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

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

1- Sorbonne Université, 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

*Corresponding author, e-mail: miguelmendezsandin@gmail.com

Abstract

Ribosomal DNA (rDNA) 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 have been focused mainly on the hypervariable regions V4 and V9 of the 18S rDNA gene. Yet, 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 two different high-throughput sequencing platforms: Illumina and Oxford Nanopore Technologies (MinION). Our analysis determined that intra-genomic variability of Nassellaria and Spumellaria is generally low, yet in some Spumellaria specimens we found two different copies of the V4 with a similarity lower than 97%. From 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 environmental rDNA metabarcoding surveys.

Introduction

Ribosomal DNA (rDNA) 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 appearance in all eukaryotes (Pawlowski et al., 2012; del Campo et al., 2018). The 18S rDNA gene, coding for the small sub-unit of the ribosome, has been widely used in molecular environmental surveys, especially the short hyper-variable 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 amount of reads (or amplicons) generated by HTS is normally classified in 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, have led to the development of different clustering methods resulting in finer-scale OTUs that focus in single nucleotide differences (Mahé et al., 2015) or in the correction of sequencing errors based in 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, that 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 a large overestimation of the environmental diversity. Several studies have argued that, quantitatively at the community level, the amount of molecular clusters largely exceeds that of morphological counts (Medinger et al., 2010; Bachy et al., 2013; Santoferrara et al., 2016), leading to hypothesis 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 hyper-variable regions (V4 and V9 most commonly used in protist) of the 18S rDNA gene. Comparing such short hyper-variable regions to, far from exhaustive reference sequences databases (Pitsch et al., 2019) may also contribute to inflating environmental diversity by misidentification of environmental clusters or lack of intra-genomic rDNA variability representation, 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 have 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 amount of reads, inexpensively and highly portable with the MinION device (Levy and Myers, 2016). Despite that the error rate of the ONT reads are improving since it was firstly 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 assess the intra-genomic variability within protists by comparing three different sequencing methods, Sanger, ONT (MinION) and Illumina. We focused our efforts in two groups of Radiolaria, the Nassellaria and Spumellaria (Polycystines). The Radiolaria is an important group of protists in the eukaryotic plankton community contributing for a major fraction of the total reads in environmental molecular surveys (de Vargas et al., 2015; Pernice et al., 2016). However, the number of morpho-species described among Radiolaria taxa (Suzuki and Not, 2015) does not match with the molecular barcodes found in environmental surveys. Nassellaria and Spumellaria environmental diversity lags far behind other radiolarian, despite of 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 among molecular markers and morphological diversity.

Material and Methods

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 3-4 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). Live specimens were imaged using an inverted microscope and thereafter transfer into 1.5 ml Eppendorf tubes containing 50 µl of molecular grade absolute ethanol and finally stored at -20 °C until DNA extraction. DNA was extracted independently for each specimen using the MasterPure Complete DNA and RNA purification Kit (Epicentre) following manufacturer's instructions.

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

Four holobionts were selected to amplify the full length of the rDNA gene (Supplementary material Fig S2), 2 of them belonging to Spumellaria: Mge17-81 (*Rhizosphaera trigonacantha*, Rhizosphaeroidea) and Mge17-82 (*Spongosphaera streptacantha*, Spongosphaeroidea); and 2 to Nassellaria: Vil325 (*Eucyrtidium cienkowskii*, Eucyrtidioidea) and Vil496 (*Eucyrtidium acuminatum*, Eucyrtidioidea). Each holobiont was PCR-amplified in three different technical replicates with the eukaryotic primers SA (AACCTGGTTGATCCTGCCAGT) / D2C-R (CCTTGGTCCGTGTTTCAAGA) (Medlin et al., 1988; Scholin et al., 1994). Amplifications were performed with the enzyme

Phusion® High-Fidelity DNA Polymerase (Finnzymes). The PCR mixture contained 1µL (~0.5ng) of DNA template with 0.35µM of each primer (final concentration), 3% of DMSO and 2X of GC buffer Phusion MasterMix (Finnzymes), in a final volume of 25µL. Amplifications were done in a SIMPLIAMP (Applied Biosystems™) with following the PCR program: initial denaturation step at 98°C for 30 s, 37 amplification cycles of 10 s at 98°C, 30 s at 55°C, 30 s at 72°C and final elongation step at 72°C for 10 minutes. For each of the 12 amplified samples, a 23-base pairs (bp) multiplex identifier or BC sequence was designed and included in one of the primers in order to identify the origin of every single read from the pooled population generated on a single run. Structures of the "BC- primers" were as follows: Primer 1: (5′- Tag + Forward primer -3′); Primer 2: (5′- Reverse primer -3′). To avoid single cell DNA sample contamination, PCR mixtures were carried out under a DNA–free vertical laminar flow hood. Negative and positive controls were added to check if there were no sample contaminations (DNA from operator, reagents or material contamination, etc.). PCR products were then purified using NucleoSpin®Gel and PCR Clean up kit (Macherey-Nagel™). Final elution product was divided in two for sequencing by Sanger (after cloning) and Oxford Nanopore technologies using the MinION device (Laver et al. 2015, Jain et al. 2015).

· Sanger sequencing:

After purification, PCR amplicons were then cloned using TOPO TA Cloning for sequencing kit (Invitrogen) following manufacturer's recommendations. After a first step of 6 min ligation into pCR™4-TOPO® vector, ligated PCR products were then transformed into TOP10 Chemically Competent *Escherichia coli* (ampicillin resistant) by heat-shock treatment at 42°C for 30 s. Cells were then incubated 2h at 37°C with rotation tray at 200 rpm with an enrichment medium. Afterwards, 80 µL of each transformant was spread on selective solid LB-agar medium (Sigma) with 50 µg/mL of ampicillin and incubated overnight at 37°C. On the following day, presence of the insert was verified in obtained colonies by PCR using the primers M13 forward and M13 reverse complementary to cloning site flanking sequences. A total of 24 clones (4 holobionts x 3 replicates x 24 clones = three 96-well plates) from each amplicon were amplified directly with GoTaq polymerase (Promega, Lyon France) in a 25 µL reaction volume using the following PCR parameters: 10 min at 95 °C, 35 amplification cycles of 30 s of denaturation at 95 °C, 30s annealing at 56°C and 1 min extension at 72 °C, with a final elongation 10 min at 72 °C. PCR products were then purified using ExoStar (Illustra™ Exostar™ 1_Step, GE Healthcare Bio-Sciences Corp.). Final PCR products were sent to Macrogen Europe for sequencing using the primers SA (AACCTGGTTGATCCTGCCAGT; Medlin et al., 1988), S69f (AAHCTYAAAGGAAHTGACGG), D1Rf (ACCCGCTGAATTTAAGCATA; Scholin et al., 1994).

· MinION library preparation and sequencing:

Each PCR product was pooled together in equimolar conditions to get 200 ng of final amplicon concentration in 45µL. Quality, size, and concentration of PCR products were assessed on a Bioanalyzer 2100 (Agilent) using a DNA1000 Labchip and on a Qbit fluorometer using a Qbit HS DNA quantification kit (Invitrogen). For this study, we optimized a protocol using Oxford Nanopore Technology (ONT) 1D^2 sequencing chemistry combined with Flowcell R 9.5.

174 Library preparation: DNA products (45 µL amplicon) were treated by an end-repair/dA
175 tailing using NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs)
176 and NEB Blunt/TA Ligase Master Mix (New England Biolabs), and D2 Adapter. Samples
177 were washed using AMPure beads (Agencourt) and 70% EtOH and re-suspended in 45
178 µL water. A first step of adapter ligation was conducted with 25 µL NEB Blunt/TA Ligase
179 Master Mix (New England Biolabs), and 2.5µL $1D^2$ Adapter, and then carried out with
180 5µL BAM and 50µL of Blunt/TA ligase. A second washing step was done using AMPure
181 beads (Agencourt) and 70% EtOH and re-suspended in 15 µL water of final elution.

182 Sequencing: Pre sequence-library was prepared following ONT protocols by mixing 12
183 µL amplicon, 2.5 µL water, 25.5 µL LBB and 35 µL RBF. Flow cell R 9.5 was inserted in
184 the MinION frame and priming port was loaded with 800 µL of priming mix (576 µL RBF
185 and 624 µL water) with SpotON cover closed. We additionally loaded 200 µL of priming
186 mix into the flowcell via the priming port, avoiding the introduction of air bubbles.
187 75µL of sample library was then added to SpotOn port via dropwise fashion. Finally, we
188 covered SpotOn and priming ports, close the MinION lid and open the MinKNOW GUI
189 software to proceed with sequencing using a half of a total flow cell R 9.5. It is
190 important to note that the flowcell was used during its last 28h (20h → 48h), after a
191 prior independent experiment (data being analyzed).

