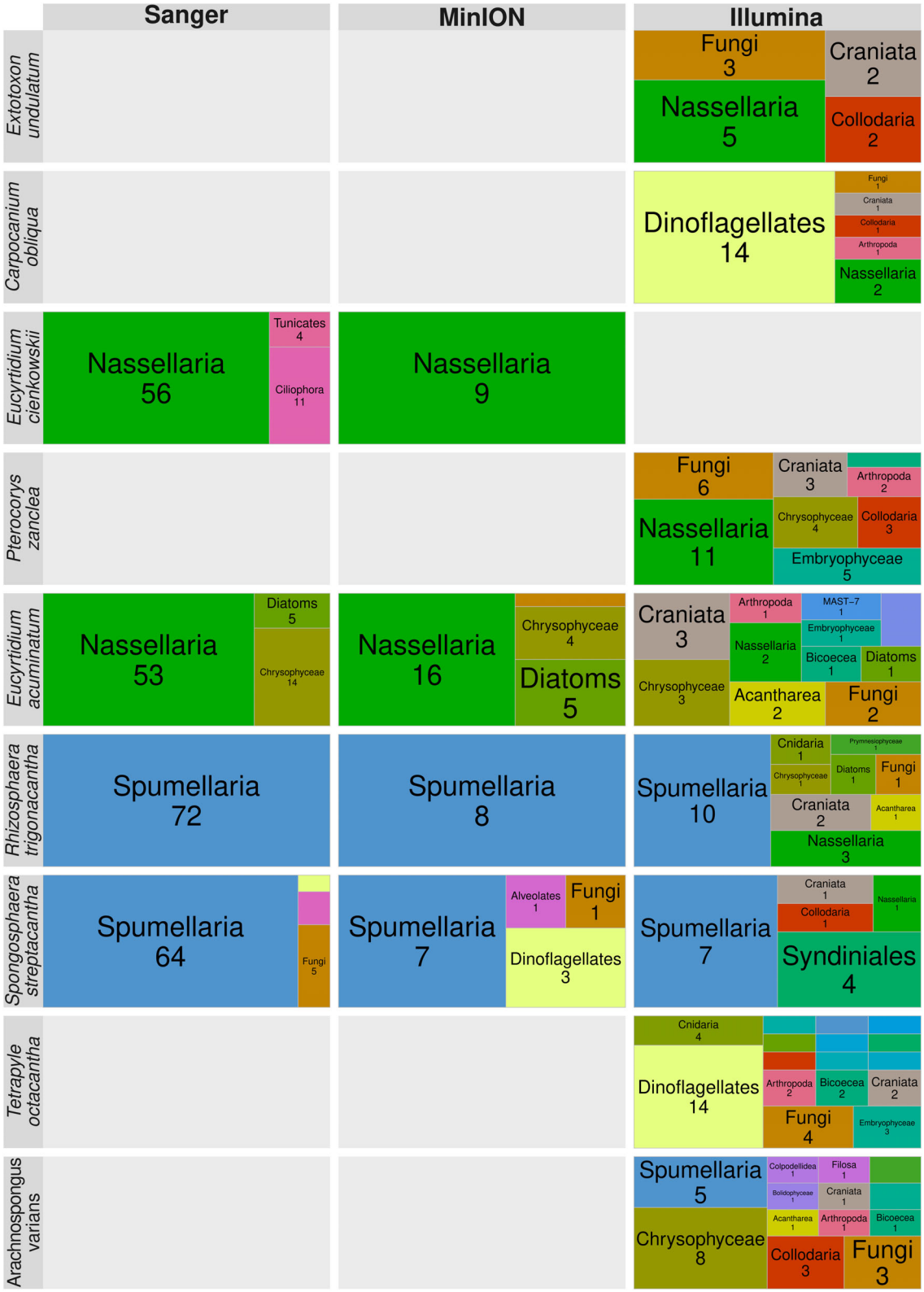

Fig. 2. Taxonomic affiliation of sequences/amplicons obtained by Sanger (after concatenation of the amplified fragments), Oxford Nanopore Technologies (MinION) and Illumina (reads processed by dada2) sequencing for each cell. The area represents the proportion of total number of unique sequences/amplicons affiliated to the specific taxonomic entity (tree map). Numbers below the taxonomic group represent the number of unique sequences/amplicons.

