***Supplementary Material***

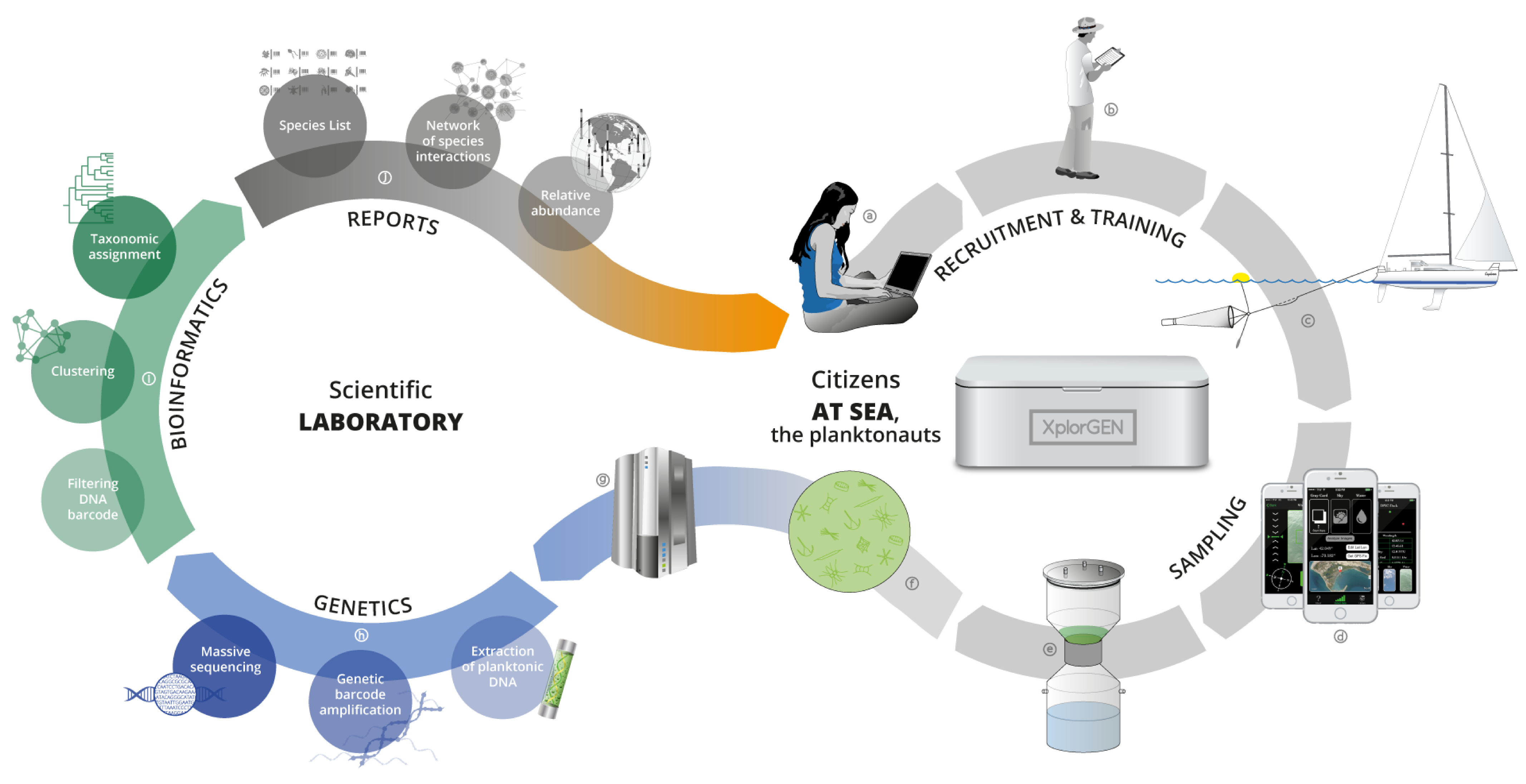
For

*Plankton Planet*: a frugal, cooperative measure of

aquatic life at the planetary scale

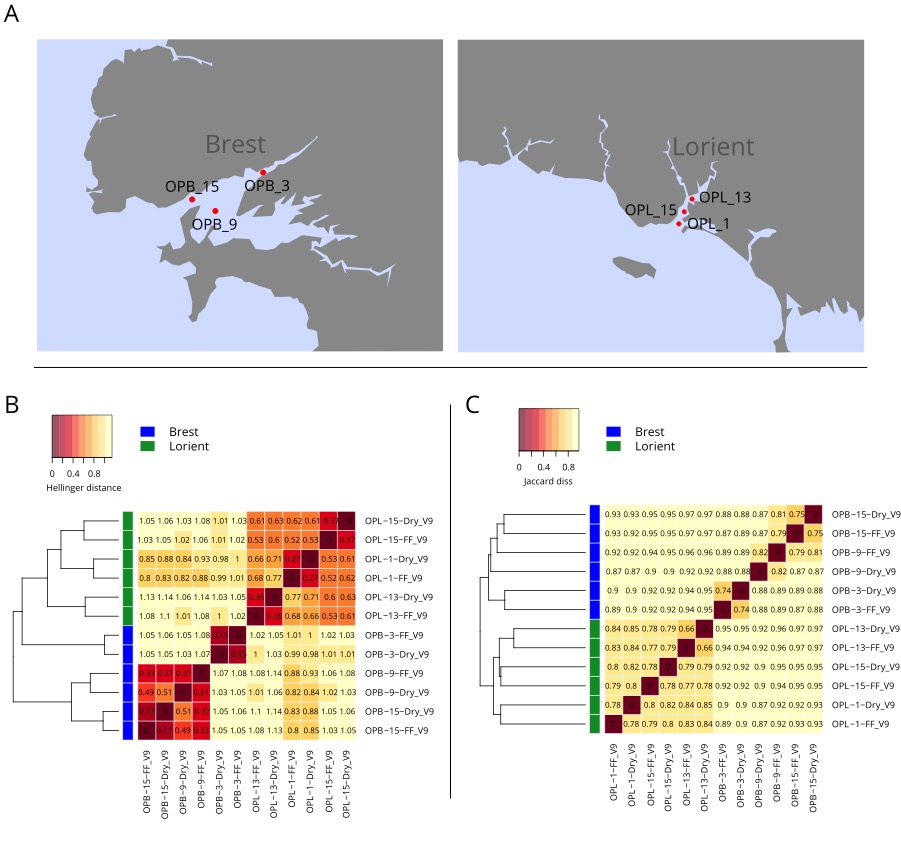
Colomban de Vargas1,2,3\*, Noan Le Bescot1,2, Thibaut Pollina2,4, Nicolas Henry1,2,3, Sarah Romac1,2,3, Sébastien Colin2,5, Nils Haëntjens6,2, Margaux Carmichael2, Calixte Berger2, David Le Guen2, Johan Decelle7, Frédéric Mahé8, Julie Poulain9,3, Emmanuel Malpot10, Carole Beaumont11, Michel Hardy12, the *Plankton Planet* core team§, Damien Guiffant2, Ian Probert1, David F. Gruber13, Andy Allen14, Gabriel Gorsky15,2, Mick Follows16, Xavier Pochon17,18, Romain Troublé19,2, Barry B. Cael20,2, Fabien Lombard15,2,21, Emmanuel Boss6,2\*, Manu Prakash4,2\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*



**Supplementary Figure 1. *Original Plankton Planet strategy***, showing the successive steps of both the seatizen (right loop, a to f) and researchers (left loop, g to j) operations, respectively at sea and in the lab.

The protocol for plankton DNA preservation by heating and desiccation was first tested during two field campaigns along the French Atlantic coast and compared with gold-standard cryo-fixation and preservation (Sup. Fig. 2).

****

**Supplementary Figure 2: *Testing the effect of dried (P2) versus flash-frozen plankton samples on DNA-based community composition***. A. Geographic location of plankton samples collected by volunteer citizen sailors along the French coast during the ‘Objectif Plankton’ initiative (<https://riem-asso.com/objectif-plancton/>). Bay of Brest, June 21st 2016, on the left, and Bay of Lorient, June 14th 2016, on the right; OPB and OPL indicate samples from Brest and Lorient bays, respectively. Each sample was preserved by both desiccation (Dry) and Flash Freezing into Liquid Nitrogen (FF), before deep Illumina sequencing of V9 SSU rDNA amplicons. B. Heatmap of the Hellinger distance (Hellinger standardization followed by Euclidian distance) between samples and sub-samples. C. Same as B, using Jaccard distances (based on OTUs presence/absence). The relatively high values of the Jaccard distances between preservation methods from the same samples comes from the fact that low abundance OTUs create a noisy background that inflates dissimilarities. The Hellinger distances are less sensitive to this issue, and the dissimilarities values range from 0.15 to 0.37 between preservation methods, while they are close to 1 between samples. The residual dissimilarities observed between subsamples are due to Illumina sequencing errors and are expected between technical replicates.

