Biological and genomic resources for the cosmopolitan phytoplanktonBathycoccus: Insights into genetic diversity and major structural variations

Louis Dennu, Martine Devic, Janaina Rigonato, Angela Falciatore, Jean-Claude Lozano, Valérie Vergé, Cédric Mariac, Olivier Jaillon, François Sabot, François-Yves Bouget

Biological and genomic resources for the cosmopolitan phytoplankton *Bathycoccus*: Insights into genetic diversity and major structural variations

Louis Dennu[1], Martine Devic*[1], Janaina Rigonato[2], Angela Falciatore[3], Jean-Claude Lozano[1], Valérie Vergé[1], Cédric Mariac[4], Olivier Jaillon[2], The Dark Edge genomics sampling team[†], François Sabot*[4] and François-Yves Bouget*[1]

* Corresponding authors : martine.devic@obs-banyuls.fr ; francois.sabot@ird.fr ; francois-yves.bouget@obs-banyuls.fr

Author affiliations :

[1] Laboratoire d'Océanographie Microbienne (LOMIC), CNRS/Sorbonne Université, UMR 7621, Observatoire Océanologique, 66650 Banyuls/mer, France.

[2] Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France.

[3] Laboratoire de Biologie du Chloroplaste et Perception de la Lumière chez les Microalgues, CNRS, UMR 7141, Sorbonne Université, Institut de Biologie Physico-Chimique, Paris, France.

[4] Diversité, adaptation et développement des plantes (DIADE), IRD/UM/CIRAD, UMR 232, Centre IRD de Montpellier, 911 avenue Agropolis, BP 604501, 34394, Montpellier Cedex 5, France.

† A list of authors and their affiliations appears at the end of the paper.

Keywords: Phytoplankton, *Mamiellophyceae*, Natural diversity, Whole genome sequencing, *Bathycoccus*, Structural variations

Word count : 7,029

1. Abstract

Population-scale sequencing has become a standard practice to explore the natural genetic diversity underlying adaptation, notably in land plants. However, current sequencing initiatives for eukaryotic phytoplankton primarily concentrate on creating reference genomes for model organisms and characterizing natural communities through metagenomics approaches. Consequently, few species have been thoroughly sequenced and intraspecific genetic diversity remains virtually undescribed, limiting our understanding of diversity and adaptation mechanisms. Here we report a biological and genomic resource to explore the genetic diversity of the cosmopolitan and ecologically important *Bathycoccus* genus. To span broad geographical and temporal scales, we selected available strains but also isolated and

genotyped strains from both the Banyuls bay (Mediterranean sea) and the Baffin bay (Arctic ocean). By combining ONT long reads and Illumina short reads technologies, we produced and annotated 28 *Bathycoccus sp. de novo* assembled genomes of high quality, including 24 genomes of *Bathycoccus prasinos* strains along a latitudinal gradient between 40° and 78° North, one reference genome of the *Bathycoccus calidus* species and 3 genomes of a yet undescribed *Bathycoccus* species named *Bathycoccus catiminus*. We assessed the genetic diversity of this genus through phylogenomic analyses and highlighted the central role of this genomic resource in providing new insights into the diversity of outlier chromosomal structures. The *Bathycoccus* biological and genomic resources offer a robust framework for investigating the diversity and adaptation mechanisms of eukaryotic phytoplankton in the Ocean.

2. Significance statement

Comparative and functional approaches for the study of eukaryotic phytoplankton and their adaptation to latitudes and seasons that rely on extensive biological and genomic resources are currently lacking. Here we report such resources and describe the natural diversity of the cosmopolitan phytoplankton *Bathycoccus*, providing insights into its species and intraspecific diversity and establishing it as a robust model for functional and ecological studies.