192 · Illumina sequencing:
193 In total 8 holobionts were selected to amplify the V4 region (~380pb) of the 18S rDNA
194 gene (Supplementary material Fig S2), 4 belonging to Spumellaria: Mge17-81 (common
195 to previous section), Mge17-82 (common to previous section), Vil480 (*Tetrapyle*
196 *octacantha*, Pylonioidea), Vil497 (*Arachnospongus varians*, Liosphaeroidea); and 4 to
197 Nassellaria: Mge17-9 (*Extotoxon undulatum*, Artostrobioidea), Mge17-124 (*Carpocanium*
198 *obliqua*, Carpocaniidae), Vil490 (*Pterocorys* cf. *zanclea*, Pterocorythoidea), Vil496
199 (common to previous section). Specimen Vil325 (from previous section) had no DNA left
200 for further experiments and therefore was not possible to include in this section. Each
201 holobiont was PCR-amplified with the eukaryotic primers TAR-EukF1 (5'-
202 CCAGCASCYGCGGTAATTCC-3') / TAR-EukR3 (5'-ACTTTCGTTCTTGATYRA-3') (Stoeck et al.
203 2010). The forward primer had a barcode adapted for Illumina sequencing and each
204 holobiont was amplified with 3 different barcode-adapted-primers to get 3 technical
205 sequencing replicates. PCR reactions contained 1x MasterMix GC Phusion High-Fidelity
206 DNA Polymerase (Finnzymes), 0.35µM of each primer, 3% dimethylsulphoxide and 1µL
207 of DNA in a final volume of 25µl. The PCR program had an initial denaturation step at
208 98°C during 30 s, followed by 15 cycles of 10 s at 98°C, 30 s at 53°C and 30 s at 72°C,
209 then 22 similar cycles but with 48°C annealing temperature, and a final elongation step
210 at 72°C for 10 min. Vil325 gave no positive reactions and therefore was no possible to
211 be included in this step. Polymerase chain reaction, in triplicates, were purified with
212 NucleoSpin Gel and PCR Clean-Up kit (Macherey-Nagel), and quantified with the Quant-
213 It PicoGreen double stranded DNA Assay kit (Invitrogen). About 500 ng of pooled
214 amplicons were sent to Fasteris (https://www.fasteris.com, Switzerland) for Illumina
215 sequencing on a Miseq nano V2 2x250.

216 Sequencing results from this study have been deposited in the European Nucleotide
217 Archive (ENA) at EMBL-EBI under accession number PRJEB47886.

218 **Quality checking and similarity of sequences/amplicons**

219 In order to compare the output of the three different sequencing results, similar and
220 broadly known methods were used to quickly check the quality of the
221 sequences/amplicons obtained. Firstly, raw sequences/amplicons were measured in
222 base pairs (bp) length. Thereafter, sequences/amplicons were compared against
223 reference sequences by local alignment (BLAST; using NCBI as reference sequences)
224 for Sanger and MinION sequencing results and global alignment (vsearch; Rognes et
225 al. 2016; using PR2 v4.14.0 database as reference sequence, Guillou et al. 2013) for
226 Illumina results. This difference in the alignment algorithm used is due to the big
227 differences in the sequence length and the method mostly used for each of the
228 sequencing results. Finally, in order to obtain sequences per holobiont, Sanger reads
229 belonging to the same replicate were concatenated, MinION reads were de-
230 multiplexed with cutadapt (Martin, 2011) and Illumina reads were clustered as
231 described in the next section.

232 **Clustering of V4 hyper-variable region**

233 Two different pipelines were used to cluster the V4 hyper-variable region: a distance-
234 based method and an error-based method, depending on the sequencing nature of
235 the sequences. The V4 region of sequences coming from Sanger and MinION
236 sequencing was extracted with cutadapt (Martin, 2011) using the primer set used to
237 amplify the V4 region (see 'Amplification and sequencing' section: TAR-EukF1 / TAR-
238 EukR3) and allowing 10% of mismatch for Sanger sequences and 20% of MinION
239 sequences. Sequence extracts were later clustered using swarm (Mahé et al., 2015)
240 with 1, 2 and 3 differences. Reads coming from Illumina sequencing were clustered
241 into ASVs using DADA2 and the pipeline described in Callahan et al. (2016). Final
242 amplicons were compared against PR2 v4.14.0 (Guillou et al. 2013). Note DADA2 was
243 not used for Sanger sequences or MinION output since the error-correction algorithm
244 relies on number of reads.

245 **Consensus sequence building**

246 Sanger sequences were aligned using MAFFT v7.395 (Katoh and Standley, 2013) with
247 the G-INS-i algorithm ('--globalpair') and 1000 iterative refinement cycles. Final
248 consensus sequences were produced with a custom script (see
249 'alignmentConsensus.py' in github.com/MiguelMSandin/IntraGenomic-variability) using
250 a 70% consensus threshold and a 30% threshold for base detection. Minion sequences
251 taxonomically assigned to Polycystines by BLAST were used to build consensus
252 sequences according to Wurzbacher et al. (2019), using an OTU clustering cutoff of 30%
253 due to the bad quality of the reads with the script *consension*
254 (https://microbiology.se/software/consension/).

255 **Phylogenetic analysis**

256 Consensus sequences coming from MinION and Sanger sequencing were aligned
257 against a reference alignment using MAFFT v7.395 (Katoh and Standley, 2013) with a G-
258 INS-i algorithm ('--globalpair') and 1000 iterative refinement cycles. The reference
259 alignment is composed of reference sequences (extracted from Decelle et al., 2012;
260 Biard et al., 2015; Sandin et al. 2019 and Sandin et al. 2021; Supplementary material

Table S1) and consensus sequences obtained by Sanger sequencing in this study. Automatic alignment outputs were manually checked and corrected in AliView v1.26 (Larsson, 2014). Final alignments were trimmed automatically using trimal (Capella-Gutiérrez et al., 2009) with a 30% gap threshold. The final data set contains XX taxa and XXXX positions. A phylogenetic analysis done by RAxML and GTR+Gamma and GTR+CAT evolutionary models with 1000 rapid bootstraps was used to infer the phylogenetic position of the consensus sequences. Final trees were visualized and edited with FigTree version 1.4.3 (Rambaut 2016).

Entropy analysis

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) with a G-INS-i algorithm and 1000 iterative refinement cycles for high accuracy. For every position of each alignment, Shannon entropy was calculated using a custom script (see 'alignmentEntropy.py' in github.com/MiguelMSandin/IntraGenomic-variability), without considering insertions or deletions ("-") due to the incompleteness of Nassellaria sequences. For Sanger and MinION sequences, a reference alignment was build including only Sanger sequences as previously described, and then MinION sequences were aligned to the reference alignment. Later Sanger sequences were removed in order to obtain 4 different alignments: Nassellaria from Sanger, Nassellaria from MinION, Spumellaria from Sanger and Spumellaria from MinION. Further entropy values were calculated for the 4 alignments, after removing columns with only gaps. The entropy of the full V4 region was measured in an independent analysis to compare the three different sequencing technologies. The full V4 region was extracted as previously described (cutadapt) from Sanger and MinION sequencing results and sequences assigned to Nassellaria and Spumellaria from the PR2 v4.14.0 (after removing highly divergent sequence assigned to Orosphaeridae and Collodaria and duplicated sequences). Final datasets of Nassellaria and Spumellaria independently included sequences from 4 different origins: reference sequences from PR2 v4.14.0 and sequences obtained in this study by Sanger, MinION and Illumina sequencing.

Results

Quality of sequencing results

In total 834 reads (representing 737,817 bases) were successfully retrieved from Sanger sequencing for the 4 holobionts (from the total 864 sequences sequenced: 4 holobionts, 3 replicates per holobiont, 24 sequences per replicate and 3 reads (primers) per sequence; see Supplementary material Fig S1 for a schematic representation). 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).

MinION sequencing resulted in 864 total sequences (representing a total of 1,645,774 bases). The low throughput obtained during the last 28h of the flowcell live (20h → 48h) contrast with the high number of sequences obtained during the first 20h (0h → 20h) in an independent experiment, where there were a total of 225,573 sequences

305 (representing 296,072,423 total bases; data being analyzed). Approximately, during the
306 last 28h of the flowcell there were sequenced 0.38% of the total sequences (or 0.56% of
307 the total bases) obtained by the same flowcell. These 864 sequences were compared
308 against NCBI database by BLAST tool and 81 had no possible matching, 185 were
309 assigned to bacteria, 593 to eukaryotes and 5 matching different domains (e.g. the
310 same sequence was matching virus, uncultured prokaryote and uncultured organism).
311 The remaining 783 sequences had an average sequence length of 2043.48 bp (±
312 1143.42) bp, with a median of 2746 bp and an average identity of 86.01 (± 2.91) % with
313 a reference sequence. De-multiplexing MinION sequences resulted in 55 sequences
314 among the different replicates. After de-multiplexing, the average sequence length
315 increased to 2641.82 bp (± 681.80) bp and the median length to 2921 bp although the
316 average identity remained at values of 86.48 (± 2.33) % similarity on average with a
317 reference sequence (Fig. 1, MinION).