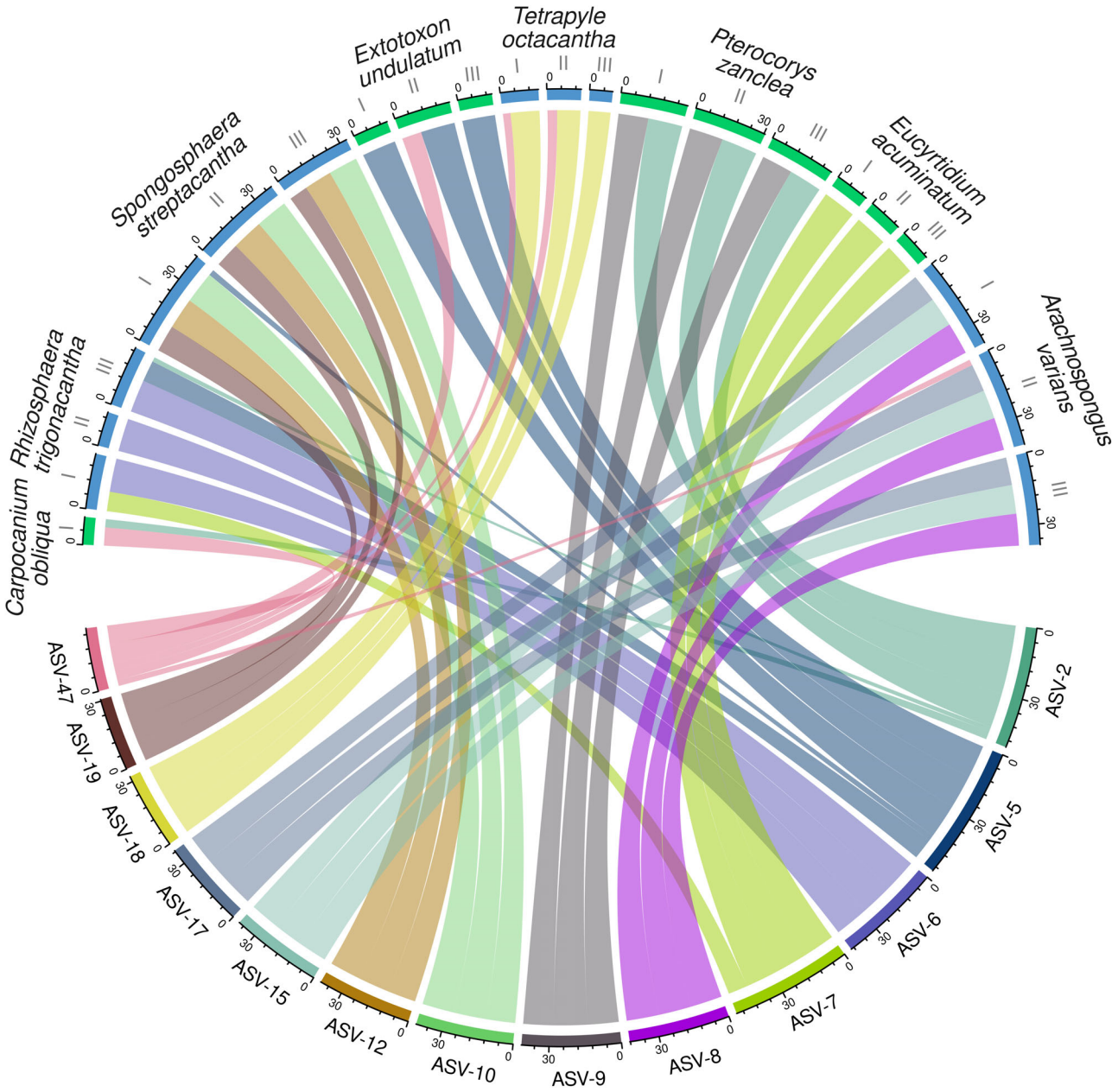

Fig. 3. Circular plot representing the log-transformed abundance of polycystines amplicons (lower half of the circles) clustered with dada2 and replicate affiliation (upper half of the circles, indicated by roman numbers under the name of the specimens). Only amplicons affiliated to polycystines present in three or more samples and with a total abundance equal to or higher than 102 reads (median) were considered.

reads). All sequences of the V4 hypervariable region extracted from the two Nassellaria specimens showed similarities within them close to 100% with a maximum of one base of difference (boxplots on Fig. 4 under unique reads). Despite these two different morpho-species showing identical V4 sequences, the ITS1 and the partial 28S rRNA gene (D1 + D2 regions) show inter-specific differences (data not shown). Similar patterns of intra-genomic variability in the V4 hypervariable region are seen in the Spumellaria *R. trigonacantha*, with one sequence found 67 times and another one found only one time. In contrast, *S. streptacantha* showed up to three different sequences relatively abundant (16, 13 and 13 reads) and eight other sequences with a lower abundance, having a similarity among them of 96.1%. It is important to note that the three most abundant sequences of *S. streptacantha* were present in the three different replicates (PCR reactions). When these sequences were clustered with the software swarm using 1 difference threshold, *Eu. acuminatum* and *Eu.*