3. Introduction

Accessing the natural diversity of a species constitutes an invaluable resource, providing insights into its evolutionary history and facilitating correlations between phenotypes and genotypes through functional studies. In the case of plants, the availability of extensive collections of natural accessions and crop cultivars from diverse environments has greatly improved our knowledge of adaptation. Furthermore, large-scale sequencing projects on widely distributed models, such as for *Arabidopsis thaliana*, and in several model crops, have provided a comprehensive framework for the study of genomic diversity, its underlying mechanisms and their impact on environmental adaptation (100 Tomato Genome Sequencing Consortium, 2014; 3,000 Rice Genomes Project, 2014; Alonso-Blanco et al., 2016). Other reference organisms have also emerged as models in population genomics, such as *Saccharomyces cerevisiae* for which extensive sequencing of available strains highlighted complex genomic structures with great diversity (Peter et al., 2018). While these comprehensive sequencing projects have greatly benefited their research communities, the study of eukaryotic phytoplankton lags behind in terms of comparable large-scale sequencing initiatives.

The global distribution and high abundance of phytoplankton species make them major actors of the planet's primary production, contributing as much as land plants (Li, 1994; Worden et al., 2004). In particular, eukaryotic phytoplankton species are characterized by their great taxonomic diversity and relatively short generation times allowing for their

ecological success in a wide range of environmental conditions in the world Ocean (Lynch et al., 1991; De Vargas et al., 2015). Microalgae also serve as valuable models for both fundamental studies and biotechnological research. Among eukaryotic phytoplankton, the class *Mamiellophyceae* stands out as unicellulars widely distributed in the oceans from poles to equator and marked by seasonal population dynamics in polar and temperate regions (Lambert et al., 2019; Leconte et al., 2020). *Mamiellophyceae* diverged at the base of the green lineage, making them attractive models for studying essential cellular functions from an evolutionary perspective (Leliaert et al., 2012; Yung et al., 2022). The early whole genome sequencing of *Ostreococcus tauri* and the implementation of genetic transformation tools, including gene targeting by homologous recombination, have established *O. tauri* as a model for the study of several cellular pathways such as the circadian clock, the cell division, iron metabolism and vitamin B1 metabolism (Derelle et al., 2006; Corellou et al., 2009; Moulager et al., 2010; O'Neill et al., 2011; Van Oijen et al., 2013; Lozano et al., 2014; Botebol et al., 2015; Paerl et al., 2017; De Barros Dantas et al., 2023). The genomic content of *O. tauri* unveiled several features shared by *Mamiellophyceae* : A compact haploid genome, a majority of monoexonic genes, limited gene redundancy and two outlier chromosomes, characterized by lower GC content and unique gene structures, respectively named big outlier chromosome (BOC) and small outlier chromosome (SOC) (Derelle et al., 2006; Grimsley et al., 2015). The BOC is putatively involved in mating mechanisms, as evidenced by the identification of two haplotypes in several *Mamiellophyceae* and potential recombination suppression, but experimental evidence for sexual reproduction in *Mamiellophyceae* is still lacking (Blanc-Mathieu et al., 2017). The SOC is hypervariable between strains and has been shown to be associated with viral resistance mechanisms (Yau et al., 2016; Blanc-Mathieu et al., 2017). While *O. tauri* offers valuable insight into the origin and evolution of biological processes in the green lineage, its low abundance in marine environments, as inferred from metagenomic dataset, hinder its broader use in investigating questions relevant to ecology and adaptation (Demir-Hilton et al., 2011).

In addition to *Ostreococcus*, the *Bathycoccaceae* family also includes the genus *Bathycoccus*, a picoalga characterized by its scales-covered cell (Eikrem & Throndsen, 1990). Mostly studied through metagenomic approaches, the two currently described *Bathycoccus* species, *Bathycoccus prasinos* (Moreau et al., 2012) and *Bathycoccus calidus* (Bachy et al, 2021), show high abundance and distinct distribution patterns in marine environments. *B. prasinos* is detected at high latitudes between temperate and polar regions, while *B. calidus* is present in warm oligotrophic waters at lower latitudes (De Vargas et al. 2015; Vannier et al., 2016; Leconte et al., 2020). *B. prasinos* also exhibit strong seasonal patterns in both temperate and arctic waters, with population growth often occurring through annual bloom events (Joli et al., 2017; Lambert et al., 2019; Devic et al., 2023). The cosmopolitan distribution of *B. prasinos* from polar environment, marked by dramatic changes in photoperiod and temperature, to temperate Mediterranean climate makes this phytoplankton a prime model for the study of both latitudinal and seasonal adaptation mechanisms. The reference genome of *B. prasinos* is 15Mb in size and composed of 19

nuclear chromosomes supporting 7,847 annotated genes (Moreau et al., 2012). It features, as for other *Mamiellophyceae*, a BOC and a SOC corresponding to chromosomes 14 and 19 respectively. Additional genomic resources comprise single-cell assembled genomes of both *Bathycoccus* species and several environmental metagenome-assembled genomes, but the variable completion rate of genomes produced by these methods currently provides a fragmented and incomplete view of the genetic diversity within the *Bathycoccus* genus (Vaulot et al., 2012; Vannier et al., 2016; Joli et al., 2017; Benites et al., 2019; Delmont et al., 2022).