318 Regarding Illumina sequencing, a total of 1,019,196 sequences (representing
319 254,799,000 bases) were sequenced in both the R1 and the R2 files. On average 12,135
320 (± 6,470) reads were obtained for each replicate. These reads were merged resulting
321 on 10,256 (± 6,278) amplicons on average per replicate and 66,913 total unique
322 amplicons. They had an average length of 375.3 (± 15.89) bp, a median of 382 bp and
323 an average identity of 99.28 (± 2.26) % with a reference sequence (Fig. 1, Illumina).

324 **Taxonomic assignation and diversity of reads**

325 Sanger sequenced amplicons of the different parts of the ribosomal genes belonging
326 to the same replicate were concatenated resulting in 287 sequences with an average
327 length of 2570.7 (± 296.97) bp. This covers about the first ~1000 bp of the 18S rDNA
328 gene (primer SA), the last ~600-800 bp of the 18S and the ITS1 rDNA gene (primer S69f)
329 and the regions D1 and D2 from the 28S rDNA gene (primer D1R).

330 In *Eucyrtidium cienkowski* (Vil325) sequences were assigned to Nassellaria (56), Ciliates
331 (11) and Tunicates (4); in *Eu. acuminatum* (Vil496) to Nassellaria (53), Chrysophyceae (14)
332 and diatoms (5); in *Rhizosphaera trigonacantha* (Mge17-81) to Spumellaria only (72) and
333 in *Spongosphaera streptacantha* (Mge17-82) the community was the most diverse with
334 Spumellaria (63), fungi (5), ciliates (Apicomplexa: 2) and dinoflagellates (1) (Fig. 2,
335 Sanger). From the de-multiplexed sequences extracted from MinION sequencing, in *E.*
336 *cienkowski* was found only Nassellaria (9); in *E. acuminatum* was found Nassellaria (16),
337 diatoms (6), chrysophyceae (4) and fungi (1); in *R. trigonacantha*, only Spumellaria (8), as
338 in Sanger sequencing, and in *S. streptacantha* a similar composition as in Sanger,
339 Spumellaria (7), dinoflagellates (3), alveolates (1) and fungi (1) (Fig. 2, MinION).

340 Amplicons of Illumina were clustered using an error-based method (dada2), in total
341 153 ASVs (with a total of 570,041 reads) were found. These ASVs had an average length
342 of 379.2 (± 8.99) bp, a median length of 378 bp and an average similarity identity of
343 99.70 (± 5.97) %. The most diverse groups are dinoflagellates (28 ASVs), Spumellaria (23
344 ASVs), Nassellaria (19 ASVs) and Fungi (14 ASVs mostly within Agaricomycotina and
345 Ustilaginomycotina) (Fig. 2, MinION). Despite the big diversity within each holobiont,
346 the number of reads is highly dominated by 1, 2 or 3 ASVs belonging to the host or the
347 symbiotic algae (Supplementary material Fig S3). Due to the great taxonomic diversity
348 found within holobionts by Illumina sequencing, only ASVs present in at least 3

samples (triplicates in the PCR), with a total abundance equal or higher than the median value of the abundance (that is 102 reads) and assigned until at least the taxonomic rank “Order” were considered. We have chosen stringent thresholds in an attempt to remove all artefacts and/or contaminations, understanding that part of the targeted 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 4 holobionts, an ASV affiliated to Acantharea present in 3 holobionts and 1/2 ASVs affiliated to Craniata present in every sample (Supplementary material Fig S4). When exploring in deeper detail the ASVs affiliated to Polycystines and their abundance and distribution among samples after applying the stringent thresholds, it is possible to find up to 3 different and highly abundant ASVs within the same cell (e.g. *Spongosphaera streptacantha* or *Arachnospngus varians*; Fig. 3). Some of the most abundant ASVs were also found in a fourth and fifth sample; as for “ASV-2”, “ASV-5” and “ASV-7”, ASVs from the Nassellaria hosts *Pterocorys zanclea*, *Extotoxon undulatum* and *Eucyrtidium acuminatum* respectively. “ASV-2” was present in the three replicates of *E. acuminatum* and in one replicate of *Carpocanium obliqua* and *Rhizosphaera trigonacantha*, a Spumellaria host, 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. This ASV, is also found in *C. obliqua*, *T. octacantha* and *A. varians*, it never appears within the same three samples from an independent holobiont and 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 rDNA 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 it is detectable the same genetic diversity within the holobiont. In total 52, 56, 59 and 68 sequences from Sanger sequencing were extracted belonging to *Eucyrtidium acuminatum*, *E. cienkowski*, *Spongosphaera streptacantha* and *Rhizosphaera trigonacantha* respectively. The same unique sequence was found to be the most abundant in both *E. acuminatum* and *E. cienkowski* (54 and 47 reads respectively) being exactly identical despite of being morphologically identified in different species, and up to 5 other unique sequences were found with a very low abundance (2, 2, 1, 1 and 1 reads, green dots on Fig. 4 under unique reads). These sequences were very similar, showing similarities within them close to 100% with a maximum of one base of difference (boxplots on Fig. 4 under unique reads). Similar patterns are seen in the Spumellaria *R. trigonacantha*, with one sequence found 67 times and only one sequence different with one read. In contrast, *S. streptacantha* shows up to 3 different reads relatively abundant (16, 13 and 13 reads) and 8 other sequences with a lower abundance, having a similarity among them of 96.1%. It is important to note, that these three sequences were sequenced in the three different

replicates (PCR reactions). When these sequences were clustered with swarm and 1 difference of threshold, *Eu*. *acuminatum* and *Eu. cienkowski* share the same amplicon. In *R. trigonacantha* the single sequence was grouped in the unique amplicon found for this holobiont, and up to 3 different amplicons were found in *S. streptacantha* (still with a similarity among these amplicons of ~96%). Increasing the difference threshold to 2 differences the amplicons in *S. streptacantha* remain comparable to 1 difference. When using a clustering of 3 differences, the number of amplicons in *S. streptacantha* decreases to 2, with low changes in the similarities among them (96.4%, representing 14 different aligned positions in 385 an 382 bp sequences long). Yet one amplicon largely dominates the other (51 reads against 8 reads). Same protocol was implemented in MinION sequences in order to extract the V4 hyper-variable region of the 18S rDNA (data not shown). In total 4 sequences were extracted from *R. trigonacantha*, 2 from *S. streptacantha* , 6 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%).

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 have low evenness and appear with low entropy (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, and also with a tendency to appear towards hyper-variable regions. In general, concatenated sequences from Sanger sequencing belonging to Nassellaria showed low Shannon entropy values among them (Supplementary material Fig S5). Towards the end of the primers the entropy slightly increased, meaning that most probably are suspected to be errors. Especially at the end of the region D2 of the 28S rDNA 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 than Nassellaria (Supplementary material Fig S5). In contrast, *Spongosphaera streptacantha* shows regions of high entropic values at around the position 750 (V4 region from the 18S rDNA gene), along the beginning of the 28S RDNA 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 S5). Alignments of sequences coming from MinION have in general a greater number of positions than alignments of sequences coming from Sanger sequencing (~922.25 bp more in 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 rDNA). These sequences had high entropy values all along the rDNA alignment. With the exception of the 18S rDNA 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 hyper-variable region of the 18S rDNA 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 5 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 hyper-variable 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 a near-0 entropy values, with the exception of a region between ~100-200 bp in Spumellaria showing higher entropy values (corresponding to *Spongosphaera streptacantha*). Illumina ASVs follow same patterns as the reference alignment, with the highest entropy values at around the position ~150. Despite the poor trends found for the full length of the rDNA for MinION sequences (Supplementary material Fig S5), 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.

We performed a last attempt to test whether the high-error rate from MinION is affecting further downstream analysis or is in the contrary possible to be addressed. Sequences coming from Sanger and MinION were clustered into consensus sequences and were aligned in a phylogenetic tree along reference sequences (Supplementary material Table S1). Despite the high error rate of MinION sequences, the random distribution of the errors and the long reads bring accurate phylogenetic information when processed into consensus sequences (Fig. 6). Consensus sequences from MinION are phylogenetically sister to the consensus sequences coming from Sanger showing high phylogenetic support and short branches. Only one sequence (from *Spongosphaera streptacantha*; Mge17-82) shows a relatively long branch matching with the holobiont host with the highest intra-genomic variability. Consensus sequences from Sanger sequencing have a similarity >99% against reference sequences except for *Eucyrtidium cienkowski* (97.32%). Consensus sequences from MinION have a similarity identity against reference sequences of 98.16%, 96.6% and 90.22% for *Eucyrtidium* sp., *Rhizosphaera trigonacantha* and *S. streptacantha* respectively (data not shown).