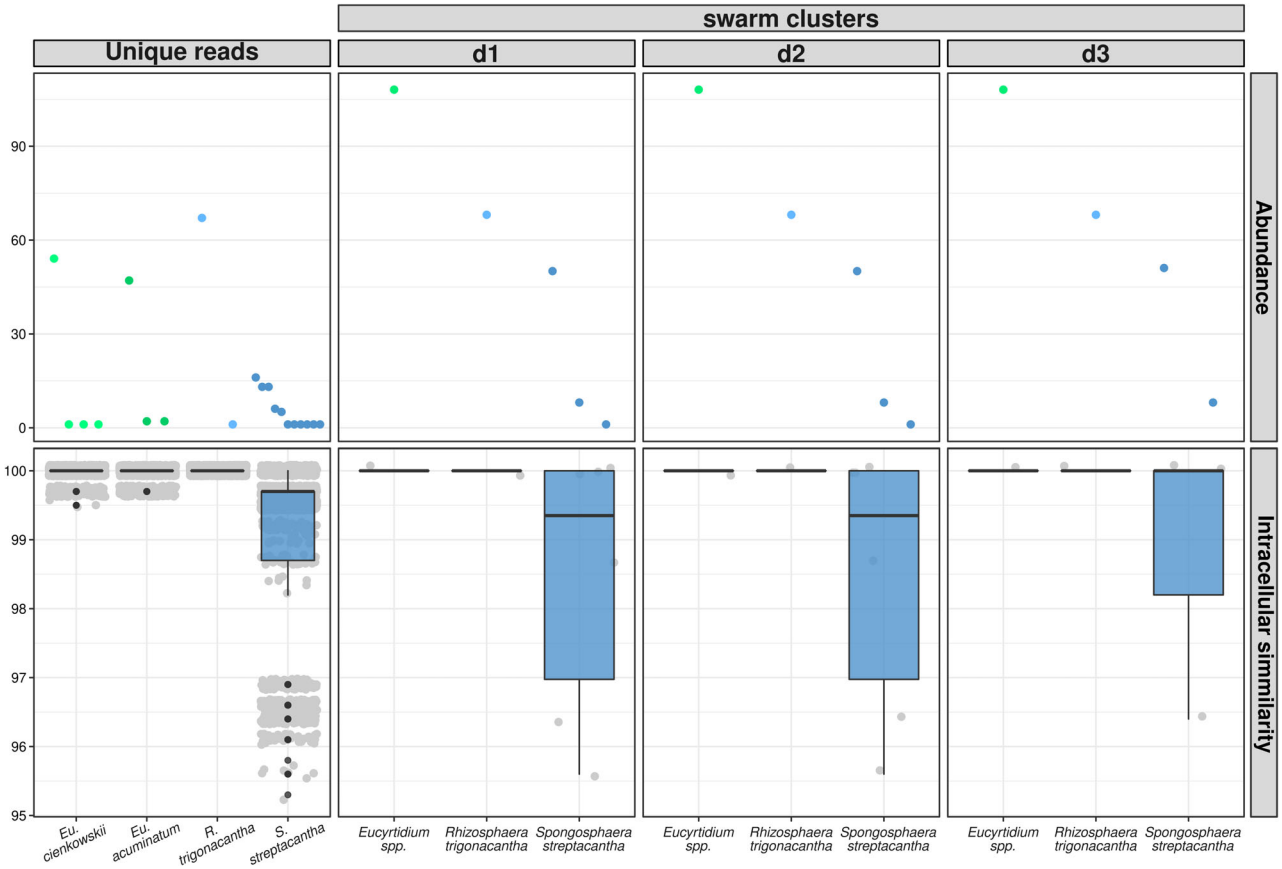

Fig. 4. Abundance and intra-genomic similarities of the V4 region extracted from Sanger sequencing results of the total unique reads and clustered with swarm with 1 (d1), 2 (d2) and 3 (d3) differences.

cienkowski shared the same amplicon. In *R. trigonacantha* the sequence with a single read was grouped in the unique amplicon found for this holobiont, and up to three different amplicons were found in *S. streptacantha* (still with a similarity among these amplicons of ~96%). Increasing the threshold to two differences did not change the outcome for *S. streptacantha*. When using a threshold of three differences, the number of amplicons in *S. streptacantha* decreases to 2, with low changes in the similarities among them (96.4%). Yet one amplicon largely dominates the other (51 reads against eight reads). The detailed exploration of such polymorphism revealed that nucleotide differences are mostly located in variable helixes of the secondary structure of the 18S rRNA gene (Supplementary material Fig. S6). Same protocol was implemented on the V4 hyper-variable region of the 18S rRNA extracted from the MinION sequences. In total four sequences were extracted from *R. trigonacantha*, two from *S. streptacantha*, six from *Eu. cienkowski* and 20 from *Eu. acuminatum*. It was not possible to cluster the sequences with swarm due to their large dissimilarities. Up to a difference threshold of 10, there were still no clusters, keeping an average intra-genomic similarity of 86.92% (±3.56%).

Discerning intra-genomic variability from sequencing errors

In order to compare the intra-genomic variability between different sequencing methods and discriminate between sequencing errors, we calculated the entropy of the alignment from sequences obtained in this study at every position. Since it is expected that errors are random, a single error in the alignment would represent low evenness resulting in low entropy values (yet not 0). On the other side, intra-genomic variability is expected to be sequenced in several replicates increasing the evenness and with it the entropy. In general, concatenated sequences from Sanger sequencing belonging to Nassellaria showed a trend of Shannon entropy values of 0 among them (Supplementary material Fig. S7). Towards the end of the amplified fragments the entropy slightly increased, meaning that most probably the sequencing error rate increases. Especially at the end of

the region D2 of the 28S rRNA gene the entropy reaches its highest values along a region of ~100 bp length, probably meaning variability among the different copies of the rDNA. Regarding Spumellaria, *Rhizosphaera trigonacantha* has a similar trend as Nassellaria (Supplementary material Fig. S7). In contrast, *Spongosphaera streptacantha* shows regions of high entropic values at around the position 750 (V4 region from the 18S rRNA gene), along with the beginning of the 28S rRNA gene and especially over a region of ~250 bp on the ITS1, showing a most probable big intra-genomic variability of the rDNA. In general MinION sequences aligned against the reference sequence with many gaps (Supplementary material Fig. S7). Alignments of sequences coming from MinION have in general a greater number of positions than alignments of sequences from Sanger sequencing (~922.25 bp more on average); greater than the length of the 5.8S and ITS2 regions of the rDNA that MinION sequences are covering in comparison to Sanger sequences (18S + ITS1 and D1 and D2 regions of the 28S rRNA gene). These sequences had high entropy values all along the rDNA alignment. With the exception of the 18S rRNA gene of Nassellaria, where there is a constant entropy trend, there are variations of the entropy depending on the region, but these variations do not match exactly those of the Sanger sequences.

The V4 hypervariable region of the 18S rRNA gene used in Fig. 4 along with that extracted from MinION sequencing and Illumina were pooled together in one alignment for the V4 region comparison. In total, 108 sequences for Nassellaria and 127 sequences for Spumellaria were extracted from Sanger sequencing, 19 Nassellaria sequences and five Spumellaria sequences were extracted from MinION sequencing, including those amplicons extracted from Illumina taxonomically assigned to Nassellaria (19) and Spumellaria (23) and completed with reference sequences from pr2 v4.14.0 (83 for Nassellaria and 671 for Spumellaria) trying to gather most of their known diversity. Final aligned dataset had a length of 478 bp for Nassellaria and 461 bp for Spumellaria (Fig. 5). Reference sequences showed a hotspot of higher Shannon diversity within the hypervariable region in both Nassellaria and Spumellaria spanning from the position ~80 until ~210 with maximum values around the position ~110–150. Nassellaria reference sequences showed two other hotspots regions peaking at around the positions ~330 and ~420, yet average entropy values are half of the first hotspot region. Regarding Spumellaria, these two last hotspot regions are less marked than for Nassellaria due to overall higher entropy values. Sanger sequences maintain near-0 entropy values, with the exception of a region between ~100 and 200 bp in Spumellaria showing higher entropy values (corresponding to *Spongosphaera streptacantha*). Illumina ASVs follow the same patterns as the reference alignment, with the highest entropy values at around the position ~150. Despite the lack of trend found for the full length of the rDNA for MinION sequences (Supplementary material Fig. S7), when focusing on the V4 region, shows similar patterns that those found in the reference alignment and Illumina, although these peaks are smoother due to the relative higher entropy values all along the alignment.