***Sampling coverage and putative cost***

****In order to compare ocean sampling cost between P2 sailing boats and classical oceanographic vessel, we measured the shortest possible route between P2 sampling sites (traveling salesman problem) by a putative oceanographic vessel. For convenience, 144 P2 sampling sites (56% of total sites) from the Pacific (83), Indian (1) and Atlantic (60) Oceans with a depth >0m according to the ETOPO1 NOAA database (Amante and Eakins, 2009) were selected. We then considered two ports as starting points for two navigation loops: Brest for the Atlantic Ocean samples plus 2 Pacific Ocean samples located at the extreme South of South America (62 samples in total); and San Diego for the other Pacific and Indian oceans samples (82 samples in total). For each group of stations, the shortest possible route that visits each station and return to the starting port was defined. To avoid crossing lands, the least cost distances between locations were computed by constraining paths to depths below 0 m. The distance matrix was computed using the lc.dist() function from the marmap R library (Pante and Simon-Bouhet, 2013). From this matrix, the traveling salesman problem was solved for each basin with the arbitrary insertion algorithm followed by a two edge exchange improvement procedure using the solve\_TSP() function with default options (TSP R library (Hahsler and Hornik, 2007)).

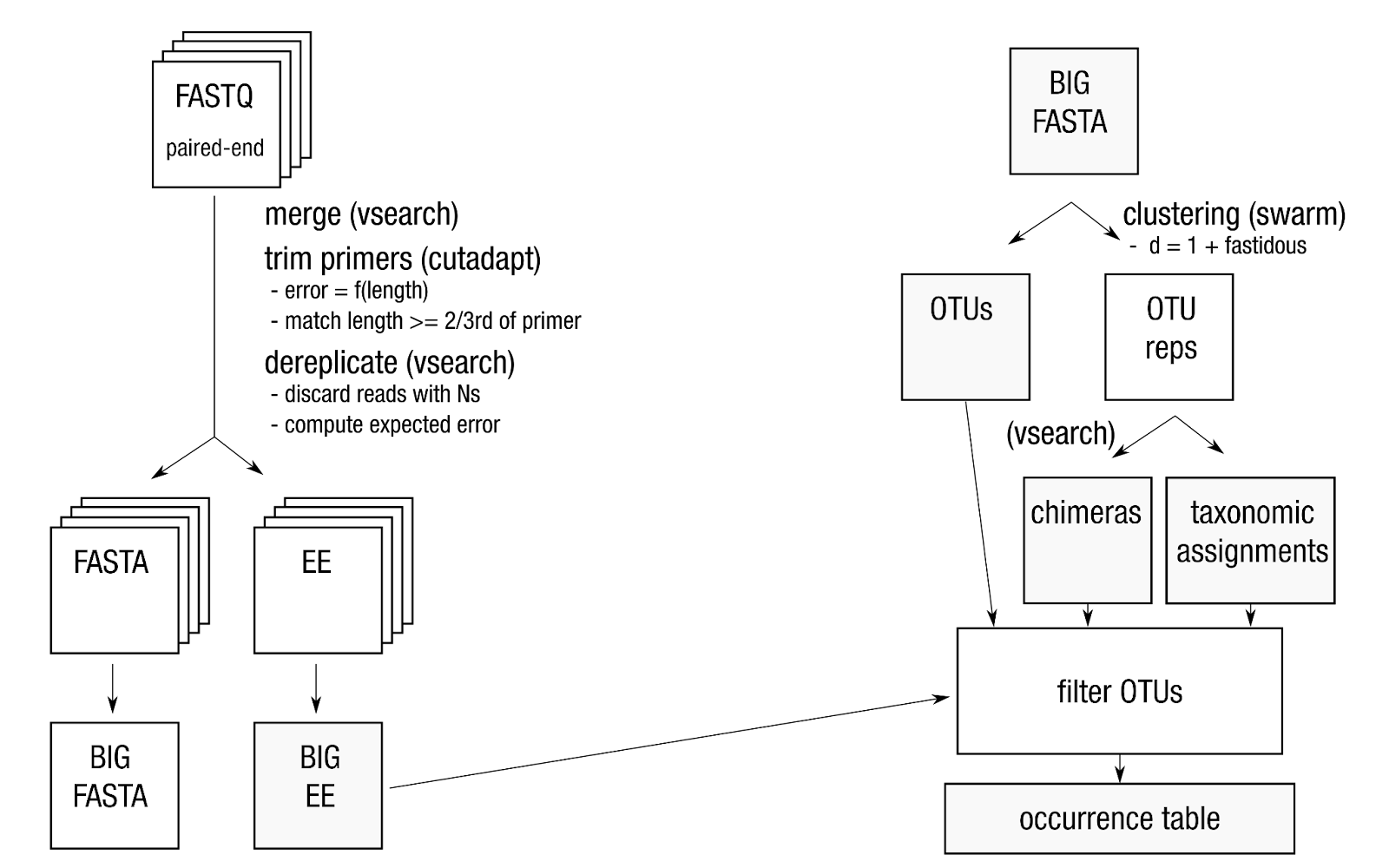
**Supplementary Figure 3*:* *Geographical coverage and putative cost of P2-pilot samples*.** Location of 233 (out of 258) ocean sites sampled by the pioneer planktonauts in 14 months during the Plankton Planet pilot project. The blue dots are the sampling sites taken into account to compute the shortest possible route by a putative single oceanographic vessel in the Indo-Pacific and Atlantic basins, respectively (‘traveling salesman problem’, see Sup. Text 1). The green dots are sampling sites considered as land sample (depth >0 m) because of their proximity to islands and the lack of resolution of the bathymetric database used to compute the shortest path without crossing the land between samples (see Sup. Text 1) . The red dots in the Mediterranean sea and Arctic ocean, correspond to sampling sites that were not integrated into the ‘traveling salesman problem’. The Atlantic loop, starting and ending in Brest, pass through 62 sites and is 44,972 km long; the Pacific loop, starting and ending in San Diego, pass through 82 sites and is 68,985 km long. Using a conventional oceanographic vessel (30,000 $US a day) travelling at 10 knots and stopping at each station for 1 hour, sampling these two loops would cost circa 7,9 million $US in total.