In this study, we aimed to draw a comprehensive landscape of *Bathycoccus* genetic diversity at both species and intraspecific (*i.e* in *B. prasinos*) levels and to provide a biological and genomic resource for future studies of molecular mechanisms underlying adaptation to environmental niches in *B. prasinos*. For this purpose we selected for sequencing a panel of *Bathycoccus* strains including (i) collection strains previously isolated in different geographic locations in north western europe, (ii) strains that have been isolated during the winter bloom of 2018-2019 in the Banyuls bay (Mediterranean sea, France) (Devic et al., 2023) and (iii) arctic strains from the Baffin bay (Arctic Ocean) we isolated and genotyped. Both ONT long reads and illumina short reads were used to generate 28 *de novo* genome assemblies of high quality and completeness. This resource was used to describe the genetic diversity of *Bathycoccus* at both species and intraspecific levels with a focus on outlier chromosomes.

4. Material and methods

Algal strains and culture conditions

Selection of *Bathycoccus sp.* strains from the Roscoff Culture Collection (RCC) and isolated strains from the Banyuls bay are detailed in Devic et al. (2023). Culture conditions of available and isolated Mediterranean strains are described in Devic et al. (2023). Arctic isolates were grown at 4°C or 15°C under constant light (10 µE).

Sampling and cell isolation

Sea water sampling at the SOLA buoy in Banyuls bay and the isolation of *Bathycoccus* strains were reported earlier by Devic et al. (2023).

Sampling in the Baffin bay was performed during the DarkEdge campaign in October, 2021 **(Figure 1A)**. For Baffin bay samples, 50 ml of sea water was filtered through a 1.2 µm pore-size acrodisc (FP 30/1.2 CA-S cat N° 10462260 Whatman GE Healthcare Sciences) and used to inoculate culture flasks with 10 ml of filtrate each. Sea water was supplemented with vitamins, NaH2PO4, NaNO3 and metal traces at the same concentration as in L1 culture medium. For samples isolated from the DE310 ice station, the salinity of the culture medium was halved by adding 10 ml of mQ water. Antibiotics (Streptomycin sulfate 100 µg/ml) were

added to half of the samples to limit bacterial growth. Finally, cultures were incubated on board either at 4° C or 15° C for 10-20 days under constant light before shipping through express carriers to Banyuls-sur-Mer (France) at 4° C or 15° C depending on the culture conditions. In the laboratory, the presence of picophytoplankton was checked using a BD accuri C6 flow cytometer. Cultures containing at least 90% of picophytoplankton were used for plating in low melting agarose 0.21% W/v as described by Devic et al. (2023). Colonies appearing after 10-15 days at 15° C, and up to a month at 4° C, were hand picked and further cultured. Individual *Bathycoccus* cells were obtained from 3 sampling zones at 4° C and 15° C with or without antibiotics **(Table S1)**.

Genotyping

The identification of *Bathycoccus sp.* strains from the Banyuls bay was performed through specific amplifications of a 614 bp fragment of the LOV-HK gene (Bathy10g02360) and 18S rDNA followed by Sanger sequencing (GENEWIZ), as reported by Devic et al. (2023). 55 isolates were unambiguously identified as *B. prasinos* by Devic et al. (2023), while 11 strains showed a low amplification signal with the LOV-HK primers but were determined to be *Bathycoccus sp.* by 18S rDNA sequencing results. We amplified the ITS2 region with 5'-GTACACACCGCCCGTCGC-3' and 5'-ATATGCTTAARTTCAGCGGGT-3' primers for these 11 strains. Sanger sequencing of the PCR products led to the identification of *B. catiminus*. Primers in the variable region of the Flavodoxin-like gene (Bathy03g02080; 5'-GCAAGAGAAGATTGAGGCGGAA-3' and 5'-CTCTGCTGCCGCTTTTGCCTCA-3') were designed for *B. catiminus* to select the most variable isolates for whole genome sequencing. Detection of *B. catiminus* in seawater samples was done by PCR amplification of environmental DNA with specific primers in the TOC1 ORF, TOCNI5B3 (5'-GGGACCCACCACAGGTTGCTGT-3') and TOCNI3B3 (5'-TACCGCGAGCAGCAACAGTAGT-3').

