Single cell ecogenomics reveals mating types of individual cells and ssDNA viral infections in the smallest photosynthetic eukaryotes

L. Felipe Felipe Benites, Nicole Poulton, Karine Labadie, Michael E. Sieracki, Nigel Grimsley, Gwenael Piganeau

Single Cell Ecogenomics reveals mating types of individual cells and ssDNA viral infections in the Smallest Photosynthetic Eukaryotes

L. Felipe Benites[1], Nicole Poulton[2], Karine Labadie[3], Mike Sieracki[2], Nigel Grimsley[1], Gwenael Piganeau[1*]

[1]*Integrative Biology of Marine Organisms (BIOM), Sorbonne University, CNRS, Oceanological Observatory of Banyuls, Banyuls-sur-Mer, France.*
[2] *Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA.*
[3] *Genoscope, Institut de Biologie François-Jacob, Commissariat à l'Energie Atomique, université Paris Saclay, Evry, France.*

Keywords: Picoeukaryotes, single amplified genome, mating type, cross-contamination, virus, Mamiellophyceae, Tara-Oceans

Summary

Planktonic photosynthetic organisms of the class Mamiellophyceae include the smallest eukaryotes (<2 µm), are globally distributed and form the basis of coastal marine ecosystems. Eight complete fully annotated 13 to 22 Mb genomes from three genera, *Ostreococcus*, *Bathycoccus* and *Micromonas*, are available from previously isolated clonal cultured strains and provide an ideal resource to explore the scope and challenges of analysing Single Cell Amplified Genomes (SAGs) isolated from a natural environment. We assembled data from 12 SAGs sampled during the Tara Oceans expedition to gain biological insights about their *in-situ* ecology, which might be lost by isolation and strain culture. Although the assembled nuclear genomes were incomplete, they were large enough to infer the mating types of four *Ostreococcus* SAGs. The systematic occurrence of sequences from the mitochondria and chloroplast, representing less than 3% of the total cell's DNA, intimates that SAGs provide suitable substrates for detection of non-target sequences, such as those of virions. Analysis of the non-Mamiellophyceae assemblies, following filtering out cross-contaminations during the sequencing process, revealed two novel 1.6 and 1.8 kb circular DNA viruses, and the presence of specific Bacterial and Oomycete sequences suggests that these organisms might co-occur with the Mamiellales.

1. Introduction

Planktonic photosynthetic eukaryotes of the class Mamiellophyceae (Chlorophyta) include the smallest eukaryotes, with bacterial-sized cells of less than 2 µm diameter. Environmental surveys based on the sequence of the highly conserved 18S ribosomal gene (18S) eukaryotic barcode (1) revealed their worldwide distribution (2)(3)(4)(5). Their ubiquity, high abundance and turnover suggest they sustain the marine ecosystems in many coastal areas (6)(7). Many flagellate species in this class were described last century before the advent of molecular biology techniques and their first application to these picoalgae (8); *Monomastix* 1912 (9), *Micromonas pusilla* 1951 (10), and *Mamiella gilva* 1964 (11), followed by *Crustomastix* and *Dolichomastix* 2005 (12). A non-flagellate member was described 1990 as *Bathycoccus prasinos* (13) and the non-scaled coccoid *Ostreococcus tauri* was described in 1995 (14). A thorough phylogenetic analysis of the nuclear and plastid encoded rDNA operon led Marin and Melkonian to define the class Mamiellophyceae, divided into three orders Mamiellales, Dolichomastigales and Monomastigales (15). While twenty-two species are presently described in AlgaeBase (16), recent environmental sequencing suggests that many more species have not been yet isolated (5)(17) so that the number of species within this class remains an open question. The relative ease of culture of strains of the picoalgae *Bathycoccus*, *Micromonas*, and *Ostreococcus* in Keller's and related media (18) fostered their development as new models for cell biology (19)(20) and environmental studies (21)(22). Complete and annotated nuclear genomes have been obtained for eight species to date: *O. tauri* RCC4221 (23), *O. lucimarinus* (24), *O.* sp. RCC809, *O. mediterraneus* (25), *Micromonas pusilla* and *M. commoda* (26), *B. prasinos* RCC1105 (27), and *O. tauri* RCC1115 (28).

Despite the progress enabled by lab experiments and sequence analyses, the sexual life cycles of these haploid picoalgae, as well as their interactions with bacteria (29) and viruses (30) *in situ* remain enigmatic. Single amplified genomes (SAGs) assemblies, produced by multiple displacement amplification (MDA) after extraction from sorted single cells have been previously reported to contain foreign DNA (31)(32). They may be viewed conceptually as metagenomes, provided they preserve a molecular signature of the ecological context of the cell. SAGs of cells directly sampled from the environment may not only foster knowledge on diversity in microorganisms (33)(34), but also open opportunities for the identification of all kinds of novel biological associations, from predation to parasitism (31)(35).

Here, we investigated 12 Mamiellales' SAGs sampled in the Indian Ocean during the Tara Oceans expedition (36) to explore the biological insights this data provides within an ecogenomic framework. Therefore, we analysed the taxonomic affiliations of the assembled sequence data to discuss the power and challenges of using SAGs to retrieve *in-situ* ecological information (1) from the target sequence data, e.g. Mamiellales sequences, about the mating type of the cells and (2) from the non-target sequences identified.

2. Material and Methods

Data

Single-cells were collected during the Tara Oceans expedition (37) from surface waters in the Indian Ocean at station 39, 41 and 46 (38) and cryopreserved using glycine betaine. Flow cytometric cell sorting, single cell lysis and whole genome amplification by Multiple Displacement Amplification (MDA) (39) were performed at the Bigelow Single Cell Genomics facility (Boothbay, Maine USA), as previously described (40)(41)(42–44)(41–43)(40–42). The resulting SAGs were screened using universal eukaryotic 18S rDNA PCR primers (45). The 12 SAGs we analysed here were selected for sequencing based on the 18S identities and were sequenced using Illumina HiSeq technology at the Genoscope (33). All SAGs were multiplexed, SAG1 to 6 and SAG7 to 12 were sequenced on two different runs. The raw Illumina data was processed as previously described (36): adapters and primers used during library construction were removed from the whole reads and both ends were trimmed for low quality nucleotides (quality value < 20). The longest sequence without adapters and with fewest low quality bases was kept. Sequences between the second unknown nucleotide (N) and the end of the read were also trimmed. Reads shorter than 30 nucleotides after trimming were discarded. These trimming steps were achieved using fastx_clean, an internal software based on the FASTX library. The reads and their mates that mapped onto run quality control sequences (Enterobacteria phage PhiX174 genome, NC_001422.1) were removed using SOAP aligner. The total number of paired-end reads from the processed raw reads ranged between 19 and 36 millions with an average of 26 million reads per SAG.

SAG assembly and taxonomic affiliation of contigs

Individual pair-end read files were assembled using SPAdes v3.9.0 (46) and the final assembly statistics were generated with Quast v4.6.3 (47). Contigs smaller than 400 bp were discarded from downstream analysis. Genome completeness was assessed with BUSCO v3.0.2 using the 303 eukaryotic single copy orthologs dataset (48) and compared to the values obtained on the available closely related sequenced genomes (Table 1).

The SAG assemblies were screened for contigs containing the 18S using a *blastn* homology search using a custom database of 18S extracted from GenBank complemented with 18S genes of Mamiellales extracted from the genome sequence (supplemental table 1). Contigs with a 18S hit were analysed with *RNAmmer* v1.2 (http://www.cbs.dtu.dk/services/RNAmmer/) and the predicted 18S sequences were aligned with Mamiellales 18SrDNA sequences using *MAFFTv7.305b* (49). Alignments were trimmed with trimAI v1.2 (-automated1 method) (50) and maximum likelihood phylogenetic trees were built under GTR model using *RAxML* v. 8.2.9 (51) and plotted with *FigTree* v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Taxonomic affiliations to separate non-target contigs (foreign DNA) from target (Mamiellophyceae) were performed as follows. Each contig was aligned against GenBank using *DIAMOND* (*blastx* option e-value 1e-5 maximum aligned sequences 100) and the 10 best hits for each contig were kept. The taxonomic affiliation of each contig was retrieved using a custom perl script that fetched the GenBank accession number from the *DIAMOND* output, and retrieved the taxonomical nodes from the GenBank taxonomy database for each hit (script available upon request). Manual inspection of taxonomic affiliation of contigs containing several open reading frames led to the correction of the taxonomic affiliation for a few contigs affiliated to Archaea on the base of the best blast hit, while most open reading frames along the contig had a best hit against Bacterial sequences. As the complete genome and organellar sequences of *Ostreococcus*, *Bathycoccus*, and *Micromonas* species are present in GenBank, contigs were considered as target if at least one of the 10 best hits belonged to a Mamiellophyceae. We considered following high order taxonomic ranks: Archaea, Bacteria, Fungi, Mamiellophyceae, Metazoa, Oomycetes, Other (when several taxa were mixed in uneven proportions within the 10 best hits), Other protists, Plants (Streptophyta), Unassigned contigs and Virus. To estimate the number of reads recruited to each taxonomic rank in each SAG assembly, we aligned the reads against each assembly with *BWA* (52) and extracted the read coverage of each contig estimated from the SAMtools *mpileup* and *depth* suite (53).