****

**Supplementary Figure 4**: Box plot representation of the percentage of prokaryotic (Archaea and Bacteria) rDNA reads P2 samples for each boat, and for *Tara* Oceans samples (micro- and meso-planktonic size fractions). 85% and 89% of P2 and *T*O samples, respectively, contain less than 1% of prokaryotic reads, the difference being explained by a few outliers. Note the higher value for a few samples from *Amasia* and *Race for Water*, which correspond to samples that were kept on board at room temperature for several days before processing.

****

**Supplementary Figure 5**: ***Details about the three plankton community clusters generated for the South-West Pacific loop samples (Fig. 7B)***. **A**: Geographic distribution of indicator OTUs among taxonomic groups for each cluster. For each OTU, the indicator values, i.e. the product of the relative frequency and relative average abundance in clusters, are computed from Hellinger transformed abundances for each of the three clusters using the indval() function (labdsv R package, 49). OTUs are considered as indicator of a specific cluster when p-value are below 0.01. **B**: Consistency of the clusters using the silhouette width (measure of how similar a sample is to its own cluster compared to other clusters). The silhouette width has been computed using the silhouette() function (cluster R package43).



**Supplementary Figure 6. Schematic workflow of the bioinformatic pipeline used to generate the 18S V9 rDNA amplicons occurrence table.**

Illumina HiSeq sequencing, base calling and demultiplexing were carried out using RTA v1.18.54, MCS 2.6 and bcl2fastq2.17, yielding a total of 305,942,637 paired-reads (2x150 bp) (min: 590,256 reads, max: 2,162,072 reads, mean: 1,662,732  reads) for the 184 samples sequenced at GenoToul, and  235,276,059 paired-reads (2x150 bp) (min: 3,119,180 reads, max: 13,536,756 reads, mean: 7,352,377 reads) for the 32 samples  sequenced at Genoscope (see above). Paired Illumina MiSeq reads were assembled with vsearch v2.8.1 (Rognes et al., 2016) using the command fastq\_mergepairs and the option fastq\_allowmergestagger. Primer clipping were performed with cutadapt v1.16 (Martin 2011) forcing a full-length match for sample tags and allowing a 2/3-length partial match for forward and reverse primers. Only reads containing both primers were retained. For each trimmed read, the expected error was estimated with vsearch’s command fastq\_filter and the option eeout. Each sample was then dereplicated, i.e. strictly identical reads were merged, using vsearch’s command derep\_fulllength, and converted to FASTA format. To prepare for clustering, the samples were poold and processed by another round of dereplication with vsearch. Files containing expected error estimates were also dereplicated to retain only the lowest expected error for each unique sequence. Clustering was performed with swarm v2.2.2 (Mahe et al., 2014), using a local threshold of one difference and the fastidious option. Operational taxonomic unit (OTU) representative sequences were then searched for chimeras with vsearch’s command uchime\_denovo (Edgar et al., 2011). In parallel, representative sequences were assigned using the stampa pipeline (https://github.com/frederic-mahe/stampa/) against a custom version (<https://doi.org/10.5281/zenodo.3768951>) of the Protist Ribosomal Reference database PR2 (Guillou *et al.*, 2013). Up to that point, reads that could not be merged, reads without tags or primers, reads shorter than 32 nucleotides and reads with uncalled bases (‘N’) had been eliminated.  To create a filtered occurrence table, clustering results, expected error values, taxonomic assignments and chimera detection results were pooled to keep only non-chimeric OTUs, OTUs with an expected error per nucleotide below 0.0002, and OTUs containing more than three reads or seen in two samples (in-house python script).