Since LOV-HK primers failed to amplify *Bathycoccus* strains from the Baffin Bay, the identification of *Bathycoccus sp.* strains was performed through specific amplification of a portion of TOC1 ORF (Bathy17g01510) with primer TOCNI5 (5'-AGGGGTTTTTGCAGAAACCGCT-3') and TOCNI3 (5'-TCTCGCATTTGATTTCGAGTCCA-3'). Intra-species diversity of Baffin bay strains was assessed using 2 markers, a fragment of the flavodoxin-like gene (Bathy03g02080) and the C-terminal region of the TIM gene (Bathy14g30100) (Devic et al., 2023), resulting in the identification of 10 multi loci genotypes **(Table S1)**.

Ultrastructure of *Bathycoccus* species

Cells were prepared according to Chrétionot-Dinet et al. (1995). Thin sections were stained with uranyl acetate and lead citrate, and observed with a 7500 Hitachi transmission electronic microscope.

Genome sequencing and assembly

Isolates from the Baffin bay were cultured at 4°C before DNA extraction, other strains were grown at 15°C. DNA extraction and Oxford Nanopore Technology (ONT) genome sequencing were performed on 28 *Bathycoccus sp.* strains as described in Devic et al. (2023). Raw ONT data were basecalled using Guppy 6.1.2 (https://nanoporetech.com) and the SUP model; reads with a PHRED score higher than 7 were retained for assembly. Quality control was performed with NanoPlot 1.19.0 (De Coster et al., 2018). Illumina pair-end sequencing (GENEWIZ & Novogene) was also performed on the same extracted DNA, yielding between 450 Mb and 2 Gb (~30-130X coverage) of sequences per strain. Paired-end sequencing of 250 bp long was used for strains from the RCC and the Banyuls bay, while 150 bp paired-end sequencing was used for strains isolated during the DarkEdge cruise.

Genome assemblies were produced using FLYE 2.9 (Kolmogorov et al., 2019) (options: --nano-hq --genome-size 15m). Assembly polishing with ONT reads was performed using MEDAKA 1.5 (https://github.com/nanoporetech/medaka) (options: -m r941_min_high_g360). Assembly correction with Illumina reads was performed using BWA 0.7.17 (Li, 2013) for read mapping, Samtools 1.9 (Danecek et al., 2021) and PILON 1.24 (Walker et al., 2014), with standard options. Scaffolding and direction of contigs on the reference genome (Strain RCC1105) was performed using RagTag 2.1.0 (Alonge et al., 2021). Unmapped contigs were designated as contaminations and excluded **(Figure S1, Appendix S1)**. Gap filling was performed using TGS-GapCloser 1.0.1 (Xu et al., 2020) with Samtools 1.9 and PILON 1.24. One round of Pilon correction was applied to corrected gaps for consistency with the global assembly correction (options: --p_round 1 --r_round 0 --min_nread 3). Abnormal chromosomal fusions due to assembly error were manually checked using D-GENIES 1.5.0 (Cabanettes and Klopp, 2018) and corrected using samtools (samtools faidx).

Assembly statistics were calculated with Assembly_stats 0.1.4 (https://github.com/MikeTrizna/assembly_stats) and QUAST 5.2.0 (Gurevich et al., 2013). Completion was assessed through BUSCO 5.4.4 (Manni et al., 2021) using the *chlorophyta_odb10* library, and quality was estimated with MERQURY 1.1 (Rhie et al., 2020) with the corresponding Illumina reads as input.