The sum of contig lengths affiliated to a given group was plotted with *ggplots* in the R environment (54).

Cross-contamination filtering

To remove contigs derived from cross-contamination during the sequencing step, all SAG assemblies from the same Illumina run were compared against all with *blastn* (e-value cut-off 1e-5). This led to the inclusion of one SAG from a MAST lineage (TOSGAG23-1, GenBank accession ERX1271190) that was multiplexed with SAGs 7-12. Each contig was flagged as candidate contaminant if its alignment length with a contig from another SAG was equal or higher than 95% of the length of its best hit, with a nucleotide identity higher than 99.5%. The number of reads recruited to each candidate contig was used to define the source of the contamination: the SAG containing the contig with the highest read coverage was defined as the source SAG, and the corresponding contigs from the other SAGs were considered as contaminants and discarded.

Annotation of non-target contigs

Because the virus contigs detected in SAGs 8 and 12 had best hits with circular viruses, the contigs were first circularized with the *Geneious* program v10.0.5 (55), and open reading frames (ORF) were predicted with *Glimmer* (56).

The conserved motifs of the viral replicase gene (57)(58) were manually searched by inspection of the predicted amino acids sequence located in the N-terminal endonuclease domain and in the C-terminal superfamily 3 helicase domain using MAFFT v7.305b (--localpair –maxiterate 1000). The best *blastp* hits to these ORFs were complemented with sequences retrieved from GenBank and the Ocean Atlas genes database (59) (supplemental table 2). Multiple sequence alignments were performed with *MAFFT* v7.305b (--localpair --maxiterate 1000) (60), and trimmed with *trimAI* v1.2 (-automated1 method) (50). For each alignment, protein evolution models were selected using *ProtTest* v3.4.2 (61) with the Akaike Information Criterion (AIC) (LG+I+G for ORF1 and RTREV+I+G+F for ORF2). Phylogenies were built by the maximum likelihood method implemented in *RAxML* v. 8.2.9 (51), and plotted in *FigTree* v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Screening assemblies for information about mating type

In *Ostreococcus* species, population genomics analyses have revealed two anciently diverged 500 kb genomic regions with supressed recombination, which are candidate mating type regions (28). As in many bipolar mating type systems, orthologous genes on this mating

type locus (MT) cluster phylogenetically by mating type, not by species (28). This is as expected if the origin of mating types predates speciation within this genus, as previously observed in the *Volvox* genus (62). We took advantage of this deep divergence between 23 orthologous genes families of the two mating types, gametologs (63), to screen the SAG assemblies for homologous sequences. Briefly, orthologous gene families were obtained following an 'all-against-all' protein sequence similarity search, performed with blastp (maximum e-value 1e-4) and gene families were delineated using OrthoFinder (version 2.1.2) (64). The dataset consisted of 23 genes in the eight sequenced Mamiellophyceae genomes (28) and contains gene families consisting exclusively of gametologs in the 8 sequenced Mamiellales genomes : *O. tauri* RCC4221 (23), *O. tauri* RCC1115 (28), *B. prasinos* RCC1105 (27), *M. commoda* RCC299, *M. pusillla* CCMP1545 (26), *O. lucimarinus* (65) and *O. spp* RCC809, *O. mediterraneus* RCC2590 (25). Protein coding sequences of these genes in the *Ostreococcus* SAG assemblies were extracted manually from the *tblastn* nucleotide alignments.

Biogeographic occurrence of Mamiellales SAG-associated sequences

The non-targeted bacterial sequences (supplemental table 3) were searched against 957 manually curated partial assemblies (metagenome-assembled genomes MAG) reconstructed from 93 TARA Oceans metagenomes (66), including those stations from the North Indian Ocean the Mamiellophyceae where the SAGs were isolated. A match against a MAG was defined as an alignment of 90% of the length of SAG-associated bacterial contig with a nucleotide identity higher than 98%. Nucleotide identity and total alignment length for each SAG associated bacterial contig were computed from blastn alignments (67).

The geographic occurrences and distribution of the non-target virus sequences was obtained by searching for homologous sequences to the two predicted protein-coding genes, a capsid and a replicase, in the OCEAN GENE ATLAS (59) using *blastp* (e-value 1e-5). Amino-acid sequences for each hit were downloaded from the database (supplemental table 2). This atlas contains a gene catalogue from 243 metagenomes and 441 metatranscriptomes from the Tara Oceans expedition, as well as the metagenomes from the Global Ocean Sequencing (GOS) expedition (68).

3. Results

Completeness of Target Assemblies

Phylogenetic analysis of 18S data retrieved from the *de novo* assembly was consistent with prior taxonomic affiliations. *Ostreococcus* SAG 18S sequences retrieved from the *de novo* assembly clustered with *O. spp* RCC809. However, while SAG9, 10, 11 sequences were 100% identical to RCC809 (and RCC143), the 18S sequence of SAG11 had one difference to RCC809. The 18S of the *Bathycoccus* SAGs were 100% identical to the 18S sequence from the *Bathycoccus prasinos* RCC1105 reference genome and the 18S found in *Micromonas* SAGs were 100% identical to the recently described *M. bravo* species (69) (Figure 1). We could not retrieve the 18S rDNA sequence from the assemblies in SAG4 and SAG6-7.

The SAG target assemblies recruit the overwhelming majority of reads, from a minimum of 82% of reads in SAG8 to 99.96% in SAG9 (Table 2). Target assemblies are fragmented and summed up to a maximum of 4.2 Mb (SAG2), corresponding to a maximum of one fourth of the expected genome size (SAG2, *Bathycoccus*, SAG10, *Ostreococcus*). This modest genome representation is also reflected by the BUSCO analysis, with a maximum of 28.7 % of represented genes (SAG9), to 1% in SAG6, where the target sequences are overwhelmingly represented by the chloroplast and mitochondrial sequences (Table 2). As a comparison, the same BUSCO analyses of the available *O. tauri, B. prasinos* and *M. commoda* genomes gave 86%, 84 and 89% respectively, suggesting that BUSCO analysis gives a conservative estimation of genome completeness in the Mamiellales. The relatively low target nuclear genome recoveries are in line with previous studies, which showed that the combination of sequences obtained from several SAGs may increase target genome recovery from 20 to 70% (70).

There was a large variation in the proportion of reads recruited to the mitochondrial (0.05 to 26%) and chloroplast (0.01 to 73%) genomes. Importantly, the assembly lengths were always large enough (> 20 kb, Table 1) to annotate several complete protein-encoding genes. This sequence data is sufficient for the unambiguous identification of contigs from both the mitochondrial and plastid genomes. However, because the percentage of recruited reads also depends on the copy number of the chloroplast and mitochondria genomes within a cell, we first estimated the nuclear to plastid or to mitochondria ratios in *Ostreococcus* from available data. In *O. tauri*, this relative copy number can be inferred from the ratio of the nuclear (28) versus organellar genome coverage (71) obtained from the same Illumina run of a haploid culture. Mitochondrial genome copy number was estimated to range between 2 and 5, chloroplast copy number ranged between 3 and 7, using the available Illumina datasets obtained from 13 independent *O. tauri* cultures (supplemental

table 4). The organellar genomes represent less than 0.5% of the nuclear genome sizes in Mamiellophyceae (Table 1), but may reach 3.5% with the maximum organellar genome copy number.

If the proportion of reads reflects the proportion of DNA in the sampled cell, the percent of reads recruited to the organellar genomes may thus be expected to range from 0.01 to 3.5%. This is the case in all *Bathycoccus* SAGs and two *Ostreococcus* SAGs (Table 2).

The complete mitochondrial (mtDNA) or chloroplastic (cpDNA) genomes could not be assembled from individual SAG data (Geneious inbuilt assembler with minimum contig overlap 100 bp and a minimum overlap Identity 90%). However, running the assembler on all contigs affiliated to cpDNA or mtDNA within each genera led to large (>30kb) assemblies of 3 mtDNA and 2 cpDNA. Manual assembly guided by the synteny with the available reference genomes (Table 1) led to 2 complete assemblies, and 3 partial assemblies, summarized in Table 3. To compare the mitochondrial genome assembly obtained for *Bathycoccus* SAGs with the available reference mitochondrial genome of *Bathycoccus prasin*os RCC1105 (27), the amino-acid identity was calculated along a concatenation of 6 orthologous genes outside the inverted repeat region (nad5, nad4, coc3, cox2, atp6 and nad1) as previously described (1). Intraspecific diversity within Mamiellophyceae organelles is yet unknown for most species, but in *O. tauri* intraspecific amino-acid divergence is below 0.1% (72). The amino-acid identity over the 1789 aa long alignment was 81%, supporting the previously published conclusion based on the analysis of the combined nuclear genome assembly from SAG1-4, that the *Bathycoccus* sequenced belong to a novel, yet uncultured, *Bathycoccu*s species (33).