Phylogenetic signal of long-read sequencing

We performed a last attempt to test whether the high-error rate from MinION is affecting further downstream analysis or could be overcome. Sequences obtained from Sanger and MinION were clustered into consensus sequences in order to resolve potential random errors and were aligned in a phylogenetic tree along with reference sequences (Supplementary material Table S2). Despite the high error rate of MinION sequences, both the random distribution of the errors and the long reads provide accurate phylogenetic information when processed into consensus sequences (Fig. 6). Consensus sequences from MinION are phylogenetically sister to the consensus sequences from Sanger showing high phylogenetic support and short branches. Consensus sequences from Sanger sequencing have a similarity >99% against reference sequences except for *Eucyrtidium cienkowski* (97.32%). Consensus sequences from MinION have a maximum similarity identity against reference sequences of 99.5%, 98.9% and 99.2% for *Eucyrtidium* sp., *Rhizosphaera trigonacantha* and *S. streptacantha* respectively (data not shown).

Discussion

Our results showed that intra-genomic rRNA gene variability of Nassellaria and Spumellaria is generally limited, as previously found in tintinnids ciliates (Bachy *et al*., 2013). Such low intra-genomic variability extends in some cases to inter-specific similarities, where the two closely related species *Eucyrtidium cienkowskii* and *Eu. acuminatum* show the same V4 hypervariable region of the 18S rRNA gene. In other species intra-genomic rDNA variability can be very important, as it is the case of *Spongosphaera streptacantha* in which we found two distinct intra-genomic V4 18S rRNA gene regions (<97% similarity, representing 14 different nucleotide positions). These polymorphisms are located in variable helixes of the 18S rRNA gene, not affecting the secondary structure, as already shown in Foraminifera (Weber and Pawlowski, 2014). Taxonomic differences among closely related groups have also been found in Oligotrich and

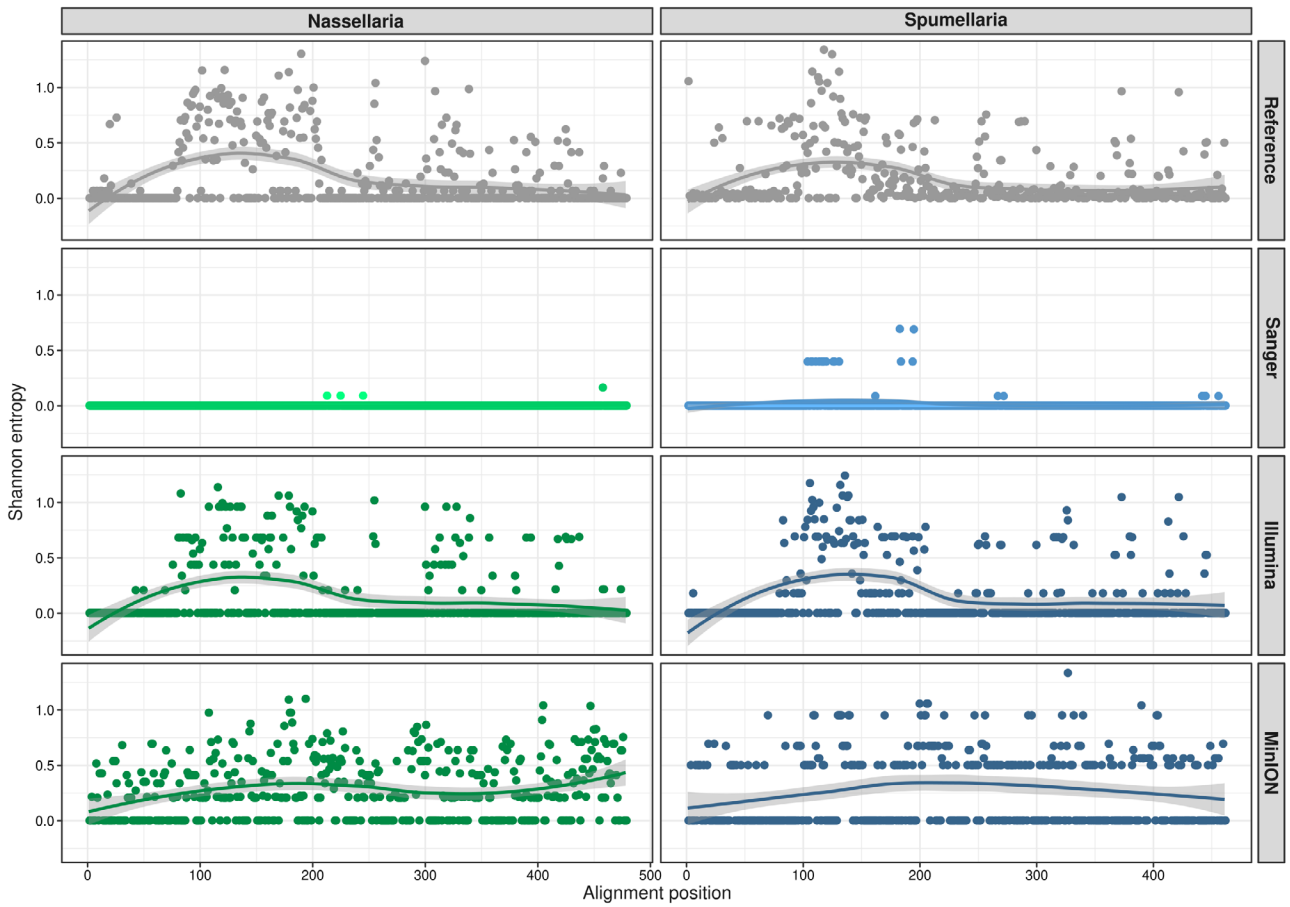

Fig. 5. Shannon entropy analysis for every position (on *x*-axis) of the V4 hypervariable region of the 18S rRNA gene for reference sequences of Nassellaria and Spumellaria (extracted from PR2 v4.14.0) and for Sanger, Illumina and Oxford Nanopore Technologies (MinION) sequencing results aligned all together within the different taxonomic groups. For further results on the near-full rDNA alignment entropy obtained by Sanger and Oxford Nanopore Technologies (MinION) sequencing see supplementary Fig. S7.