Genome Annotation

Genome assemblies were individually submitted to RepeatModeler 2.0.1 (https://github.com/Dfam-consortium/RepeatModeler) to build strain specific repeat libraries, which were subsequently concatenated in a non-redundant repeat library using CD-hit (Fu et al., 2012) (90% identity and coverage). This library was used with RepeatMasker 4.1.2

(https://github.com/rmhubley/RepeatMasker) on all samples to produce soft masked assemblies.

Available Illumina RNAseq whole transcriptome of *Bathycoccus prasinos* strain RCC4752 (SRX554258) (Keeling et al., 2014) was mapped upon the RCC4752 genome assembly using HISAT2 2.1.0 (Kim et al., 2015), and transcripts inferred using STRINGTIE 1.3.4 (Shumate et al., 2022). Plant orthologous proteins database 10 (Kriventseva et al., 2019) was used as protein evidences for annotation

Training of GENEMARK-ETP 4.71 (Brůna et al., 2023) and AUGUSTUS 3.3.3 (Stanke et al., 2008) *ab initio* prediction models were performed through BRAKER3 2.1.6 (Gotoh, 2008; Lomsadze et al., 2014; Buchfink et al., 2015; Gabriel et al., 2023), using transcriptome mapping data and the plant protein database as evidence. Training of SNAP (Korf, 2004) *ab initio* prediction model was performed after a first round of MAKER 2.31.9 (Cantarel et al., 2008; Campbell et al., 2014), using genome assembled transcripts, available *de novo* assembled transcripts (https://www.ncbi.nlm.nih.gov/Traces/wgs?val=HBMR01) and plant protein database as evidence. GENEMARK-ETP, AUGUSTUS and SNAP prediction models were merged through a second round of Maker to produce *de novo* structural annotation.

Phylogenetic analysis

BUSCO (Manni et al., 2021) analysis with *chlorophyta_odb10* library was computed for each assembled genome and on *O. tauri* and *M. pusilla* reference genomes to be used as outgroups. This resulted in 1,201 predicted genes shared across all genomes. Amino acid sequences were aligned for each gene using MAFFT 7.520 (options: --auto --maxiterate 1000) (Katoh and Standley, 2013) and converted to nucleic acid alignment using PAL2NAL 14 (Suyama et al., 2006) to ensure accurate coding sequence alignment.

Sequence alignments were concatenated in a partitioned supermatrix using catfasta2phyml.pl script (https://github.com/nylander/catfasta2phyml). Maximum likelihood phylogenetic inferences were computed through a partitioned analysis for multi-gene alignments using IQ-TREE 2.2.0.3 (Chernomor et al., 2016; Minh et al., 2020), with ModelFinder for automatic model finding (Kalyaanamoorthy et al., 2017) and 1,000 ultrafast bootstrap (Hoang et al., 2018) (options: -B 1000 -m MFP+MERGE).

Comparative genomic

Complete genomes dotplot comparison were computed using D-GENIES 1.5.0 (Cabanettes and Klopp, 2018) with Minimap2 2.24 (Li, 2018) as aligner and "Few repeats" option.

Genome assembly alignments were performed with MUMmer 3.1 (Kurtz et al., 2004) nucmer

(options: --maxmatch -c 500 -b 500 -l 200), alignments were then filtered for identity (<90%) and length (<100bp) using delta-filter (options: -m -i 90 -l 100). Synteny, single nucleotide polymorphism and structural variations were identified from the MUMmer output by SYRI 1.6 (Goel et al., 2019) and visualized with Plotsr 0.5.3 (Geol and Schneeberger, 2022).

Mapping of metagenomic datasets

Metagenomic reads from the TARA Ocean and TARA Polar Circle (De Vargas et al., 2015) campaigns were mapped to genome sequences using BWA mem 0.7.17 (Li, 2013). Samtools 1.13 (Danecek et al., 2021) was used to recover mapped reads (samtools view, options: -F 4) and to remove duplicates to avoid bias due to PCR artifacts. Using bamFilters (https://github.com/institut-de-genomique/bamFilters), mapped reads were filtered out for low-complexity bases (>75%), high-complexity bases (<30%), coverage (<80%) and identity (<95%). Remaining reads were retained for further analysis. Relative abundance was computed for each sample as the number of reads mapped normalized by the number of reads sequenced.