Mating type inference in *Ostreococcus* SAGs

Contigs from the target fraction in the SAGs assemblies were searched for genes located within the candidate *MT-* and *MT+* gene families defined in the available reference genomes. Since the sequences of the two mating types have not yet been identified in *Bathycoccus* and *Micromonas*, the inference of mating types is not yet possible in these genera. The amino-acid sequence of orthologous genes could be retrieved for 6 out of the 23 gene families (Figure 2 A). This low gene number is not surprising given the low nuclear genome coverage of each SAG (Table 1), but gene family 16 (GF16 – encoding for a diacylglycerol glucosyltransferase) had matches with 3 out of the 4 *Ostreococcus* SAGs.

Comparison of the amino-acid sequence of these genes with genes from *Ostreococcus* sp. RCC809 indicate that the four *Ostreococcus* SAGs have the same mating type as RCC809, that is *MT+* (Figure 2B).

Taxonomic Diversity of Non-target contigs

The non-target assemblies recruited 0.01 to 6% of the reads (Table 1). Taxonomic affiliation of total non-target contigs is provided in Figure 3A and varies greatly between SAGs. For eight SAGs, most non-target reads were recruited to Bacteria, for 2 SAGs, most non-target reads were recruited to Oomycetes, and for 2 SAGs, most reads were recruited to sequences of viral origin. Sequence comparison between non-target assemblies revealed 100% identical sequences shared between SAGs coming from different stations, but sequenced on the same Illumina lane, a suspicious signature suggesting cross-contamination during the amplification and sequencing step (73). This led us to define a cross-contamination filter procedure for all SAGs sequenced within the same run. To find out which of the SAGs was the most likely source of the contamination, we used the number of reads recruited to non-target contigs. The SAG containing the most sequenced reads was considered as the source SAG of the non-target contig. The taxonomic affiliation of non-target contigs following filtering is presented in Figure 3B. Basically, this led to a 2 to 7-fold reduction of the total non-target assembly length depending on the SAG (Table 2). Following cross-contamination filtering, the maximum percent of reads recruited to non-target sequences dropped from 6% to 1% in SAG8, and became almost negligible (0.002%) in SAG10. The taxonomic affiliation of these filtered non-target sequences could be divided into three groups: Bacteria, Oomycetes/Fungi and Viruses, in order of decreasing frequency.

Most non-target reads were recruited to Bacteria for 11 SAGs and to virus for one SAG (SAG12) (Figure 3B). The non-target bacterial contigs had a large phylogenetic spread (Figure 4A) and while the most common phylum belonged to the Proteobacteria, there was no evidence of any specific bacteria - SAG association from this analysis. To gain more information about the origin of the 83 non-target bacterial sequences, we estimated their similarity with 957 manually curated partial assemblies (metagenome-assembled genomes MAGs) reconstructed from 93 TARA Oceans metagenomes (66). Over one third (32 out of 83) bacterial contigs had more than 98% identity over 90% of their length with a TARA MAG, suggesting indeed that these bacteria belong to the most abundant bacteria sequenced within the TARA Ocean sampling (Figure 4A). However, only 3 bacterial contigs

had a hit against a MAG from the North Indian Ocean, while most hits were against MAGs from the Mediterranean Sea and South East Pacific Ocean. Therefore, except for the 3 bacterial contigs assembled from SAG3 and SAG4, the bacteria sampled together with the SAG do not represent the most dominant bacteria at the sampling site.

The cross-contamination filtering removed most of Oomycetes affiliated non-target contigs (Figure 3B), and the source was one uncultured marine Stramenopile (MAST) SAG (70) sequenced in the same lane as SAG7 to SAG12. The class Oomycetes is also from the Stramenopiles group and form a distant, yet phylogenetically related clade to MASTs, which currently lack representation in reference databases (74). Following the cross-contamination filtration step, non-target Oomycete contigs were affiliated to four families (Figure 4B): Pythiaceae (2 contigs), Albuginaceae (3 contigs) and the more abundant Peronosporaceae (30 contigs) and Saprolegniaceae (33 contigs). Given the available genome size estimations in Oomycetes (75) or MASTs (70), which are larger than 20 Mbp, the size of the largest target Mamiellales, *Micromonas* (Table 1), the quantity of Oomycetes DNA present within the SAGs precludes any evidence of co-sampling, but rather suggests random capture of DNA during the cell sorting process.

The least frequent non-target sequence group was affiliated to viruses. Analysis of the contigs affiliated to viruses in SAG8 (*Micromonas*) and SAG12 (*Ostreococcus*) led us to assemble two novel circular viral genomes, which we named "Mamiellales SAG associated circular virus" (MACV). Genome sizes varies between 1586 bp (SAG12 – Ostreococcus Figure 5A) and 1788 bp (in SAG8 – *Micromonas*), both coding for two ORFs, a replicase gene (ORF1) and putative capsid gene (ORF2). The two genomes are 100% identical, except for a 202 bp deletion in the shorter version, leading to smaller ORF1 sequence. The experimental analysis performed would not discriminate whether these sequences arose from single or double stranded DNA. The read coverage of the SAG8 and SAG12 viruses is 50x and 7893x respectively. Assuming MDA amplification did not induce coverage biases between the virus and nuclear genomes, the same average coverage would be expected for viral and nuclear nucleotides. Compared to the average coverage of the nuclear genome, the SAG8 associated virus had a 10-fold lower coverage, while the SAG12 associated virus had a 5-fold higher coverage. This supports the notion that the virus was present in multiple copies (5 copies) within the cytoplasm of SAG12, suggesting an infected state, while there were fewer copies, and may have been under-amplified, in SAG8. The majority of circular DNA viruses were “Circular Replication-associated protein Encoding Single-Stranded DNA”

(CRESS) viruses (76). They have been found within diverse hosts and marine metagenomes, even though there is not much information regarding their infection strategies. To explore the similarity of the Mamiellales associated circular virus to other CRESS DNA viruses, we performed Maximum likelihood phylogenetic analyses of their predicted capsid (Figure 5C) and predicted replicase (Figure 5D) proteins, together with their retrieved blast hits (e-value cut off 1E-5) both in GenBank and in the TARA-Oceans Gene Atlas. Capsid homologous sequences were found in seven TARA-Oceans metagenomes (Figure 5B) including station 41 in the Indian Ocean, the same station where the *Ostreococcus* SAG12 was sampled. The phylogenies of both genes suggested that Mamiellales associated circular viruses formed a well-supported monophyletic clade, which included circo-like viruses, metagenomic sequences, and plant ssDNA viruses (Yerba mate virus and Banana bunchy top virus). Sequences from other ssDNA viruses clades, such as genes encoded by Bacilladnaviruses (from Diatoms), belonged to divergent branches suggesting that Mamiellales associated circular viruses might form a divergent group of a novel ssDNA virus family.

4. Discussion

SAGs have so far been previously used to increase our knowledge of the uncultivable multitude of marine microorganisms (40)(74). The use of SAGs to discover novel species does not necessitate high genome coverage as only a handful of highly conserved genes is informative to delineate a species. However, partial genome coverage may constitute a greater challenge for the use of SAGs in evolutionary studies. To increase the genome coverage of a target species, data from genetically close SAGs may be merged together (33)(45). This approach has been successfully applied by Vannier et al (33) to the 4 *Bathycoccus* SAGs and resulted in an assembly estimated to contain 64% of the complete genome. The amino-acid identity between this novel assembly and the reference *B. prasinos* RCC1105 genome led the authors to conclude that these 4 SAGs belonged to a cryptic novel *Bathycoccus* species, TOSAG39, which is supported here by the phylogenetic divergence of their mitochondrial genome sequences. Although combining SAGs produces a more complete genome, individual information about the environment of the sorted single cells, which could indicate potential biological interactions, or even stages of infections in the case of parasites, may be lost.

Although the individual SAGs were estimated to represent a maximum of 27% of the complete genome sequence in *Ostreococcus*, a search for the presence of 23 gene families

located in the mating-type region in Mamiellales enabled us to retrieve orthologous genes for 3 *Ostreococcus* SAGs. Amino-acid identity between the SAG sequences and the sequences from available genomes indicated that the candidate mating type of the 3 *Ostreococcus* SAGs was *MT*+, like the candidate mating type of the closest reference genome from strain *O. sp* RCC809 (28). The relative prevalence of the *MT*+ and *MT*- in natural *Ostreococcus* blooms remains an open question. To the best of our knowledge, the co-occurrence of strains of the two mating types from the same sample has not yet been reported, but the maximum number of strains isolated from the same sampling point is less than 3 (77). The minimal frequency of sexual reproduction has been estimated indirectly in *Ostreococcus tauri* from the population polymorphism spectrum, 1 sexual meiosis for 100,000 clonal divisions (28). This low rate may be explained either by a low encounters rate of *MT*+ and *MT*- strains in their natural environment, or by a low rate of outcrossing. The identification of informative sequences from the *MT* locus in 3 of the 4 SAGs demonstrates the utility of the single cell genomics approach for estimating the frequencies of the two mating types within a bloom, provided that a sufficient number of cells is analysed. While the presence of identical *MT*s in 3 SAG is consistent with clonal reproduction within a bloom, the sample size was too small for a proper estimation of the level of clonality within a bloom.