Discussion

Our results showed that intra-genomic rDNA gene variability of Nassellaria and Spumellaria is generally limited, showing in some cases the same V4 rDNA hyper-variable region between different morpho-species (yet belonging to the same genus; *E. cienkowskii* and *E. acuminatum*), as previously found in tintinnids ciliates (Bachy et al., 2013). Although in some species intra-genomic rDNA gene variability can be very

important, as it is the case of *Spongosphaera streptacantha* in which we found two distinct intra-genomic V4 rDNA regions (~97% similarity, representing 14 different nucleotide positions). Such taxonomic differences among closely related groups have also been found in Oligotrich and 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). However, in Acantharia the intra-genomic variability could also become important, finding up to 3 different OTUs (V9: clustered at 97% and present in 2 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 between 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 tend to be small whereas Acantharia and Collodaria have several nuclei (Suzuki et al., 2009). In the former case, they tend to show low intra-genomic variability, whereas in the later case can be relatively high (Decelle et al., 2014).

Despite the low intra-genomic variability found within the hosts, Illumina sequencing has identified *a priori* a very diverse host gene variability; finding in some hosts 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 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). Other ASVs present in a fourth or fifth sample (i.e.: “ASV-2”, “ASV-5” and “ASV-7” in Fig. 3) also question technical issues such as cross-contamination during sequencing (Kircher et al., 2012) or tag-jumping during library preparation (Schnell et al., 2015). 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) since 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). 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 low abundant (*i.e.* that of Nassellaria) and have 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 figure S4). Probably, DNA from Acantharia and Collodaria are more prone to be successfully amplified regarding 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

must assess the effect on differences in shell or cell architecture regarding DNA extraction and amplification.

These results highlight the need for a careful interpretation of environmental metabarcoding surveys due to the taxonomic bias found among different 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 over-estimation for some other taxa (e.g.: Collodaria and Craniata). Most of these biases and technical problems cannot be identified and sorted out in environmental metabarcoding surveys, most frequently due to the lack of replicates, leading to poor estimations of the environmental molecular diversity. For example, the presence of “ASV-2”, “ASV-5” or “ASV-7” in a fourth and fifth sample would have been considered as low abundant ASVs in those specific samples in global environmental metabarcoding surveys lacking replicates. This emphasizes the need for considering replication in metabarcoding surveys to ensure an accurate estimation of the environmental microbial diversity (Prosser, 2010), being technical replicates more important than biological replicates. Besides, understanding other aspects may help the correct interpretation of metabarcoding surveys, 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. 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.

Illumina sequencing showed to be the method with the highest sequencing depth. After clustering into ASVs, the error rate showed to be the lower from the three methods and randomly distributed. Besides, most of the suspected artefacts produced during Illumina sequencing were removed applying stringent filters. Sanger sequencing also showed a limited error rate, yet singletons were found in the V4 rDNA region of every sample sequenced, and were later corrected when clustering (Fig. 4). However, towards the end of the primers the error rate increased substantially. For this reason, we suggest that filtering of Illumina ASVs should be done more importantly based in abundance than identity thresholds, especially for the V9 region that is located near the end of general and specific primers. On the other side, MinION is the technique with the highest error rate, finding alignment similarities of the 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. Although, 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). Despite the high error rate of raw reads from MinION found in our analysis, consensus sequences showed accurate phylogenetic signal with high bootstrap supports. Therefore it turns out that the length of the reads over-competes the randomly generated sequencing errors and brings accurate phylogenetic information. In our study we have used the last 28h of the MinION

sequencing device, largely affecting the life and quality of the nanopores from the MinION flowcell in detecting the ion signal. It is therefore recommended to stop the flowcell according to dessired acquisition of reads and not time sequenced when using different projects in the same flowcell; as seen in Srivathsan et al. (2021) where they have successfully re-used a flow cell up to four times. Yet, when looking at the holobiont community sequenced in our study, the results obtained with MinION are comparable to those obtained by Sanger sequencing (Fig. 2).

ONT have the advantage of directly sequence the DNA strand with a theoretical 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. This also brings the possibility to work with absolute reads, and not relative values providing a more quantitative picture of the genetic community. Although rDNA copy number will still be an issue when comparing genetic and morphological diversity. ONT's accuracy is improving since it was firstly 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. 2020; consensus sequence building: Pomerantz et al., 2018; Wurzbacher et al., 2019). Therefore, if the error rate is overcome, MinION could generate fruitful results in the near future for metabarcoding surveys by taking advantage of the extensive work of the 18S reference sequences already done 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 of 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.

Acknowledgments

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 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 for constructive comments on earlier versions of this manuscript.

References

Amaral-zettler, L.A., Mccliment, E.A., Ducklow, H.W., 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., 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., Not, F. (2017) Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. Nat. Publ. Gr. 1–42

Biard, T., Pillet, L., Decelle, J., Poirier, C., Suzuki, N., 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., 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., Holmes, S.P. (2016) DADA2: High resolution sample inference from amplicon data. Nat. Methods 13:581–583

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973

Caron, D.A., Hu, S.K. (2018) Are We Overestimating Protistan Diversity in Nature? Trends Microbiol. 27(3):197-205

Cui, J., Shen, N., Lu, Z., Xu, G., Wang, Y., 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., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. Science. 348:6237, 1261605

Decelle, J., Romac, S., Sasaki, E., Not, F., Mahé, F. (2014) Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-Throughput Sequencing. PloS One 9(8):e104297

Decelle, J., Suzuki, N., Mahé, F., Vargas, C. De, Not, F. (2012) Molecular Phylogeny and Morphological Evolution of the Acantharia (Radiolaria). Protist 163(3):435-50

Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L.F., Nenarokov, S., Massana, R., Guillou, L., Simpson, A., Berney, C., de Vargas, C., Brown, M.W., Keeling, P.J., Parfrey, L.W. (2018) EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLoSBiol 16(9):e2005849

Forster, D., Dunthorn, M., Mahé, F., Dolan, J.R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Edvardsen, B., Egge, E., Eikrem, W., Gobet, A., Kooistra, W.H.C.F., Logares, R., Massana, R., Montresor, M., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Richards, T.A., Santini, S., Sarno, D., Siano, R., Vaulot, D., Wincker, P., Zingone, A., Vargas, C. de, Stoeck, T. (2016) Benthic protists: the under-charted majority. FEMS Microbiol. Ecol. 92(8):fiw120

Gong, J., Dong, J., Liu, X., 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(3):369-79

Gong, W., 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., 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., Boutte, C., Burgaud, G., De Vargas, C., Decelle, J., Del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R. (2013) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41(Database issue):D597-604

666 Jain, M., Olsen, H.E., Paten, B., Akeson, M. (2015) Improved data analysis for the MinION
667 nanopore sequencer. Nat Methods 12:351–356
668 Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bass, D., Burki, F. (2019)
669 Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically
670 resolve environmental diversity. Mol Ecol Resour. 20:429–443
671 Katoh, K., Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7:
672 Improvements in performance and usability. Mol. Biol. Evol. 30,4:772–780
673 Kircher, M., Sawyer, S., Meyer, M. (2012) Double indexing overcomes inaccuracies in multiplex
674 sequencing on the Illumina platform. Nucleic Acids Res. 40(1):e3
675 Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P. (2010) Wrinkles in the rare biosphere:
676 Pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ.
677 Microbiol. 2(1):118-23
678 Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large data
679 sets. Bioinformatics. 30(22):3276-3278
680 Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J. (2015)
681 Assessing the performance of the Oxford Nanopore technologies MinION. Biomol Detect
682 Quantif. 3:1–8
683 Levy, S.E., Myers, R.M. (2016) Advancements in Next-Generation Sequencing. Annu. Rev.
684 Genomics Hum. Genet. 31;17:95-115
685 Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M. (2015) Swarmv2: Highly-scalable
686 and high- resolution amplicon clustering. PeerJ 3:e1420
687 Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing
688 reads. EMBnet.journal 17.1,10–12
689 Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R.,
690 Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D.,
691 Guillou, L., Jaillon, O., Kooistra, W.H.C.F., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski,
692 J., Pernice, M.C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R.,
693 Simon, N., Stoeck, T., Vaulot, D., Zingone, A., Vargas, C. de (2015) Marine protist diversity in
694 European coastal waters and sediments as revealed by high-throughput sequencing.
695 Environ. Microbiol. 17(10):4035-49
696 Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C., Boenigk, J. (2010)
697 Diversity in a hidden world: Potential and limitation of next-generation sequencing for
698 surveys of molecular diversity of eukaryotic microorganisms. Mol. Ecol. 19 Suppl 1(Suppl
699 1):32-40
700 Medlin, L., Elwood, H.J., Stickel, S., Sogin, M.L. (1988) The Characterization of enzymatically
701 amplified eukaryotic 16S-like rRNA-coding regions. Genetics 71(2):491-9
702 Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K.,
703 Knight, R. (2019) Establishing microbial composition measurement standards with reference
704 frames. Nat. Commun. 10(1):2719
705 Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C. dric, Bowser, S.S., Cepicka, I.,
706 Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirku, M.,
707 Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukes, J., Mann, D.G., Mitchell, E.A.D.,
708 Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G.B., Smirnov, A. V., Spouge, J.L., Stern,
709 R.F., Stoeck, T., Zimmermann, J., Schindel, D., Vargas, C. de (2012) CBOL Protist Working
710 Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS
711 Biol 10(11):e1001419
712 Pernice, M.C., Giner, C.R., Logares, R., Perera-bel, J., Acinas, S.G., Duarte, C.M., Gasol, J.M.,
713 Massana, R. (2016) Large variability of bathypelagic microbial eukaryotic communities across
714 the world's oceans. ISME J. 10(4):945-58