For *Bathycoccus sp.* species biogeography, complete assembled genomes of strains RCC4222, RCC716 and G8 were respectively used as reference for *B. prasinos*, *B. calidus* and *B. catiminus*. For big outlier chromosome haplotypes, Chr14:232230-630064 genome segment of strain RCC4222 and Chr14:231394-743201 genome segment of strain A8 were respectively used as reference for BOC A and BOC B outlier regions, according to loss of synteny.

5. Results

Biological resource

The biological resource used in this study includes 256 strains of *Bathycoccus sp*. from different geographic locations including the Mediterranean Sea, English Channel, North Sea, Arctic Ocean and Indian Ocean. These comprise seven strains of *B. prasinos* selected from the Roscoff Culture collection (RCC5417, RCC1613, RCC685, RCC1615, RCC1868, RCC4222, RCC4752) and initially sequenced using ONT sequencing to identify polymorphic indel markers (Devic et al., 2023). These markers were used to characterize the genetic diversity of 66 *Bathycoccus sp.* strains isolated in the bay of Banyuls-sur-mer (Mediterranean sea, France) during the 2018/2019 winter bloom, allowing the selection of seven representative *B. prasinos* isolates for sequencing (G11, C2, G2, E2, A8, B8, A1) (Devic et al., 2023). Similarly, seawater and ice samplings were conducted in three stations of the Baffin bay as part of the DarkEdge cruise in October 2021. After incubation of microbial communities (sizes < 1.2 µm) at 4°C or 15°C, picophytoplankton clones were obtained in semi-solid low-melting agarose. Genotyping using *B. prasinos* specific primers of the TOC1 gene revealed that all 183 isolates corresponded to *B. prasinos*. Further genotyping using two

polymorphic indel markers led to the identification of 10 multi-loci genotypes representative of the *B. prasinos* diversity in the Baffin bay. One isolate of each genotype was retained for sequencing (A818, B218, B518, C218, E318, H718, D119, H44, A727, A827) **(Table S1)**.

In total, we selected for sequencing 24 strains of *B. prasinos* with origins covering a latitudinal gradient from 40°N to 78°N **(Table 1)**. The only strain of the recently described species *Bathycoccus calidus* (RCC716) isolated from the Indian ocean was also included (Bachy et al., 2021). Finally, 11 isolates from the Banyuls bay showed weak amplifications of the LOV histidine-kinase marker. Sequencing of the 18S rDNA confirmed their identity as *Bathycoccus sp.* but variations in the ITS2 region revealed that they were phylogenetically distinct from *B. prasinos*, *B. calidus* and from *Bathycoccus* ITS2 environmental sequences recovered from the Kara Sea (Belevich et al., 2021) **(Figure S2)**. Three of these isolates were also selected for sequencing (C3, G5 and G8).

*G*enomic resources

The 28 *Bathycoccus sp.* strains, corresponding to 24 *B. prasinos*, one *B. calidus* and three yet undescribed *Bathycoccus sp*, display a wide geographical distribution ranging from arctic to equatorial regions and seasonal patterns in the Banyuls bay (Devic et al., 2023) **(Figure 1A, Table 1)**. These strains were sequenced using both ONT and Illumina to produce long and short high quality reads respectively, ensuring contiguity and low error rates in the final assemblies. Genome assembly was conducted using a pipeline outlined in **Figure 1B**. ONT reads of approximative depth ranging from 12 to 300X, assuming a genome size of 15Mb (Moreau et al., 2012), were used for a genome assembly using FLYE and a post-assembly correction using Medaka. Illumina reads (30 to 130X depth) were used for final assembly polishing using Pilon. Scaffolding and orientation of the created contigs upon the reference genome from the strain RCC1105 (Moreau et al., 2012) was performed through RagTag.

RCC4222, a clonal strain of the *B. prasinos* reference strain RCC1105 isolated in 2006 (Moreau et al., 2012), was sequenced in 2018 to check clonal conservation of the genome structures and to test the assembly pipeline. No major structural variations were identified compared to the original reference **(Figure 2A)**. The chromosome 19, a small outlier chromosome (SOC) described in *Bathycoccaceae* as an hypervariable structure with a high intraspecific diversity, was fully conserved between RCC4222 and RCC1105 (Moreau et al., 2012; Blanc-Mathieu et al., 2017). This indicates that the *B. prasinos* genome is stable in culture and validates our pipeline. Overall, reassembly of the reference using long reads improved continuity, as evidenced by the higher N50 (from 663 to 937 kb), higher genome size (from 15.04 to 15.19 Mb), and a reduction of contig number from 41 to 21, corresponding to the 19 nuclear chromosomes in addition to the the chloroplastic and mitochondrial genomes. **(Figure 2B)**. The BUSCO completion scores remained comparable between assemblies, confirming furthermore the high completion of the resequenced

reference genome **(Figure 2C)**.