The estimations of the percentages of reads recruited to the mitochondrion and chloroplast genomes in each SAG attest the ability of single cell sequencing to identify novel candidate associations in the smallest eukaryotic microalgae of the order Mamiellales. Indeed, while organellar copy numbers may reach several thousands of copies in humans (78) or *Arabidopsis* (79), bacterial-sized *Ostreococcus* have a reduced number of organellar genomes, with an average copy number per cell estimated at 4 and 5 for the mtDNA and cpDNA, respectively. As a consequence, the systematic identification of mtDNA and cpDNA sequences from the SAGs provides support to the ability of this strategy to detect genomic sequences of any intracellular symbionts, provided there is no DNA amplification bias as a consequence of differences in genome architecture or cell wall structure between the symbiont and the organellar genome. Alternatively, these sequences might arise from random capture of amplified dissolved DNA in the marine seawater around the isolated picoalgal cells. In a typical sorting run the droplet containing a single cell may contain about 2.8 nL (80), or 2.8×10^6 µm3. The sample volume within a drop is around 1/1000 of the total volume, or 2.8×10^3 µm3, whereas a single *Ostreococcus* cell contains about 0.5 µm3

of cytoplasm, giving about 6×10^3 fold more seawater than cell volume, ample space for the "phycosphere" (81) or dissolved DNA (82) that may be sequenced with the targeted algal cell.

To distinguish between random capture of dissolved DNA and intracellular symbionts, relative read coverage of the non-target to the target genomes provides a rule of the thumb about the likely origin of the non-target sequences. In the present data, we found that most of the non-target sequence data can be traced back to cross-contamination from another SAG, sequenced in the same Illumina run. Filtering out cross-contaminants, the amount of non-target DNA identified – except for the virus sequence - is not compatible with co-sampling of a Mamiellales and a bacteria or Oomycete cell, and rather suggests random capture of dissolved DNA. These associated non-target sequences may nevertheless provide information about the diversity of microorganisms present in the cell's environment. Indeed, Mamiellales are auxotrophic for B vitamins (83) and have been found to co-occur with diverse bacteria in lab cultures (29)(84), though no obligate specific association has been yet observed. The taxonomic diversity of the candidate associated bacterial contigs retrieved from the SAG data is thus not surprising.

The presence of a novel CRESS-like DNA virus came as a surprise. Single-stranded DNA (ssDNA) viruses with circular genomes, with sizes from ~1700 bases, are the smallest viruses known to infect eukaryotes (85). Several previous metagenomic studies have identified many novel genomes similar to ssDNA circular viruses through data-mining of public marine metagenomes (85)(86), but with no hints about their putative hosts. Previously described CRESS viral hosts range from plants (87) (Geminiviridae, Nanoviridae), bleached kelp macroalgae (88), vertebrates (89) (Circoviridae), fungi, insects (Genomoviridae) (90), and crustaceans (91). In protists, CRESS viruses were discovered in SAGs targeting Picobiliphytes (31) and in Bacillariophyceae diatoms, where they have been isolated and their lytic effect could be demonstrated (92). The Mamiellales associated circular viruses described here are among the smallest viral genomes described to date infecting the smallest photosynthetic eukaryotes.

In conclusion, sorting of these 12 Mamiellales SAGs by flow cytometry relied on the choice of appropriate volumes to favour isolation of one cell, and the "phycosphere" around cell had been subjected to turbulent mixing with the cytometer sheath fluid, thus limiting the suitability of this technique for identification of cell-to-cell interactions. Independent ways

of finding such associations are necessary, one of which may be to pick off single cells using micromanipulation (see (35) in this issue), but the technical challenge of performing this for picoeukaryotes < 2 μm in size, such those in the Mamiellales, remain very challenging. Conventional culture-based approaches of strains and viruses are required to demonstrate the outcome of interactions by co-culture (93) and remain an essential tool for deciphering both intercellular and virus-host interactions in aquatic environments and interpreting environmental sequencing data. Nevertheless, we showed here that - provided careful filtration of contaminations between dataset is performed - SAGs provide *in-situ* ecological information that would be lost in conventional isolation and culture processes, which select for the fastest growing cells in laboratory culture medium.

Acknowledgments

We would like to thank Tom Delmont and Olivier Jaillon from the Genoscope for information about the SAGs, Lilian Caesar for help with the figures, and the Genotoul platform (www.genotoul.fr) for cluster availability for bioinformatics analysis and Genoscope for sequencing. We are grateful to Fabien Burki for making a perl script for extracting taxonomic affiliation from DIAMOND outputs available.
Tara Oceans (which includes both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (http://oceans.taraexpeditions.org). We further thank the commitment of the following sponsors: CNRS (in particular Groupement de Recherche GDR3280 and the Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans-GOSEE), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, The French Ministry of Research, and the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL* Research University (ANR-11-IDEX-0001-02), Gordon and Betty Moore Foundation (award #3790) and the US National Science Foundation (awards OCE#1536989 and OCE#1829831) to MBS, as well as the Ohio Supercomputer for computational support. We also thank the support and commitment of agnès b. and Etienne Bourgois, the Prince Albert II de Monaco Foundation, the Veolia Foundation, Region Bretagne, Lorient Agglomeration, Serge Ferrari, Worldcourier, and KAUST. The global sampling effort was enabled by countless scientists and crew who sampled aboard the Tara from 2009-2013, and we thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expeditions. We are also grateful to the countries who graciously granted sampling permissions. The authors declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the analyses, publications, and ownership of data are free from legal entanglement or restriction by the various nations whose waters the Tara Oceans expeditions sampled in. This article is contribution number XX of Tara Oceans.

LFB was funded by the EU Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2015-675752. We would also like to acknowledge three anonymous referees for constructive comments on a previous version of this manuscript.

References

1. Piganeau G, Eyre-Walker A, Jancek S, Grimsley N, Moreau H. How and why DNA barcodes underestimate the diversity of microbial eukaryotes. PLoS ONE. 2011;6(2):e16342.
2. Piganeau G, Desdevises Y, Derelle E, Moreau H. Picoeukaryotic sequences in the Sargasso sea metagenome. Genome Biol. 2008;9(1):R5.
3. Vaulot D, Lepère C, Toulza E, De la Iglesia R, Poulain J, Gaboyer F, et al. Metagenomes of the picoalga Bathycoccus from the Chile coastal upwelling. PLoS ONE. 2012;7(6):e39648.
4. Monier A, Worden AZ, Richards TA. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. Environ Microbiol Rep. 2016 Aug;8(4):461–9.
5. Tragin M, Vaulot D. Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. Sci Rep. 2018 Sep 19;8(1):14020.
6. Not F, Latasa M, Marie D, Cariou T, Vaulot D, Simon N. A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. Appl Environ Microbiol. 2004;70(7):4064–72.
7. Worden AZ, Nolan JK, Palenik B. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. Limnol Oceanogr. 2004 Jan 1;49(1):168–79.
8. Simon N, LeBot N, Marie D, Partensky F, Vaulot D. Fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes to identify small phytoplankton by flow cytometry. Appl Environ Microbiol. 1995 Jul;61(7):2506–13.
9. Scherffel A. Zwei neue, trichocystenartige Bildungen führende Flagellaten. Archiv für Protistenkunde. 27:94–128.
10. Knightjones EW, Walne PR. Chromulina-Pusilla Butcher, a Dominant Member of the Ultraplankton. Nature. 1951;167(4246):445–6.
11. Moestrup øjvind. Further studies on Nephroselmis and its allies (Prasinophyceae). II. Mamiella gen. nov., Mamiellaceae fam. nov., Mamiellales ord. nov. Nordic Journal of Botany. 1984;4(1):109–21.
12. Manton I. Dolichomastix (Prasinophyceae) from arctic Canada, Alaska and South Africa: a new genus of flagellates with scaly flagella. Phycologia. 1977 Dec 1;16(4):427–38.
13. Eikrem W, Throndsen J. The Ultrastructure of Bathycoccus Gen-Nov and Bathycoccus-Prasinos Sp-Nov, a Nonmotile Picoplanktonic Alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. Phycologia. 1990;29(3):344–50.
14. Chretiennot-Dinet MJ, Courties C, Vaquer A, Neveux J, Claustres H, Lautier J, et al. A new marine picoeukaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). Phycologia. 1995;34(4):285–92.
15. Marin B, Melkonian M. Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. Protist. 2010 Apr;161(2):304–36.
16. Guiry MD, Guiry GM. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; 2017.
17. Tragin M, Vaulot D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. Scientific Reports. 2019 Mar 26;9(1):5190.
18. Keller MD, Selvin RC, Claus W, Guillard RRL. MEDIA FOR THE CULTURE OF OCEANIC ULTRAPHYTOPLANKTON1,2. Journal of Phycology. 1987 Dec 1;23(4):633–8.
19. Ral JP, Derelle E, Ferraz C, Wattebled F, Farinas B, Corellou F, et al. Starch division and partitioning. A mechanism for granule propagation and maintenance in the picophytoplanktonic green alga Ostreococcus tauri(1[w]). Plant Physiology. 2004;136(2):3333–40.
20. O'Neill JS, van Ooijen G, Dixon LE, Troein C, Corellou F, Bouget FY, et al. Circadian rhythms persist without transcription in a eukaryote. Nature. 2012;469(7331):554–8.
21. Sanchez-Ferandin S, Leroy F, Bouget FY, Joux F. A new, sensitive marine microalgal recombinant biosensor using luminescence monitoring for toxicity testing of antifouling biocides. Appl Environ Microbiol. 2012;79(2):631–8.