Peritrich Ciliates (Gong *et al*., 2013). Most of this intra-genomic variability is, however, overlooked due to the presence of a highly repeated copy that predominates over the low abundant copies, as previously found in Nassellaria or among other orders of Radiolaria such as Acantharia (Decelle *et al*., 2014). Yet, in Acantharia the intra-genomic variability could also become important, finding up to three different OTUs (V9: clustered at 97% and present in two replicates; Decelle *et al*., 2014). Similar studies have shown a relationship between the intra-genomic variability and the number of macronuclei (Zhao *et al*., 2019) and the rDNA copy number in ciliates (Gong and Marchetti, 2019) and in alveolates (Medinger *et al*., 2010). That could explain the taxonomic differences in the intra-genomic variability of Radiolaria, since both Nassellaria and Spumellaria have only one nucleus that tends to be small, whereas Acantharia and Collodaria have several nuclei (Suzuki *et al*., 2009). In the former case, the species tend to show low intra-genomic variability, whereas in the latter case the intra-genomic variability can be relatively high (Decelle *et al*., 2014).

Illumina sequencing has identified *a priori* a very diverse holobiont host gene variability; finding in some Radiolaria specimens up to 10–11 different ASVs (Fig. 2). And when comparing the holobiont community obtained by the three different sequencing methods, Illumina shows a notably more diverse holobiont community. The unexpected presence of taxonomic groups such as Acantharia, Collodaria or Craniata within holobionts constituted by Nassellaria and Spumellaria as hosts, questions the reliability of the so-called 'rare' ASVs in environmental studies. Part of this rare biosphere has been suggested as artefacts inflating diversity estimates (Kunin *et al*., 2010; Bachy *et al*., 2013). Technical issues such as cross-contamination during sequencing or tag-jumping during library preparation also question the rare biosphere over-estimating it (Kircher *et al*., 2012; Schnell *et al*., 2015); as seen in ASVs present in an additional fourth and fifth replicate (i.e. 'ASV-2', 'ASV-5' and 'ASV-