De novo assembly of the 28 *Bathycoccus sp.* strains resulted in assembly statistics that were either superior to or on par with the reference genome **(Data S1)**. These assemblies showed an average of 24 contigs per assembly, with a mean N50 of 920 kb, including six assemblies without any identified gap (21 contigs). These high continuity assemblies have genome sizes ranging from 15 Mb to 16 Mb **(Figure 2B),** and a BUSCO completion score between 94.10% and 97.10%, with a median value of 96.80% **(Figure 2C)**.

A non-redundant repeat library, generated from all *Bathycoccus sp.* genomes with RepeatModeler and CD-hit, was used for repeat annotation *via* RepeatMasker, resulting in the annotation of ~10.36 ± 0.82% of each genome as repeated sequences **(Figure S3, Table S2)**. However, only a few of them could be associated with known structures, including LINEs, LTR elements and transposons, while the majority of classified sequences corresponded to simple repeats (~3.22 ± 0.18% of sequences) **(Table S2)**. A structural annotation of coding sequences was performed for each genome using available RNAseq data from *B. prasinos* strain RCC4752 and plant orthologous protein data to train three gene prediction models through the BRAKER3 and Maker pipelines (GENEMARK-ETP, Augustus and SNAP). The output of each model was then integrated using the Maker annotation pipeline into a complete structural annotation **(Figure 1B, Figure S3)**. The structural annotation predicted 7,478 (± 217) genes, with 7,253 annotated in the strain RCC4222 while 7,900 coding genes were initially annotated in the reference (Moreau et al., 2012). This discrepancy may be due to a lower ratio of mono-exonic genes (73% of predicted transcripts) in our annotation, resulting from the integration of genome assembled transcripts from RNAseq data **(Table S3)**.

Phylogenomic of the *Bathycoccus* genus

With this dataset we investigated the phylogenomics of the *Bathycoccus* genus. Maximum likelihood phylogenetic distances were inferred from nucleic acid alignments of genes shared between all strains. To avoid potential annotation biases between species caused by the predominance of *B. prasinos* transcriptomic data in the annotation model training, a set of 1,201 shared BUSCO genes from the *chlorophyta_odb10* database were used. The inferred phylogenetic tree showed clear separations between strains of *B. prasinos*, *B. calidus* and the three *Bathycoccus* genome sequences from the Banyuls bay. This putative *Bathycoccus sp.* showed an early divergence in the phylogenetic tree that could potentially correspond to a new species **(Figure 3A)**. Cell ultrastructure determined by electronic microscopy revealed characteristic features of the *Bathycoccus* genus with a single mitochondria, a single chloroplast containing a single starch granule, and scales at the cell surface that have been described earlier for *B. prasinos* (Moreau et al., 2012) and *B. calidus* (Bachy et al., 2021) **(Figure 3B).** We thus named this cryptic species as *Bathycoccus catiminus*.

The mapping of metagenomic data from TARA Ocean and TARA Polar Circle cruises were

conducted on a representative genome for each species. This analysis confirms the clear latitudinal and depth separations between *B. prasinos* and *B. calidus* populations, initially described by Vannier et al. (2016), with *B. prasinos* found in temperate regions at the surface, and *B. calidus* in tropical water at the deep chlorophyll maximum (DCM). *B. prasinos* was also abundant (up to 4% of sequences) in Arctic waters. *B. catiminus* was also detected in metagenomic data, and shows geographical distribution patterns similar to *B. prasinos*, however at much lower levels of abundance than *B. prasinos* and *B. calidus (up to 0,5% of* sequences). However, the comparison of mapped metagenomic reads indicates a high level of cross-mapping between *B. prasinos* and *B. catiminus*, with only ~9% of reads being specific to *B. catiminus* in all stations. This results in approximately 0.5% of horizontal coverage distributed along all chromosomes of *B. catiminus* and suggests that its abundance is ~10 times lower than initially computed **(Figure S4)**.