22. Demory D, Baudoux A-C, Monier A, Simon N, Six C, Ge P, et al. Picoeukaryotes of the Micromonas genus: sentinels of a warming ocean. ISME J. 2019 Jan;13(1):132–46.
23. Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget F-Y, Carré I, et al. An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. BMC Genomics. 2014;15:1103.
24. Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 2007;104(18):7705–10.
25. Yau S, Krasovec M, Rombauts S, Groussin M, Benites LF, Vancaester E, et al. Virus-host coexistence in phytoplankton through the genomic lens. bioRxiv. 2019 Jan 1;513622.
26. van Baren MJ, Bachy C, Reistetter EN, Purvine SO, Grimwood J, Sudek S, et al. Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. BMC Genomics. 2016 Mar 31;17:267.
27. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. Genome Biol. 2012;13(8):R74.
28. Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, et al. Population genomics of picophytoplankton unveils novel chromosome hypervariability. Science Advances. 2017 Jul 1;3(7):e1700239.
29. Abby SS, Touchon M, De Jode A, Grimsley N, Piganeau G. Bacteria in Ostreococcus tauri cultures - friends, foes or hitchhikers? Front Microbiol. 2014;5:505.
30. Yau S, Grimsley N, Moreau H. Molecular ecology of Mamiellales and their viruses in the marine environment. Perspectives in Phycology. 2015 Oct 21;83–9.
31. Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science. 2011 May 6;332(6030):714–7.
32. Bhattacharya D, Price DC, Bicep C, Bapteste E, Sarwade M, Rajah VD, et al. Identification of a Marine Cyanophage in a Protist Single-cell Metagenome Assembly. J Phycol. 2013 Feb;49(1):207–12.
33. Vannier T, Leconte J, Seeleuthner Y, Mondy S, Pelletier E, Aury J-M, et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. Scientific Reports. 2016 Nov 30;6:37900.
34. Strassert JFH, Karnkowska A, Hehenberger E, Del Campo J, Kolisko M, Okamoto N, et al. Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. ISME J. 2018 Jan;12(1):304–8.
35. Galindo LJ, Torruella G, Moreira D, Eglit Y, Simpston A, Voelker E, et al. Combined cultivation and single-cell approaches to the phylogenomics of nucleariid amoebae, close relatives of Fungi. Proc R Soc London B Biological sciences. 2019;in press.
36. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data. 2017 Aug 1;4:170093.
37. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A holistic approach to marine eco-systems biology. PLoS Biol. 2011;9(10):e1001177.
38. Tara Oceans Consortium C, Tara Oceans Expedition P. Registry of all stations from the Tara Oceans Expedition (2009-2013) [Internet]. PANGAEA; 2015. Available from: https://doi.org/10.1594/PANGAEA.842237
39. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002 Apr 16;99(8):5261–6.
40. Stepanauskas R, Sieracki ME. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc Natl Acad Sci U S A. 2007 May 22;104(21):9052–7.
41. Martinez-Garcia M, Brazel D, Poulton NJ, Swan BK, Gomez ML, Masland D, et al. Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. ISME J. 2012 Mar;6(3):703–7.
42. Stepanauskas R, Sieracki ME. Matching phylogeny and metabolism in the uncultured marine

bacteria, one cell at a time. Proceedings of the National Academy of Sciences of the United States of America. 2007 May 22;104(21):9052–7.
43. Martinez-Garcia M, Brazel D, Poulton NJ, Swan BK, Gomez ML, Masland D, et al. Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. The ISME journal. 2012 Mar;6(3):703–7.
44. Mangot J, Logares R, Sanchez P, Latorre F. Accessing to the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Scientific Reports. 2017;7.
45. Mangot J-F, Logares R, Sanchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep. 2017 Jan 27;7:41498.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455–77.
47. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
48. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.
49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
50. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.
51. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.
52. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010 Mar 1;26(5):589–95.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
54. R Core Team. R: A language and environment for statistical computing. [Internet]. R Foundation for Statistical Computing, Vienna, Austria.; 2015. Available from: http://www.R-project.org/
55. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012 Jun 15;28(12):1647–9.
56. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 1998 Jan 15;26(2):544–8.
57. Gorbalenya AE, Koonin EV, Wolf YI. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett. 1990 Mar 12;262(1):145–8.
58. Krupovic M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol. 2013 Oct;3(5):578–86.
59. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018 Jan 25;9(1):373.
60. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research. 2005;33(2):511–8.
61. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012 Jul 30;9(8):772.
62. Ferris PJ, Pavlovic C, Fabry S, Goodenough UW. Rapid evolution of sex-related genes in Chlamydomonas. Proc Natl Acad Sci USA. 1997 Aug 5;94(16):8634–9.
63. Garcia-Moreno J, Mindell DP. Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD. Mol Biol Evol. 2000 Dec;17(12):1826–32.
64. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. 2015 Aug 6;16(1):157.
65. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote

Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA. 2007 May 1;104(18):7705–10.
66. Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nature Microbiology. 2018 Jul 1;3(7):804–13.
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
68. Villar E, Vannier T, Vernette C, Lescot M, Cuenca M, Alexandre A, et al. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. Nucleic Acids Res. 2018 Jul 2;46(W1):W289–95.
69. Simon N, Foulon E, Grulois D, Six C, Desdevises Y, Latimier M, et al. Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species M. pusilla (Butcher) Manton & Parke and of the Species M. commoda van Baren, Bachy and Worden and Description of Two New Species Based on the Genetic and Phenotypic Characterization of Cultured Isolates. Protist. 2017 Sep 14;168(5):612–35.
70. Mangot J-F, Logares R, Sanchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep. 2017 Jan 27;7:41498.
71. Blanc-Mathieu R, Sanchez-Ferandin S, Eyre-Walker A, Piganeau G. Organellar inheritance in the green lineage: insights from Ostreococcus tauri. Genome Biol Evol. 2013;5(8):1503–11.
72. Blanc-Mathieu R, Sanchez-Ferandin S, Eyre-Walker A, Piganeau G. Organellar inheritance in the green lineage: insights from Ostreococcus tauri. Genome Biol Evol. 2013;5(8):1503–11.
73. Lejzerowicz F, Pawlowski J, Esling P. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. Nucleic Acids Research. 2015 Feb 17;43(5):2513–24.
74. Seeleuthner Y, Mondy S, Lombard V, Carradec Q, Pelletier E, Wessner M, et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat Commun. 2018 Jan 22;9(1):310.
75. McGowan J, Byrne KP, Fitzpatrick DA. Comparative Analysis of Oomycete Genome Evolution Using the Oomycete Gene Order Browser (OGOB). Genome Biology and Evolution. 2018 Dec 11;11(1):189–206.
76. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. Adv Virus Res. 2019;103:71–133.
77. Grimsley N, Pequin B, Bachy C, Moreau H, Piganeau G. Cryptic sex in the smallest eukaryotic marine green alga. Mol Biol Evol. 2010;27(1):47–54.
78. Miller FJ, Rosenfeldt FL, Zhang C, Linnane AW, Nagley P. Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. Nucleic Acids Res. 2003 Jun 1;31(11):e61–e61.
79. Morley SA, Nielsen BL. Chloroplast DNA Copy Number Changes during Plant Development in Organelle DNA Polymerase Mutants. Front Plant Sci. 2016 Feb 4;7:57–57.
80. Sieracki M, Poulton N, Crosbie N. Automated Isolation Techniques for Microalgae. In: Algal Culturing Techniques. 2005. p. 101–16.
81. Seymour JR, Amin SA, Raina J-B, Stocker R. Zooming in on the phycosphere: the ecological interface for phytoplankton-bacteria relationships. Nat Microbiol. 2017 May 30;2:17065.
82. Deflaun MF, Paul JH, Davis D. Simplified method for dissolved DNA determination in aquatic environments. Appl Environ Microbiol. 1986 Oct;52(4):654–9.
83. Paerl RW, Bouget F-Y, Lozano J-C, Vergé V, Schatt P, Allen EE, et al. Use of plankton-derived vitamin B1 precursors, especially thiazole-related precursor, by key marine picoeukaryotic phytoplankton. The Isme Journal. 2016 Dec 9;11:753.
84. Lupette J, Lami R, Krasovec M, Grimsley NH, Moreau H, Piganeau G, et al. Marinobacter dominates the bacterial community of the Ostreococcus tauri phycosphere in culture. Front Microbiol. 2016;7:1414.
85. Rosario K, Duffy S, Breitbart M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol. 2009 Oct;90(Pt 10):2418–24.