Pitsch, G., Bruni, E.P., Forster, D., Qu, Z., Sonntag, B., Stoeck, T., 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., Barrio-Amorós, C.L., Salazar-Valenzuela, D., Prost, S. (2018) Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. Gigascience 7(4):giy033

Prosser, J.I. (2010) Replicate or lie. Environ. Microbiol. 12(7):1806-10

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

Rang, F.J., Kloosterman, W.P., de Ridder, J. (2018) From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19(1):90

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ. Oct 18;4:e2584

Sandin, M.M., Pillet, L., Biard, T., Poirier, C., Bigeard, E., Romac, S., Suzuki, N., Not, F. (2019) Time Calibrated Morpho-molecular Classification of Nassellaria (Radiolaria). Protist 170(2):187-208

Sandin, M.M., Biard, T., Romac, S., O'Dogherty, L., Suzuki, N., Not, F. (2021) A Morpho-molecular Perspective on the Diversity and Evolution of Spumellaria (Radiolaria). Protist. 172(3):125806

Santoferrara, L.F., Grattepanche, J.D., Katz, L.A., 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., Gilbert, M.T.P. (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. Mol. Ecol. Resour. 15(6):1289-303

Scholin, C. a, Herzog, M., Sogin, M., Anderson, D.M. (1994) Identification of group-specific and strain- specific genetic-markers for globally distributed Alexandrium (Dinophyceae). II. Sequence analysis of a fragment of the LSU rRNA gene. J. Phycol. 30:6

Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R. (2021) ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. BMC Biol. 19(1):217

Stoeck, T., Bass, D., Nebel, M., Christen, R., Meredith, D. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol. 19 Suppl 1:21-31

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

Suzuki, N., Ogane, K., Aita, Y., Kato, M., Sakai, S., Kurihara, T., Matsuoka, A., Ohtsuka, S., Go, A., Nakaguchi, K., Yamaguchi, S., Takahashi, T., Tuji, A. (2009) Distribution Patterns of the Radiolarian Nuclei and Symbionts Using. Bull. Natl. Mus. Nat. Sci., Ser. B 35(4):169–182

Thornhill, D.J., Lajeunesse, T.C., Santos, S.R. (2007) Measuring rDNA diversity in eukaryotic microbial systems: How intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. Mol. Ecol. 16(24):5326-40

Tyler, A.D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R., Corbett, C.R. (2018) Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci. Rep. 8, 10931

Weiner, A.K.M., Morard, R., Weinkauf, M.F., Darling, K.F., André, A., Quillévéré, F., Ujiie, Y., Douady, C.J., de Vargas, C., Kucera, M. (2016) Methodology for single-cell genetic analysis of planktonic foraminifera for studies of protist diversity and evolution. Fontiers Mar. Sci. 3:1–15

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Toepfer, A., Chin, C.-S., Alonge, M., Mahmoud, M., Qian, Y., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W. (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., Holt, K.E. (2018) Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLOS Comput. Biol. 14(11):e1006583

Wick, R.R., Judd, L.M., 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., Kagami, M., Nilsson, R.H. (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., Zhang, L., Meng, X, Mafofo, J., Zaher, W.A., Li, Y., Qiao, N. (2020) Fast-Bonito: A Faster Basecaller for Nanopore Sequencing. bioRxiv 2020.10.08.318535

Yabuki, A., Toyofuku, T., Takishita, K. (2014) Lateral transfer of eukaryotic ribosomal RNA genes: An emerging concern for molecular ecology of microbial eukaryotes. ISME J. 8(7):1544-7

Zhao, F., Filker, S., Xu, K., Li, J., Zhou, T., 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

Tables and Figure Legends

Figure 1. Boxplot summarizing sequencing results of Sanger, Oxford Nanopore Technologies (MinION) and Illumina sequencing; top row shows the sequence length (in base pairs; 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.

Figure 2. Taxonomic affiliation of sequences/amplicons obtained by Sanger (after concatenation of the primers), 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.

Figure 3. Circular plot representing the log transformed abundance of polycystines amplicons (lower half of the circles) clustered with dada2 and sample affiliation (upper half of the circles). Only amplicons affiliated to polycystines, present in 3 or more samples and with a total abundance equal or higher than 102 reads (median) were considered.

Figure 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.

Figure 5. Shannon entropy analysis for every position (on *x* axis) of the V4 hyper-variable region of the 18S rDNA 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 figure S5.

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

819 **Supplementary Data**

820 **Table S1**. List of publicly available sequences used in Fig. 6.

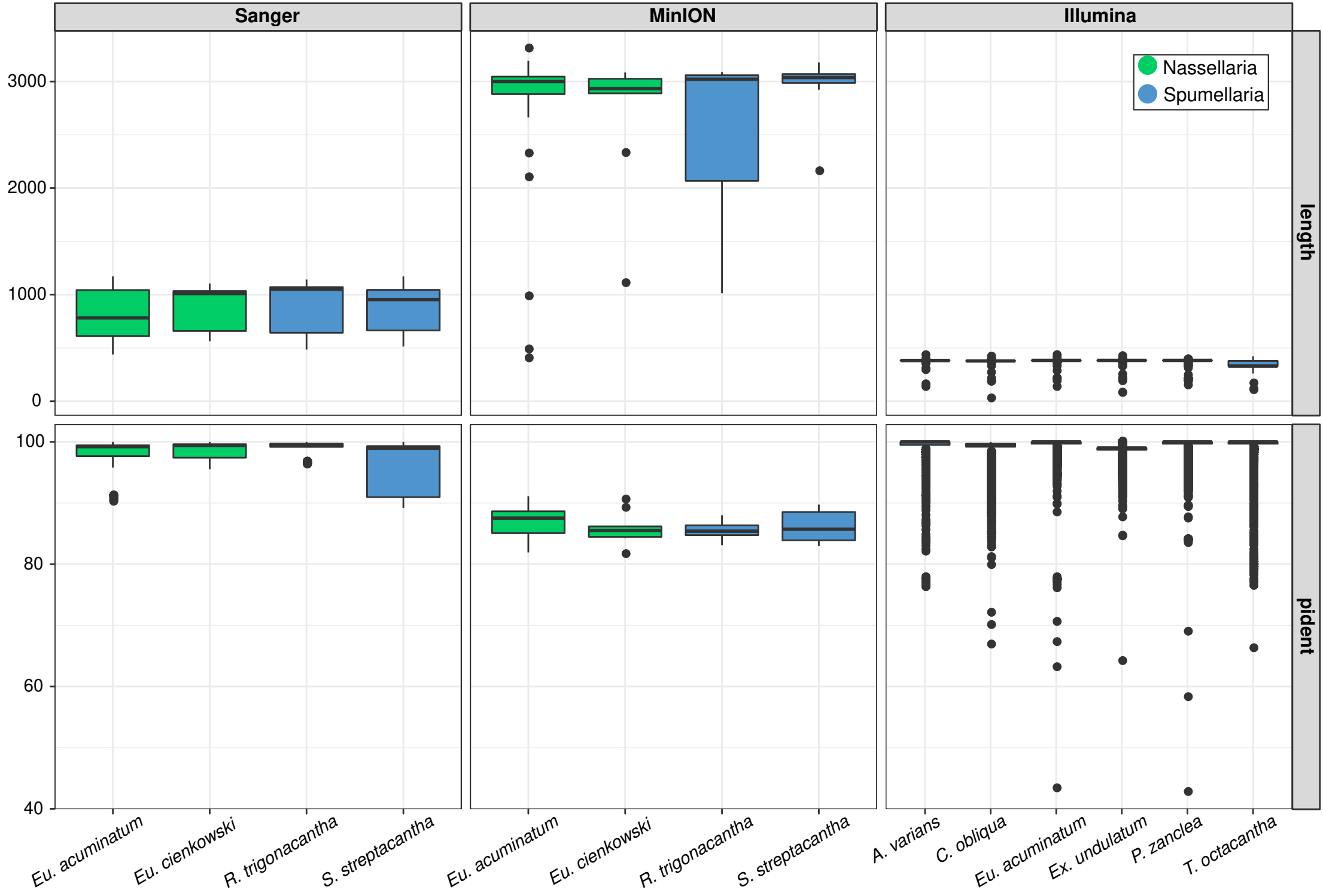
821 **Figure S1.** Schematic representation of the study design.