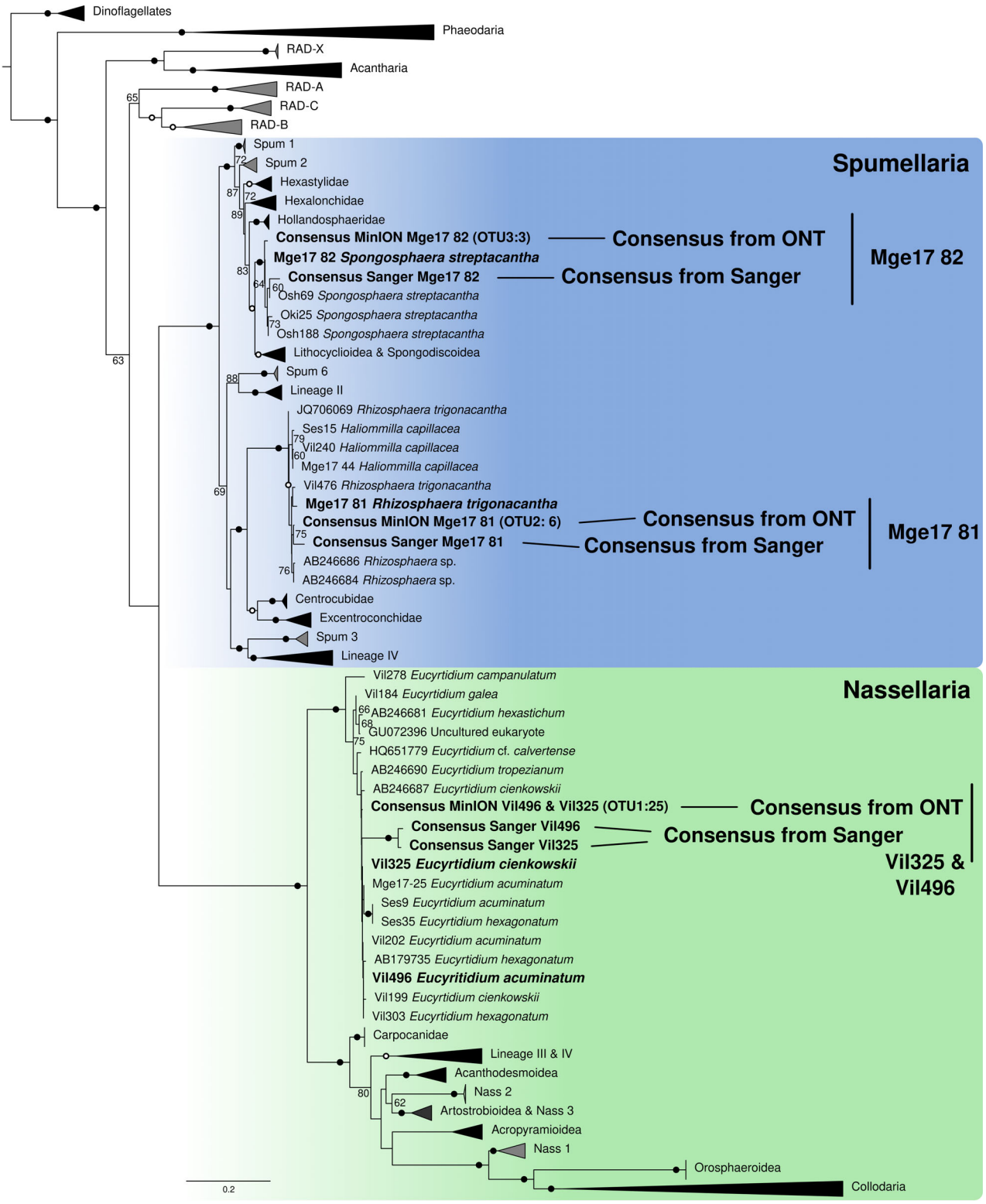

Fig. 6. Molecular phylogeny of consensus Oxford Nanopore Technologies (MinION) sequences, consensus of Sanger results and reference sequences inferred from the concatenated complete 18S and partial 28S (D1–D2 regions) rRNA genes. In bold are shown specimens used in this study (consensus) and reference sequences previously obtained for the same specimens. Numbers after the OTU name of sequences obtained by MinION represent how many raw sequences the consensus was built. The tree was obtained by using a phylogenetic maximum likelihood method implemented in RAxML using the GTR + CAT model of sequence evolution and 1000 rapid bootstraps (BS, shown at the nodes). Bootstrap values below 60 are not shown. Black circles indicate BS ≥99%. Hollow circles indicate BS ≥90%.

7' in Fig. 3). Other ASVs have had full similarities against a reference sequence and have passed the stringent abundance filters we have applied (e.g. Craniata in fig. 2 or 'ASV-47' affiliated to Collodaria in Fig. 3), suggesting the presence of environmental DNA (eDNA). All holobionts were collected in the north-western Mediterranean Sea, yet at varying depths, localities and dates. Numbers of reads have been found to be correlated with the number of nuclei and the cell size (Biard *et al*., 2017; Pitsch *et al*., 2019). And our results indicate that eDNA contamination is showing a bias towards organisms with a higher copy number, such as Collodaria (Biard *et al*., 2017) or the metazoans Craniata. A problem that might be accentuated when the targeted DNA is relatively scarce (i.e. that of Nassellaria) and has to 'compete' for the available space in the flow-cell during sequencing, resulting in large differences of relative abundances, as seen in *Carpocanium obliqua* and *Tetrapyle octacantha* (Fig. 2 and supplementary Fig. S4). Probably, DNA from Acantharia and Collodaria are more prone to be successfully amplified compared to that of Nassellaria and Spumellaria, as seen in differences of DNA amplification success among distinct Foraminifera taxa (Weiner *et al*., 2016). In this study we have used general eukaryotic primers that are equally binding these radiolarian groups. Therefore, further analysis would be needed to assess the possible effect of shell or cell architecture regarding potential impact on DNA extraction and amplification.

These results highlight the need for a careful interpretation of diversity through environmental metabarcoding surveys due to the unpredictable bias found among various taxa (e.g. Weiner *et al*., 2016; Gong and Marchetti, 2019). In some cases, Illumina sequencing from eDNA tends to lead to an under-representation of the environmental diversity (e.g. *Carpocanium obliqua* and *Tetrapyle octacantha* in Fig. 2) and to an overestimate for other taxa (e.g. Collodaria and Craniata). For example, the presence of 'ASV-2', 'ASV-5' or 'ASV-7' in fourth and fifth additional replicates would have been considered as low abundant ASVs in those specific samples in global environmental metabarcoding surveys lacking replicates. Such biases emphasize the need to consider technical replication in metabarcoding surveys to ensure an accurate estimation of the environmental microbial diversity (Prosser, 2010). Besides, understanding other aspects is especially important when working with specific taxa, such as the estimation of the rDNA copy number (Biard *et al*., 2017; Gong and Marchetti, 2019), the analysis of differences in relative abundance (Morton *et al*., 2019) or even exploring previous steps such as DNA extraction bias, amplification primers or intracellular architecture, may contribute to a more accurate interpretation of metabarcoding surveys. However, a big part of the overestimated diversity might also be due to sequencing errors, as previously proposed (Bachy *et al*., 2013; Decelle *et al*., 2014), since the 50% less abundant ASVs accounted for less than 5% of the total reads meaning a high presence of singletons and low abundant clusters. Recently developed tools, such as the 'LULU' algorithm (Frøslev *et al*., 2017), proved to be efficient identifying intra-genomic variability or early PCR errors by post-clustering similar ASVs/OTUs based on co-occurrence patterns.

Illumina sequencing is the method with the highest sequencing depth. After clustering into ASVs, the error rate is the lowest from the three methods used and is randomly distributed. Besides, most of the suspected artefacts produced during Illumina sequencing were removed by applying stringent filters and/or post-clustering methods. Sanger sequencing also showed a limited error rate, yet singletons were found in the V4 18S rRNA gene region of every replicate sequenced, and were later corrected when clustering (Fig. 4). However, towards the end of the amplified fragments the error rate increased substantially. Due to the potential presence of sequencing errors in reference sequences, we suggest that filtering of Illumina ASVs should be done principally based on abundance rather than on identity thresholds. MinION is known to be a sequencing technology with the highest error rate, providing alignment similarities of raw reads around 94%–97% in bacteria (Tyler *et al*., 2018) better than our results with an average of 86% for the best scoring sequences. Yet, recent studies have generated consensus sequences decreasing considerably the error rate and reaching results comparable or better than those obtained by Sanger sequencing (Pomerantz *et al*., 2018; Wurzbacher *et al*., 2019), as seen in our study (>98.9% similarity identity to reference sequences). Despite the high error rate of raw reads from MinION found in our analysis, consensus sequences showed accurate phylogenetic signal with high bootstrap supports. It turns out that the length of the reads overcompetes the randomly generated sequencing errors and brings accurate phylogenetic information. In our study we have used the last 28 h of the MinION sequencing device, largely affecting the life and quality of the nanopores from the MinION flow cell in detecting the ion signal. It is therefore recommended to stop the flow cell according to desired acquisition of reads and not time-sequenced when using different projects in the same flow cell; as seen in Srivathsan *et al*. (2021) where they have successfully re-used a flow cell up to four times.