86. Labonte JM, Suttle CA. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J. 2013 Nov;7(11):2169–77.
87. Zerbini FM, Briddon RW, Idris A, Martin DP, Moriones E, Navas-Castillo J, et al. ICTV Virus Taxonomy Profile: Geminiviridae. J Gen Virol. 2017 Feb;98(2):131–3.
88. Beattie DT, Lachnit T, Dinsdale EA, Thomas T, Steinberg PD. Novel ssDNA Viruses Detected in the Virome of Bleached, Habitat-Forming Kelp Ecklonia radiata. Frontiers in Marine Science. 2018;4:441.
89. Breitbart M, Delwart E, Rosario K, Segales J, Varsani A, Consortium IR. ICTV Virus Taxonomy Profile: Circoviridae. J Gen Virol. 2017 Aug;98(8):1997–8.
90. Krupovic M, Ghabrial SA, Jiang D, Varsani A. Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch Virol. 2016 Sep;161(9):2633–43.
91. Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, et al. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ. 2018;6:e5761.
92. Shirai Y, Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom Chaetoceros tenuissimus Meunier. Appl Environ Microbiol. 2008 Jul;74(13):4022–7.
93. Amin SA, Hmelo LR, van Tol HM, Durham BP, Carlson LT, Heal KR, et al. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. Nature. 2015 Jun 4;522(7554):98–101.

Tables

Table 1: Assembly statistics of 12 SAGs and statistics of reference genomes from the order Mamiellales

Sample - station	**Bathycoccus - 39**				**Micromonas - 46**				**Ostreococcus - 41**			
SAG	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Assembly size (Mb)	3.25	4.42	3.36	3.72	0.89	0.56	0.58	0.53	0.85	3.46	1.54	2.71
N50 (kb)	13.6	17.4	19.5	20.2	3.17	2.47	9.67	6.91	6.44	37.5	17.8	36.4
GC%	45.7	47.0	46.6	46.5	40.0	39.3	41.6	39.1	52.2	57.1	52.1	56.6
Contigs	3157	3129	3461	2913	2864	1894	1043	1788	2914	2479	2207	2364
Contigs (> 400 bp)	1048	1241	990	991	691	470	272	275	410	737	546	655
Reference Genome	*B. prasinos* RCC1105				*M. commoda* RCC299				*O. tauri* RCC4221			
Nuclear (GC%)	15 Mb (48.0)				21 Mb (63.8)				13 Mb (59.0)			
Mitochondria (bp) (GC%)	43,614 (40.1)				47,425 (34.6)				44,237 (38.2)			
Chloroplast (bp) (GC%)	72,700 (43.3)				72,585 (38.8)				67,681 (39.9)			

Table 2: Summary statistics of assemblies per taxonomic affiliation categories, assembly length and percent of reads recruited are provided for Target (Mamiellophyceae) and Non Target assemblies (Bacteria, Eukaryote, Archaebacteria and virus).

* Tara-Oceans Expedition station number

		Target sequences						Non Target sequences				BUSCO
		Assembly total length (% of reads recruited)						Total length (% of reads recruited)				
T.-O.*	SAG	Total		Mitochondria		Chloroplast		Total		Filtered		
Station		x10^3 bp	(%)	bp	(%)	bp	(%)	bp	(%)	bp	(%)	(%)
39	SAG1	3,001	(99.89)	40,481	(1.46)	33,423	(0.03)	101,751	(0.09)	29,645	(0.03)	11.9
39	SAG2	4,166	(99.47)	49,421	(0.05)	21,954	(0.01)	89,862	(0.28)	22,815	(0.07)	17.2
39	SAG3	3,130	(99.91)	46,597	(0.11)	35,348	(0.05)	90,824	(0.06)	46,751	(0.03)	15.2
39	SAG4	3,509	(99.96)	54,044	(0.08)	38,960	(0.04)	118,507	(0.02)	32,028	(0.005)	16.8
46	SAG5	441	(90.32)	54,577	(3.36)	51,112	(65.97)	228,342	(2.15)	122,933	(1.16)	2.0
46	SAG6	271	(98.26)	37,972	(20.66)	64,962	(48.70)	135,783	(0.26)	33,796	(0.06)	1.0
46	SAG7	255	(96.12)	34,404	(0.28)	53,356	(73.30)	136,620	(2.02)	26,962	(0.39)	2.3
46	SAG8	187	(82.95)	44,825	(14.67)	42,081	(0.21)	144,189	(6.09)	31,813	(1.34)	3.0
41	SAG9	712	(99.86)	43,157	(1.58)	40,737	(0.15)	41,906	(0.08)	6,307	(0.01)	28.7
41	SAG10	3,180	(99.88)	47,961	(1.11)	93,684	(0.79)	55,576	(0.01)	11,764	(0.002)	22.8
41	SAG11	1,225	(92.74)	28,820	(19.66)	101,901	(7.98)	142,887	(1.78)	20,493	(0.25)	7.6
41	SAG12	2,250	(98.85)	43,041	(0.15)	84,203	(10.60)	79,669	(1.11)	11,876	(0.16)	15.5

Table 3: Summary statistics of cpDNA and mtDNA manual assemblies from combined SAG sequence data.

	Chloroplast genome (cpDNA)		Mitochondrial genome (mtDNA)	
	assembly size (Kb)	assembly name	assembly size (Kb)	assembly name
SAG1-4	-	-	48.8	Ba_mt_TOSAG39 complete
SAG5-8	54.7	Mi_cp_TOSAG46 partial	33.3	Mi_mt_TOSAG46 partial
SAG9-12	68.9	Os_cp_TOSAG41 complete	39.1	Os_mt_TOSAG41 partial

Figure captions

FIG1 - Maximum likelihood phylogeny (under GTR model) of 18S rDNA sequences retrieved from the SAG assemblies of Bathycoccus sp. (SAG1-3), Micromonas sp. (SAG5 and 8) and Ostreococcus sp. (SAG9-12). Mamiellales reference genomes are marked with a red star, and SAGs are marked with orange star. Rapid bootstrap support values are indicated in the key (100 replicates), represented by colored circles in branches. The final alignment contained 1750 nucleotides.

FIG2 - Mating type inference from amino-acid identity with MT+ and MT- orthologous gene families (GF) in Ostreococcus SAGs (a) Expected taxonomic affiliation for MT- and MT+ strains (b) Presence/absence matrix of hits to gene families in each SAGs (c) Maximum likelihood phylogeny (LG+G model) of the SAGs sequence corresponding to GF16 with orthologous genes from *Ostreococcus* species.

FIG3 - Taxonomic affiliations of non-target sequences. (a) Percent of raw read number (b) Percent of reads following cross-contamination read filtering.

FIG4 - Taxonomic affiliation inferred from best blast hits (1e-5 cut-off) of Bacterial contigs (a) at level of the phylum, Oomycetes contigs at level of family (b), sum of contig lengths per taxa in each SAG is indicated by bubble size. Length of contig recruited to TARA Microbial Assembled Genomes (MAGs) from TARA metagenomes is indicated by bubble color.

FIG5 – Genome structure, biogeography and phylogeny of Mamiellales associated circular virus. (a) Genome assembly of 1,586 bp virus encoding two open reading frames ; ORF1 is a replicase and ORF2 a putative capsid protein. (b) Biogeographic distribution of putative homologous sequences encoding to ORF2 (putative capsid) in the Ocean Gene Atlas (OGA) database per TARA station (triangles) and number of hits per station (bubble size) (b). (c) Maximum Likelihood Phylogeny of ORF2 (putative capsid – RTREV+I+G+F protein

evolution model) and (d) ORF1 (replicase – LG+I+G protein evolution model) with homologous sequences retrieved from GenBank and OGA database.