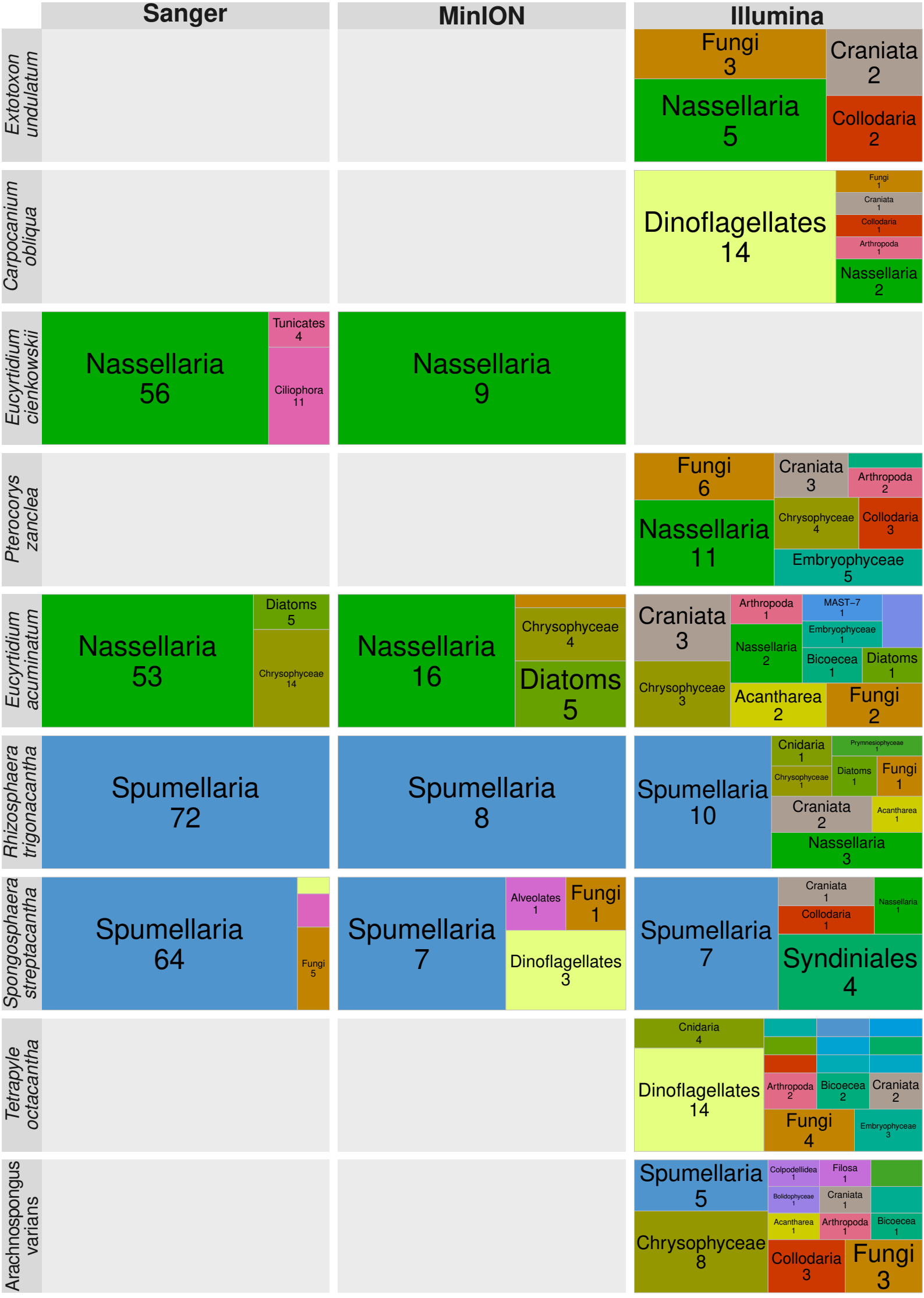
822 **Figure S2**. Light microscopy images of specimens used in this study. On the right of
823 each specimen it is indicated whether it was used (tick mark) or not (cross mark) for
824 Sanger + Oxford Nanopore Technologies (MinION) sequencing and/or Illumina, and in
825 grey, a brief description on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

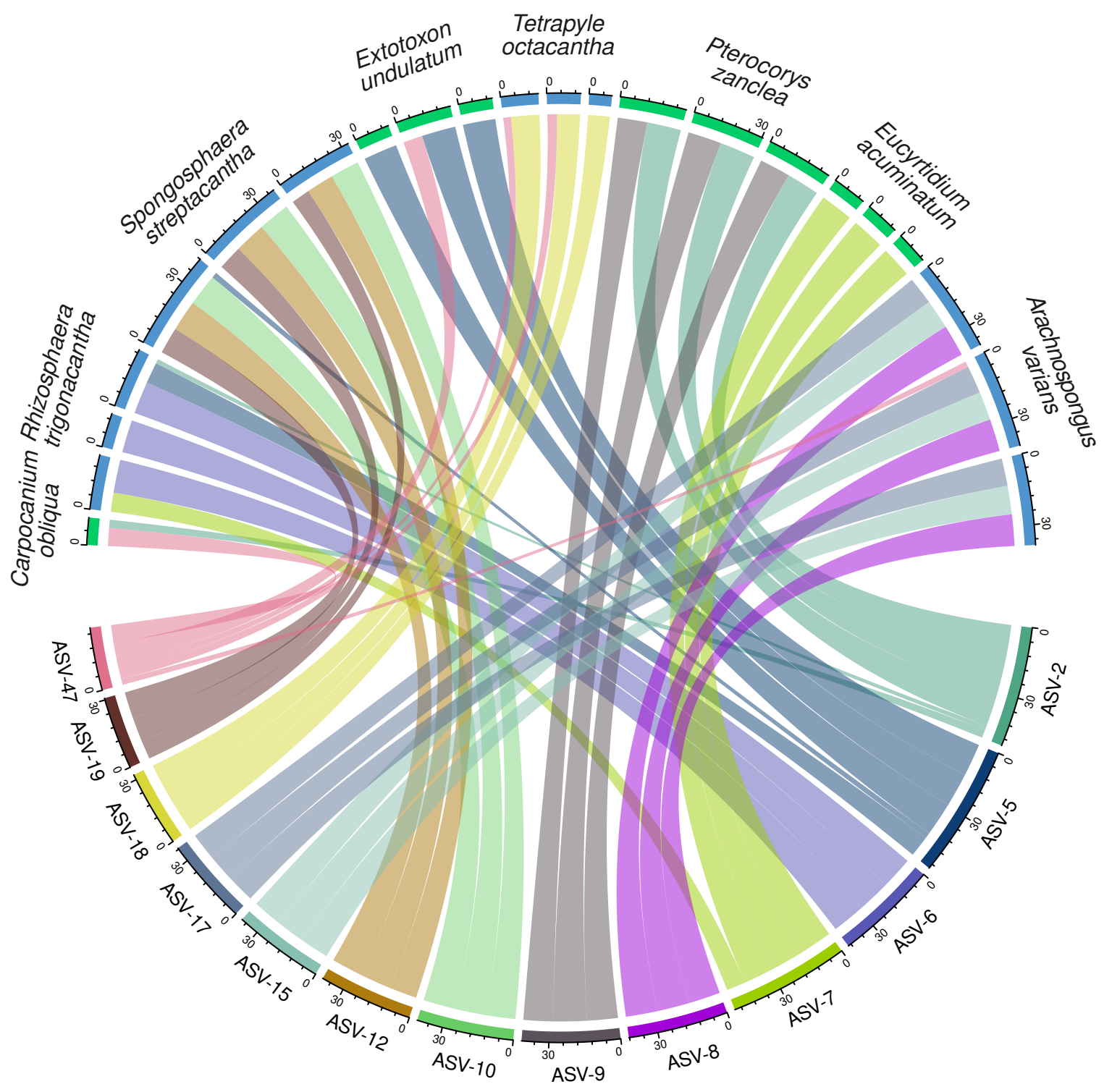
827 **Figure S3**. Tree map of the total number of reads (abundance) of every ASV affiliated to
828 a taxonomic group for each cell obtained by Illumina sequencing. ASVs were processed
829 by dada2.