ONT have the advantage of directly sequencing the DNA strand with no need of PCR amplification, and therefore removing biases associated in these steps that most likely produced part of the artificial environmental diversity found in Illumina generated data. This also

brings the possibility to work with absolute reads, and not relative values providing a more quantitative picture of the genetic community. Yet, rDNA copy number will still be an issue when comparing genetic and morphological diversity. ONT's accuracy is improving since it was first released (Rang *et al*., 2018), and new methods for correcting and analysing ONT results are coming along (e.g. demultiplexing: Wick *et al*., 2018; base-calling: Wick *et al*., 2019, Xu *et al*., 2021; consensus sequence building: Pomerantz *et al*., 2018; Wurzbacher *et al*., 2019; and polishing: Vaser *et al*., 2017). Therefore, MinION could generate fruitful results in the near future for metabarcoding surveys by taking advantage of the extensive 18S reference database available and the high variability and taxonomic resolution of the ITS and the 28S (as seen in environmental studies performed with PacBio, Jamy *et al*., 2019). The high cost-effectiveness of ONT (Cui *et al*., 2020) and the portability and sequencing depth and length of MinION device show fruitful perspectives for its implementation in environmental phylogenetic studies. Along with the high accuracy of circular consensus sequencing of PacBio (Wenger *et al*., 2019) for the high-throughput establishment of reference sequences, environmental phylogenetic studies could move towards the full rDNA surveys.

Acknowledgements

This work was supported by the IMPEKAB ANR 15-CE02-0011 grant and the Brittany Region ARED C16 1520A01. We would like to thank the MOOSE cruise and program for the opportunity of sampling and the facilities given on board, as well as John Dolan for hosting us multiple times at the Laboratoire d'Océanographie of Villefranche-sur-Mer. We are greatly thankful to Charles Bachy and Michal Karlicki for fruitful discussions on the intra-genomic variability and the analysis of Oxford Nanopore Technologies sequencing results, Nicolas Henry for help on the analysis and interpretation of Illumina sequencing results and Anna Karnkowska and two anonymous reviewers for constructive comments on earlier versions of this manuscript.

References

Amaral-zettler, L.A., Mccliment, E.A., Ducklow, H.W., and Huse, S.M. (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.

Bachy, C., Dolan, J.R., López-García, P., Deschamps, P., and 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.

Biard, T., Bigeard, E., Audic, S., Poulain, J., Stemmann, L., and Not, F. (2017) Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. *ISME J* **11**: 1331–1344.

Biard, T., Pillet, L., Decelle, J., Poirier, C., Suzuki, N., and Not, F. (2015) Towards an integrative morpho-olecular classification of the Collodaria (Polycystinea, Radiolaria). *Protist* **166**: 374–388.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005) Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc B Biol Sci* **360**: 1935–1943.

Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., and Holmes, S.P. (2016) DADA2: high resolution sample inference from amplicon data. *Nat Methods* **13**: 581–583.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Caron, D.A., and Hu, S.K. (2018) Are we overestimating Protistan diversity in nature? *Trends Microbiol* **27**: 197–205.

Cui, J., Shen, N., Lu, Z., Xu, G., Wang, Y., and Jin, B. (2020) Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods* **16**: 85.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., *et al*. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

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

Decelle, J., Suzuki, N., Mahé, F., De Vargas, C., and Not, F. (2012) Molecular phylogeny and morphological evolution of the Acantharia (Radiolaria). *Protist* **163**: 435–450.

Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L.F., Nenarokov, S., Massana, R., *et al*. (2018) EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol* **16**: e2005849.

Forster, D., Dunthorn, M., Mahé, F., Dolan, J.R., Audic, S., Bass, D., *et al*. (2016) Benthic protists: the under-charted majority. *FEMS Microbiol Ecol* **92**: fiw120.

Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., and Hansen, A.J. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat Commun* **8**: 1188.

Gong, J., Dong, J., Liu, X., and Massana, R. (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**: 369–379.

Gong, W., and Marchetti, A. (2019) Estimation of 18S gene copy number in marine eukaryotic plankton using a next-generation sequencing approach. *Front Mar Sci* **6**: 219.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., *et al*. (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.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12**: 351–356.

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., *et al*. (2019) Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Mol Ecol Resour* **20**: 429–443.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kircher, M., Sawyer, S., and Meyer, M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**: e3.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **2**: 118–123.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J. (2015) Assessing the performance of the Oxford Nanopore technologies MinION. *Biomol Detect Quantif* **3**: 1–8.

Levy, S.E., and Myers, R.M. (2016) Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* **31**: 95–115.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015) Swarmv2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**: e1420.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., *et al*. (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17**: 4035–4049.

Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C., and 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**: 32–40.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., *et al*. (2019) Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 2719.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C.d., *et al*. (2012) CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol* **10**: e1001419.

Pernice, M.C., Giner, C.R., Logares, R., Perera-bel, J., Acinas, S.G., Duarte, C.M., *et al*. (2016) Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J* **10**: 945–958.

Pitsch, G., Bruni, E.P., Forster, D., Qu, Z., Sonntag, B., Stoeck, T., and Posch, T. (2019) Seasonality of planktonic freshwater ciliates: are analyses based on V9 regions of the 18S rRNA gene correlated with morphospecies counts? *Front Microbiol* **10**: 248.

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L.A., *et al*. (2018) Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**: giy033.

Prosser, J.I. (2010) Replicate or lie. *Environ Microbiol* **12**: 1806–1810.

Rambaut, A. (2016) *FigTree version 1.4.3*. URL http://tree.bio.ed.ac.uk/software/figtree/.