Figures

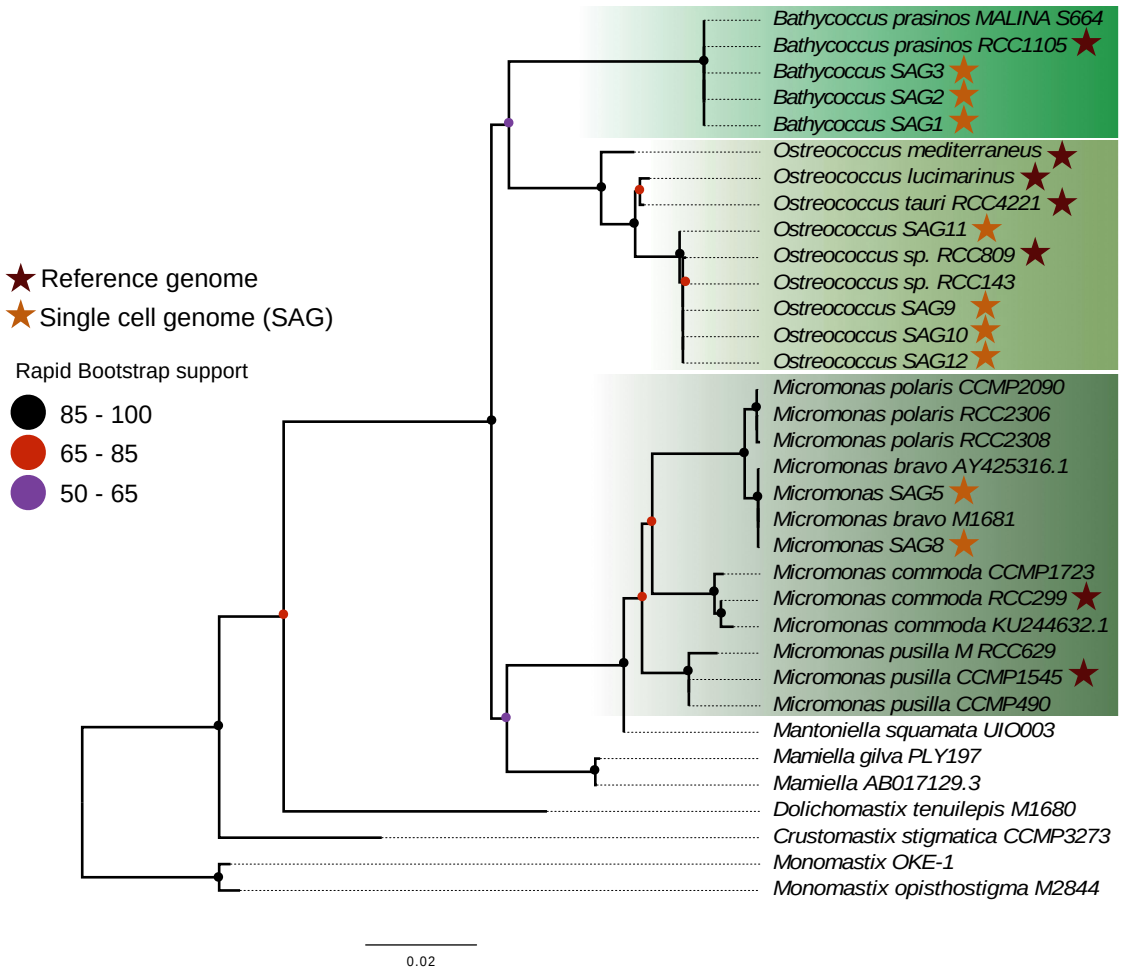

FIG1 - Maximum likelihood phylogeny (under GTR model) of 18S rDNA sequences retrieved from the SAG assemblies of Bathycoccus sp. (SAG1-3), Micromonas sp. (SAG5 and 8) and Ostreococcus sp. (SAG9-12). Mamiellales reference genomes are marked with a red star, and SAGs are marked with orange star. Rapid bootstrap support values are indicated in the key (100 replicates), represented by colored circles in branches. The final alignment contained 1750 nucleotides.

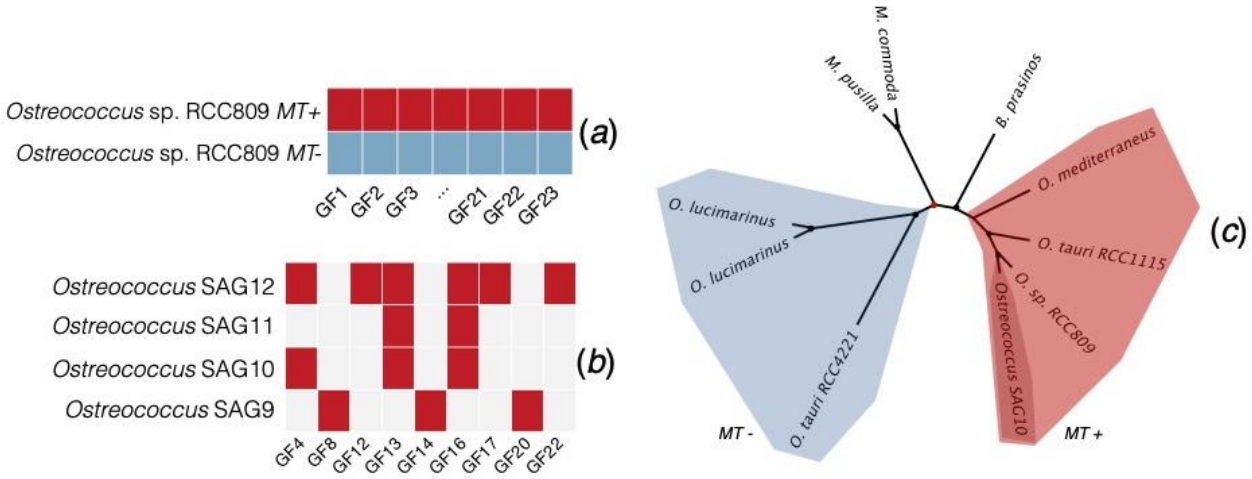

FIG2 - Mating type inference from amino-acid identity with *MT*+ and *MT*- orthologous gene families (GF) in *Ostreococcus* SAGs (a) Expected taxonomic affiliation for *MT*- and *MT*+ strains (b) Presence/absence matrix of hits to gene families in each SAGs (c) Maximum Likelihood phylogeny (LG+G model) of the SAGs sequence corresponding to GF16 with orthologous genes from *Ostreococcus*, *Micromonas* and *Bathycoccus* species.

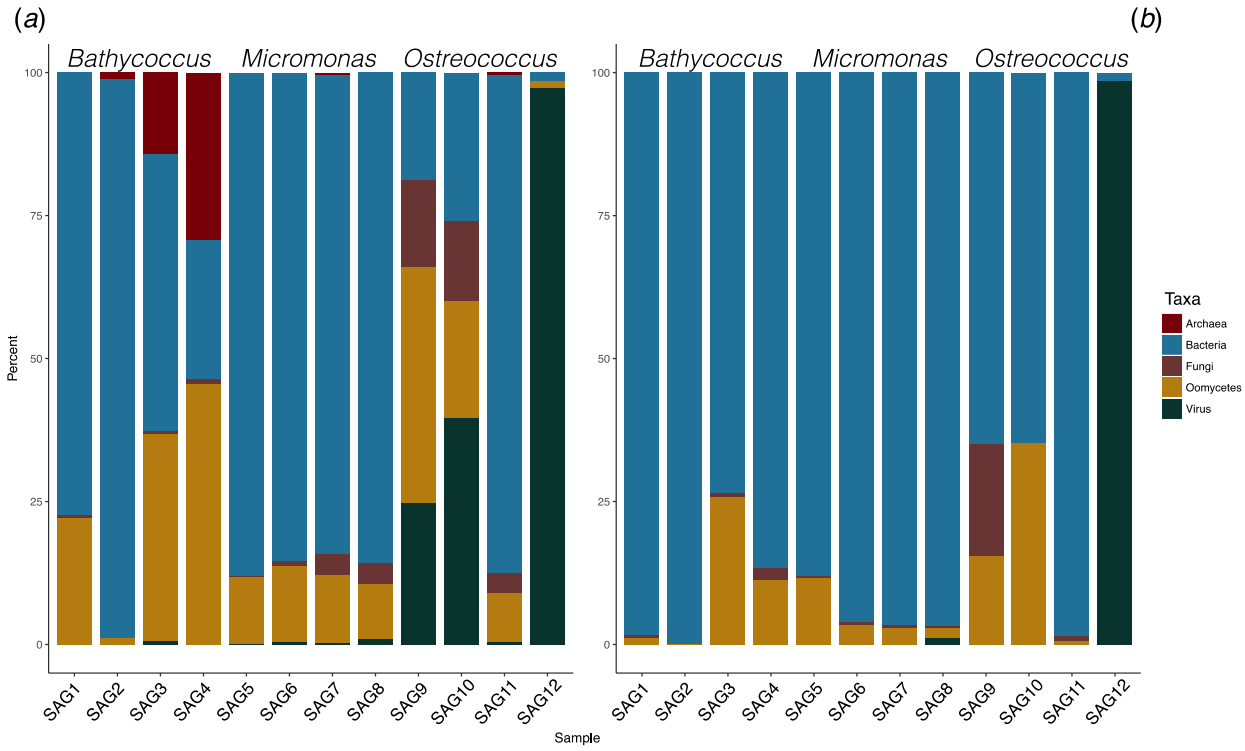

FIG3 - Taxonomic affiliations of non-target sequences. (a) Percent of raw read number (b) Percent of reads following cross-contamination read filtering.