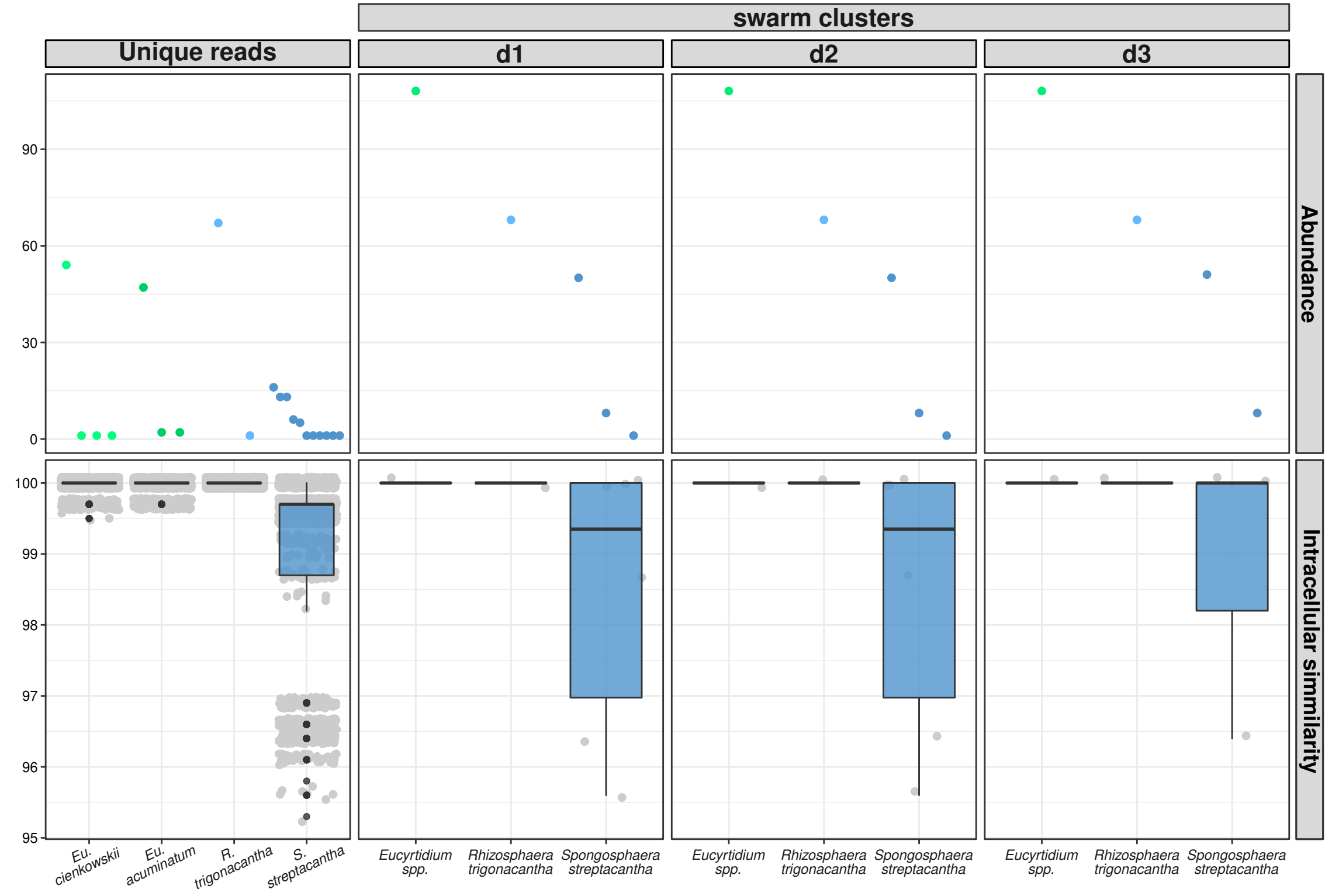
830 **Figure S4**. Tree map of every ASV affiliated to a taxonomic group (left column) and
831 their total number of reads (abundance; right column) for each cell obtained by
832 Illumina sequencing. ASVs were processed by dada2. Only ASVs present in 3 or more
833 samples and with a total abundance equal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

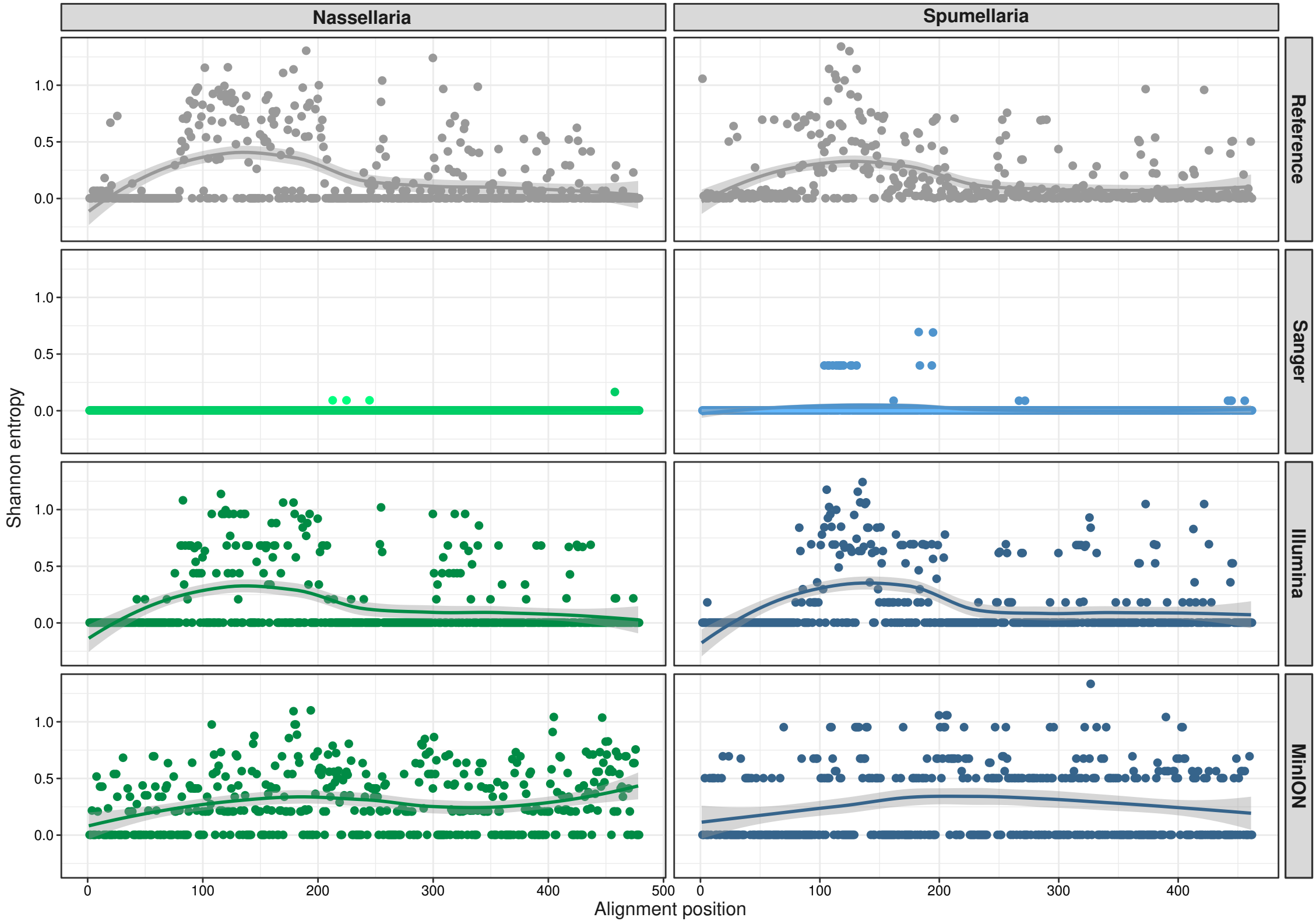
835 **Figure S5**. Shannon entropy analysis for every position (on *x* axis) of Sanger and Oxford
836 Nanopore Technologies (MinION) sequencing results aligned independently for
837 Nassellaria (green) and Spumellaria (blue). Lines represent the tendency of the entropy
838 for each alignment. Black arrows in Sanger boxes represent direction and approximate
839 position of the primers used for Sanger sequencing. Insertions (“-“) were not
840 considered due to the differences in length of the sequences and the high amount of
841 insertions produced by the raw sequences from MinION.