Rang, F.J., Kloosterman, W.P., and de Ridder, J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**: 90.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Sandin, M.M., Biard, T., Romac, S., O'Dogherty, L., Suzuki, N., and Not, F. (2021) A morpho-molecular perspective on the diversity and evolution of Spumellaria (Radiolaria). *Protist* **172**: 125806.

Sandin, M.M., Pillet, L., Biard, T., Poirier, C., Bigeard, E., Romac, S., *et al*. (2019) Time calibrated morpho-molecular classification of Nassellaria (Radiolaria). *Protist* **170**: 187–208.

Santoferrara, L.F., Grattepanche, J.D., Katz, L.A., and Mcmanus, G.B. (2016) Patterns and processes in microbial biogeography: do molecules and morphologies give the same answers? *ISME J* **10**: 1779–1790.

Schnell, I.B., Bohmann, K., and Gilbert, M.T.P. (2015) Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* **15**: 1289–1303.

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S.N., Wong, J., *et al*. (2021) ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biol* **19**: 217.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stoeck, T., Bass, D., Nebel, M., Christen, R., and Meredith, D. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**: 21–31.

Suzuki, N., and Not, F. (2015) Biology and ecology of Radiolaria. In *Marine Protists*, Ohtsuka, S., Suzaki, T., Horiguchi, T., Suzuki, N., and Not, F. (eds). Tokyo: Springer.

Suzuki, N., Ogane, K., Aita, Y., Kato, M., Sakai, S., Kurihara, T., *et al*. (2009) Distribution patterns of the radiolarian nuclei and symbionts using DAPI-fluorescence. *Bull Natl Mus Nat Sci Ser B* **35**: 169–182.

Thornhill, D.J., Lajeunesse, T.C., and Santos, S.R. (2007) Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Mol Ecol* **16**: 5326–5340.

Tyler, A.D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R., and Corbett, C.R. (2018) Evaluation of Oxford nanopore's MinION sequencing

device for microbial whole genome sequencing applications. *Sci Rep* **8**: 10931.
Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746.
Weber, A.A., and Pawlowski, J. (2014) Wide occurrence of SSU rDNA intragenomic polymorphism in foraminifera and its implications for molecular species identification. *Protist* **165**: 645–661.
Weiner, A.K.M., Morard, R., Weinkauf, M.F., Darling, K.F., André, A., Quillévéré, F., *et al*. (2016) Methodology for single-cell genetic analysis of planktonic foraminifera for studies of protist diversity and evolution. *Front Mar Sci* **3**: 1–15.
Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., *et al*. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.
Wick, R.R., Judd, L.M., and Holt, K.E. (2018) Deepbinner: demultiplexing barcoded Oxford nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* **14**: e1006583.
Wick, R.R., Judd, L.M., and Holt, K.E. (2019) Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biol* **20**: 129.
Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., *et al*. (2019) Introducing ribosomal tandem repeat barcoding for fungi. *Mol Ecol Resour* **19**: 118–127.
Xu, Z., Mai, Y., Liu, D., He, W., Lin, X., Xu, C., *et al*. (2021) Fast-Bonito: a faster basecaller for nanopore sequencing. *AILSCI* **1**: 100011. https://doi.org/10.1016/j.ailsci.2021.100011
Yabuki, A., Toyofuku, T., and Takishita, K. (2014) Lateral transfer of eukaryotic ribosomal RNA genes: an emerging concern for molecular ecology of microbial eukaryotes. *ISME J* **8**: 1544–1547.
Zhao, F., Filker, S., Xu, K., Li, J., Zhou, T., and Huang, P. (2019) Effects of intragenomic polymorphism in the SSU rRNA gene on estimating marine microeukaryotic diversity: a test for ciliates using single-cell high-throughput DNA sequencing. *Limnol Oceanogr* **17**: 533–543.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supporting Information.

Fig. S1. Schematic representation of the study design.

Fig. S2. Light microscopy images of specimens used in this study. On the right of each specimen it is indicated whether it was used (tick mark) or not (cross mark) for Sanger + Oxford Nanopore Technologies (MinION) sequencing and/or Illumina, and in grey, a brief description on location and time of isolation. Scale bar (when available; black/white bar) represents 50 μm.

Fig. S3. Tree map of the total number of reads (abundance) of every ASV affiliated to a taxonomic group for each cell obtained by Illumina sequencing. ASVs were processed by dada2.

Fig. S4. Tree map of every ASV affiliated to a taxonomic group (left column) and their total number of reads (abundance; right column) for each cell obtained by Illumina sequencing. ASVs were processed by dada2. Only ASVs present in 3 or more samples and with a total abundance equal or higher than 102 reads (median) were considered.

Fig. S5. Circular plot representing the post-clustering (by 'LULU') of polycystines amplicons clustered with dada2 shown in Figure 4.

Fig. S6. Schematic representation of the 18S rRNA gene 2-dimensional structure (obtained with rnacentral.org/r2dt) from *Spongosphaera streptacantha* (Mge17-82), highlighting the most important regions and positions of intra-genomic variability. Nucleotide size of the highlighted regions represents the relative proportion of the given nucleotide at the given position in the alignment ('x' represents a gap) among the sequences obtained by Sanger sequencing. Nucleotides with an orange background refer to the 14 different positions found in the V4 hypervariable region.

Fig. S7. Shannon entropy analysis for every position (on *x* axis) of Sanger and Oxford Nanopore Technologies (MinION) sequencing results aligned independently for Nassellaria (green) and Spumellaria (blue). Lines represent the tendency of the entropy for each alignment. Black arrows in Sanger boxes represent direction and approximate position of the primers used for Sanger sequencing. Insertions ('-') were not considered due to the differences in length of the sequences and the high amount of insertions produced by the raw sequences from MinION.

Table S1. Number of sequences obtained by the three different sequencing methods used in this study Sanger, MinION and Illumina for every holobiont, specimen and replicate.

Table S2. List of publicly available sequences used in Fig. 6.