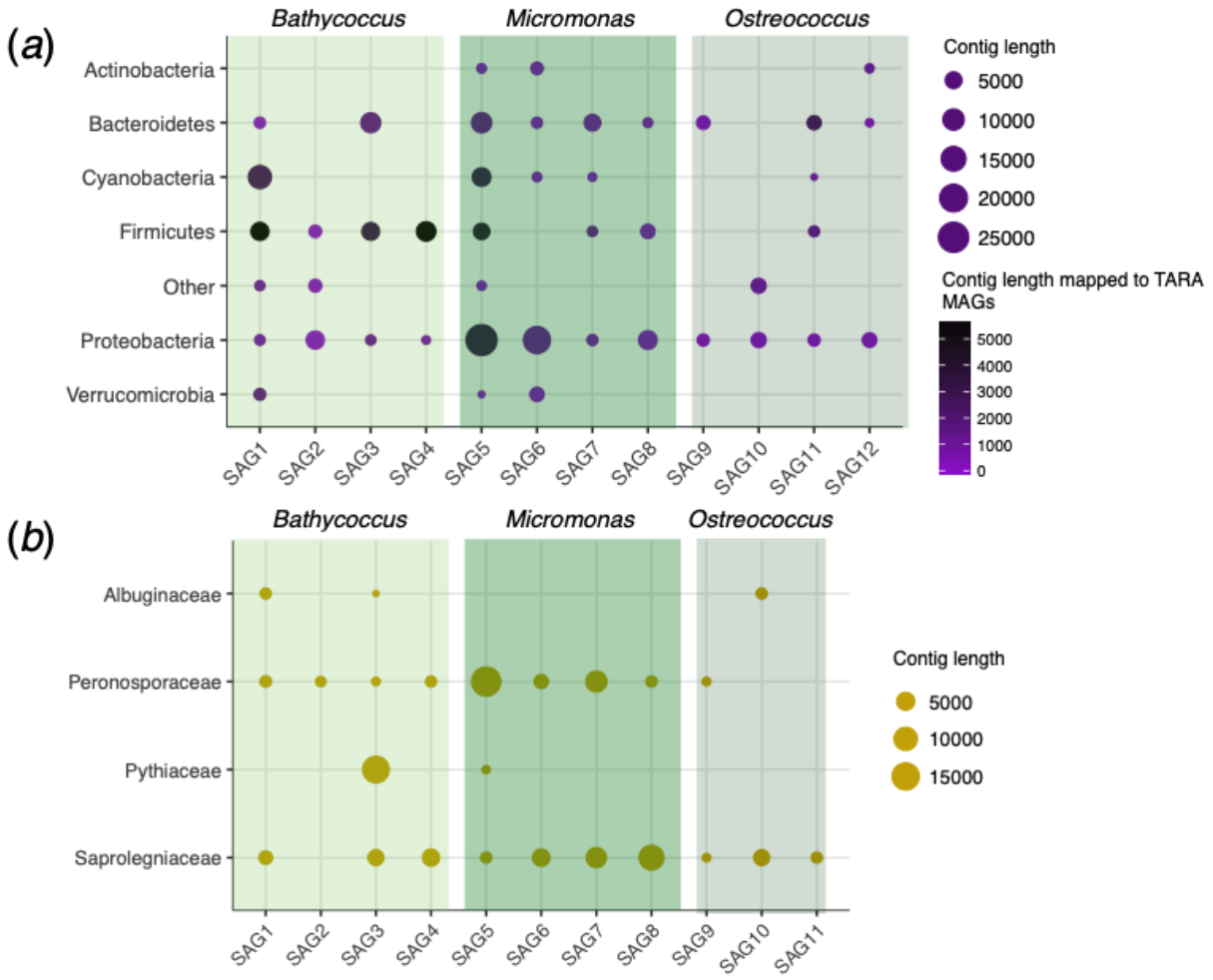

FIG4 - Taxonomic affiliation inferred from best blast hits (1e-5 cut-off) of Bacterial contigs (a) at level of the phylum, Oomycetes contigs at level of family (b), sum of contig lengths per taxa in each SAG is indicated by bubble size. Length of contig recruited to TARA Microbial Assembled Genomes (MAGs) from TARA metagenomes is indicated by bubble color.

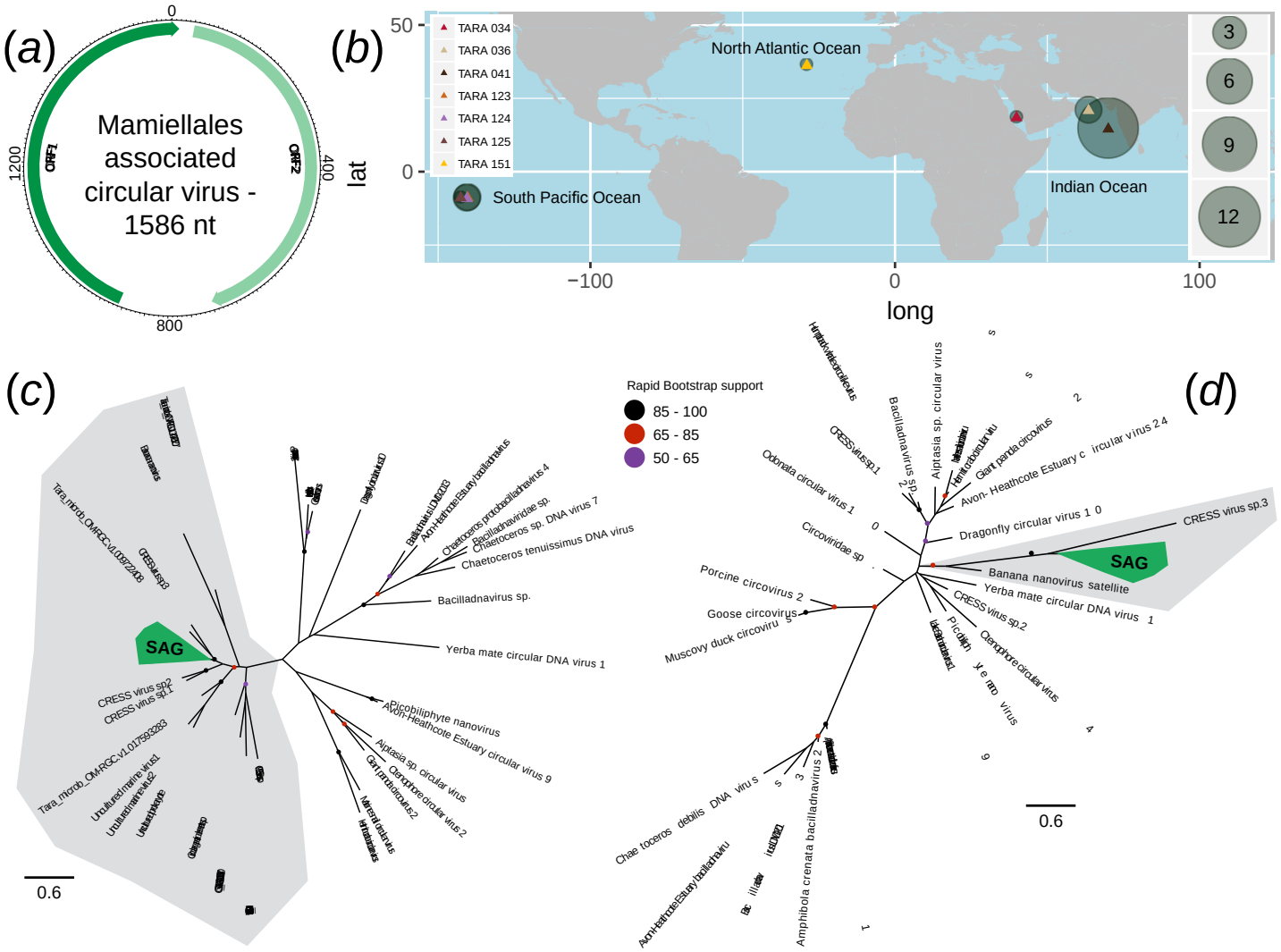

FIG5 – Genome structure, biogeography and phylogeny of Mamiellales associated circular virus. (a) Genome assembly of 1,586 bp virus encoding two open reading frames ; ORF1 is a replicase and ORF2 a putative capsid protein. (b) Biogeographic distribution of putative homologous sequences encoding to ORF2 (putative capsid) in the Ocean Gene Atlas (OGA) database per TARA station (b). (c) Maximum Likelihood Phylogeny of (putative capsid – RTREV+I+G+F protein evolution model) and (d) ORF1 (replicase – LG+I+G protein evolution model) with homologous sequences retrieved from GenBank and OGA database.

Supplementary material

Supplementary material table 1. GenBank accession numbers of 18S sequences used for phylogeny.

Supplementary material table 2. GenBank accession numbers for virus sequences used in Fig.5.

Supplementary material table 3. Taxonomic affiliation of non-target bacterial sequences for each SAG.

Supplementary material table 4. Plastid, mitochondrion and nuclear genome coverage in 12 *Ostreococcus tauri* strains and estimation of mitochondrial and plastid genome copy number per haploid nuclear genome.

Additional Information

Ethics
Not applicable.

Data Accessibility
SAGs assemblies analysed in this manuscript have been submitted to GenBank under BioProject PRJNA549236 with accession numbers VIAS00000000 (SAG1); VIAT00000000 (SAG2); VIAU00000000 (SAG3); VIAV00000000 (SAG4); VIAW00000000 (SAG5); VIAX00000000 (SAG6); VIAY00000000 (SAG7); VIAZ00000000 (SAG8); VIBA00000000 (SAG9); VIBB00000000 (SAG10); VIBC00000000 (SAG11); VIBD00000000 (SAG12). Mitochondrial assemblies have been submitted to GenBank under accession MN243825 (Os_mt_TOSAG41), MN243826 (Mi_mt_TOSAG46), plastidial assemblies under accession MN243827 (Os_cp_TOSAG41) and MN243828 (Mi_cpTOSAG46).

Authors' Contributions
LFB and GP performed bioinformatics analyses. NP and MS performed single cell sorting and prior species' taxonomic affiliation. KL performed SAG sequencing. GP, LFB and NG designed the study. All authors contributed to manuscript writing and editing.

Competing Interests
We have no competing interests.