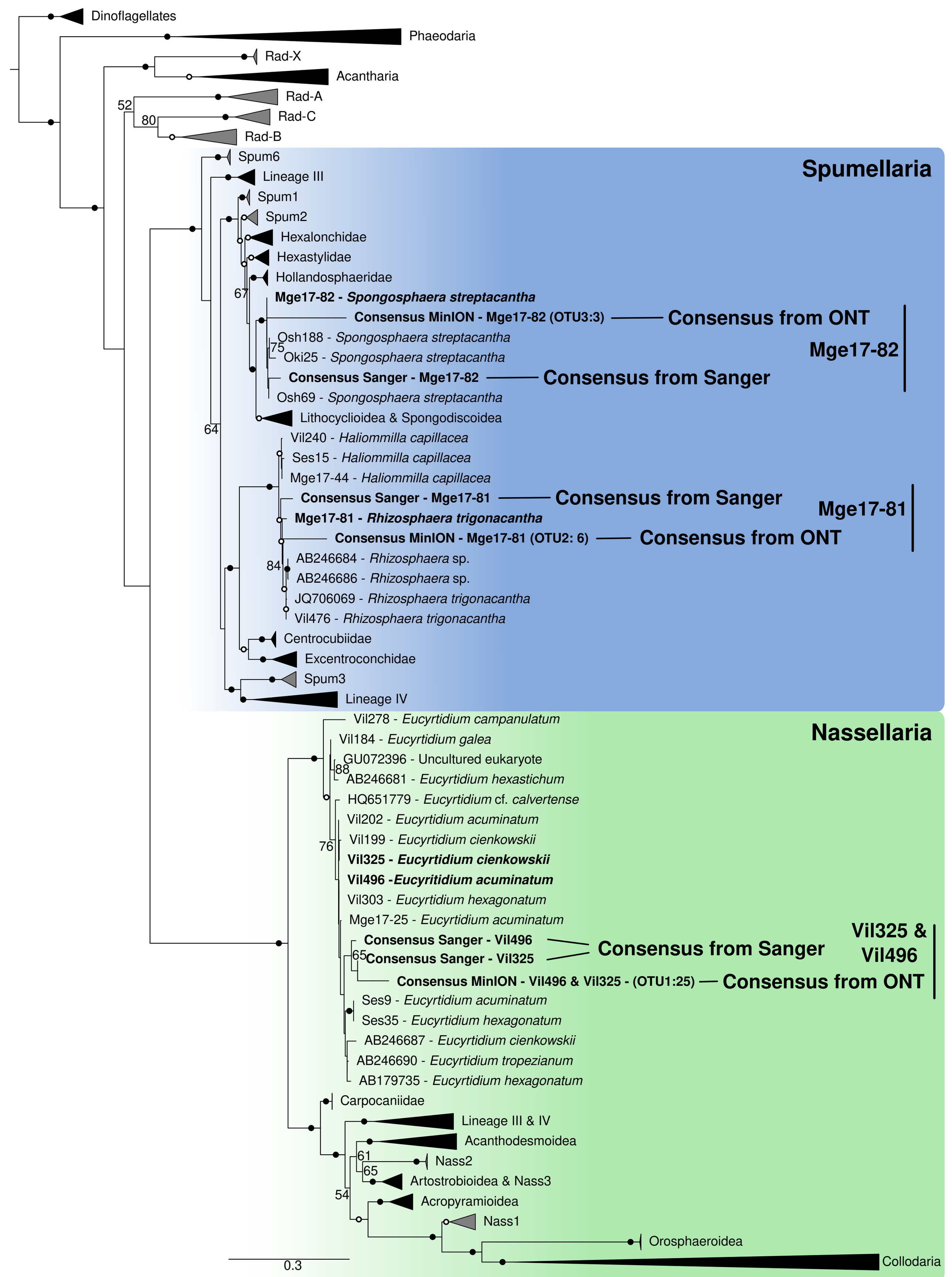

Sanger
MinION
Illumina
Nassellaria
Spumellaria
length
pident
3000
2000
1000
0
100
80
60
40
Eu. acuminatum
Eu. cienkowski
R. trigonacantha
S. streptacantha
Eu. acuminatum
Eu. cienkowski
R. trigonacantha
S. streptacantha
A. varians
C. obliqua
Eu. acuminatum
Ex. undulatum
P. zanclea
T. octacantha


Sanger
MinION
Illumina
Extotoxon undulatum
Fungi 3
Craniata 2
Nassellaria 5
Collodaria 2
Carpocanium obliqua
Dinoflagellates 14
Fungi 1
Craniata 1
Collodaria 1
Arthropoda 1
Nassellaria 2
Eucyrtidium cienkowskii
Nassellaria 56
Tunicates 4
Ciliophora 11
Nassellaria 9
Pterocorys zanclea
Fungi 6
Craniata 3
Arthropoda 2
Nassellaria 11
Chrysophyceae 4
Collodaria 3
Embryophyceae 5
Eucyrtidium acuminatum
Nassellaria 53
Diatoms 5
Chrysophyceae 14
Nassellaria 16
Chrysophyceae 4
Diatoms 5
Craniata 3
Arthropoda 1
MAST−7 1
Nassellaria 2
Embryophyceae 1
Bicoecea 1
Diatoms 1
Chrysophyceae 3
Acantharea 2
Fungi 2
Rhizosphaera trigonacantha
Spumellaria 72
Spumellaria 8
Spumellaria 10
Cnidaria 1
Prymnesiophyceae 1
Chrysophyceae 1
Diatoms 1
Fungi 1
Craniata 2
Acantharea 1
Nassellaria 3
Spongosphaera streptacantha
Spumellaria 64
Fungi 5
Spumellaria 7
Alveolates 1
Fungi 1
Dinoflagellates 3
Spumellaria 7
Craniata 1
Nassellaria 1
Collodaria 1
Syndiniales 4
Tetrapyle octacantha
Cnidaria 4
Dinoflagellates 14
Arthropoda 2
Bicoecea 2
Craniata 2
Fungi 4
Embryophyceae 3
Arachnospongus varians
Spumellaria 5
Colpodellidea 1
Filosa 1
Bolidophyceae 1
Craniata 1
Acantharea 1
Arthropoda 1
Bicoecea 1
Chrysophyceae 8
Collodaria 3
Fungi 3


Extotoxon undulatum
Tetrapyle octacantha
Pterocorys zanclea
Eucyrtidium acuminatum
Arachnospongus varians
Spongosphaera streptacantha
Rhizosphaera trigonacantha
Carpocanium obliqua
ASV-47
ASV-19
ASV-18
ASV-17
ASV-15
ASV-12
ASV-10
ASV-9
ASV-8
ASV-7
ASV-6
ASV-5
ASV-2


swarm clusters
Unique reads
d1
d2
d3
Abundance
Intracellular simmilarity
90
60
30
0
100
99
98
97
96
95
Eu. cienkowskii
Eu. acuminatum
R. trigonacantha
S. streptacantha
Eucyrtidium spp.
Rhizosphaera trigonacantha
Spongosphaera streptacantha


Nassellaria
Spumellaria
Reference
Sanger
Illumina
MinION
Shannon entropy
1.0
0.5
0.0
0
100
200
300
400
500
Alignment position


Dinoflagellates
Phaeodaria
Rad-X
Acantharia
Rad-A
Rad-C
Rad-B
52
80
Spumellaria
Spum6
Lineage III
Spum1
Spum2
Hexalonchidae
Hexastylidae
Hollandosphaeridae
67
Mge17-82 - Spongosphaera streptacantha
Consensus MinION - Mge17-82 (OTU3:3)
Consensus from ONT
Osh188 - Spongosphaera streptacantha
75
Oki25 - Spongosphaera streptacantha
Mge17-82
Consensus Sanger - Mge17-82
Consensus from Sanger
Osh69 - Spongosphaera streptacantha
Lithocyclioidea & Spongodiscoidea
64
Vil240 - Haliommilla capillacea
Ses15 - Haliommilla capillacea
Mge17-44 - Haliommilla capillacea
Consensus Sanger - Mge17-81
Consensus from Sanger
Mge17-81
Mge17-81 - Rhizosphaera trigonacantha
Consensus MinION - Mge17-81 (OTU2: 6)
Consensus from ONT
84
AB246684 - Rhizosphaera sp.
AB246686 - Rhizosphaera sp.
JQ706069 - Rhizosphaera trigonacantha
Vil476 - Rhizosphaera trigonacantha
Centrocubiidae
Excentroconchidae
Spum3
Lineage IV
Nassellaria
Vil278 - Eucyrtidium campanulatum
Vil184 - Eucyrtidium galea
GU072396 - Uncultured eukaryote
88
AB246681 - Eucyrtidium hexastichum
HQ651779 - Eucyrtidium cf. calvertense
Vil202 - Eucyrtidium acuminatum
76
Vil199 - Eucyrtidium cienkowskii
Vil325 - Eucyrtidium cienkowskii
Vil496 -Eucyritidium acuminatum
Vil303 - Eucyrtidium hexagonatum
Mge17-25 - Eucyrtidium acuminatum
Consensus Sanger - Vil496
Consensus Sanger - Vil325
Consensus from Sanger
Vil325 & Vil496
65
Consensus MinION - Vil496 & Vil325 - (OTU1:25)
Consensus from ONT
Ses9 - Eucyrtidium acuminatum
Ses35 - Eucyrtidium hexagonatum
AB246687 - Eucyrtidium cienkowskii
AB246690 - Eucyrtidium tropezianum
AB179735 - Eucyrtidium hexagonatum
Carpocaniidae
Lineage III & IV
Acanthodesmoidea
61
Nass2
65
Artostrobioidea & Nass3
54
Acropyramioidea
Nass1
Orosphaeroidea
Collodaria
0.3