The presence of *B. catiminus* was also experimentally investigated through PCR in water samples from the Banyuls bay between January and March 2019. Species specific primers designed from sequences of the *B. catiminu*s TOC1 ORF were used. *B. catiminus* was detected in Banyuls bay water between January and March 2019, with a stronger amplification signal in February **(Figure 3C)**, in addition to its presence in December when the B. catiminus strains were isolated.

Phylogenetic diversity of *Bathycoccus prasinos*

The phylogenetic diversity of the *Bathycoccus prasinos* species was inferred from the 24 sequenced genomes using 1,201 conserved genes. The resulting phylogenetic tree revealed two main branches, separating the strain isolated from the North of Baffin bay during the DarkEdge cruise in October 2021 from all other strains, including the arctic strain RCC5417 isolated from the south of Baffin bay in 2016. However, this last strain diverges early from the cluster of temperate strains. Similarly, Mediterranean strains from the Banyuls bay isolated in 2018 and 2019 were more closely related to each other than to the RCC4222 strain isolated from the bay of Banyuls in 2006. Comparably, the strain RCC4752 isolated in 1986 in the Gulf of Naples did not cluster together with other Mediterranean strains. Geographical origin of strains from the English Channel and the North sea were not resolved in our phylogenetic tree **(Figure 4)**. Within geographic basins, subgroups were observed including 3 strains in the Banyuls bay (A1, G11 and A8) and 5 strains in the Baffin bay (C218, B218, A818, A827, A717). A second phylogenetic analysis was performed based on 69 shared genes located on chromosome 14, the BOC of *B. prasinos*. This analysis revealed that strains within the aforementioned subgroups clustered together, independently of their geographic origin. This distinguishes 2 haplotypes of the BOC that were named BOC A and BOC B **(Figure 4, Figure S5)**.

Big outlier chromosome haplotypes and polymorphism

The genomic diversity between BOC haplotypes of chromosome 14 was investigated by alignment of chromosome sequences between strains. The identification of syntenic regions confirms that the haplotype specific sequences are restricted to the outlier region, featured by a lower GC content. This chromosome segment corresponds to a ~400 (BOC A) to 500 kb (BOC B) non-syntenic region between the two haplotypes (i.e. ~2/3 of the chromosome length), with major inversion and duplication events, as well as other smaller structural rearrangements **(Figure 5A, Figure S6A)**. Since several strains were sequenced for each BOC haplotype (16 for BOC A, 8 for BOC B), single nucleotide polymorphism as well as structural variation density along chromosome 14 could be determined. Polymorphism density was computed on a 20 kb sliding window along chromosome 14, and compared to the genome-wide polymorphism to detect local divergences in sequence diversity. Lower than average polymorphism (local density/average density < 1) was detected in the outlier region of both identified haplotypes and, overall, lower polymorphism density of this region could be seen compared to the common regions located at both extremities of chromosome 14 **(Figure 5B, Figure S7)**. This clear pattern of localized reduced polymorphism could not be seen on any other non outlier chromosomes **(Figure S8)**.

The geographical distribution of both BOC haplotypes in the world Ocean was assessed by mapping of TARA Ocean and TARA polar circle metagenomic dataset on their respective outlier sequences. In all stations defined by relatively abundant *B. prasinos* sequences, the BOC A haplotype, corresponding to the haplotype of the reference strain RCC4222, was predominant, with a mapping ratio ranging from 68% to 97% of all mapped reads. This imbalance seemed less strict in polar circle stations, with the BOC B haplotype representing up to 32% of mapped reads **(Figure 5C)**. This difference in BOC ratios between temperate and arctic regions was further amplified in the haplotype ratio of isolated strains, with BOC A corresponding to the majority in strains isolated from the Banyuls bay (65.5% of strains), while BOC B was predominant in the Baffin bay (82% of strains) **(Figure S6B)**.

Genomic diversity of the small outlier chromosome

Another characteristic feature of *Mamiellophyceae* genomes is the small outlier chromosome (SOC), corresponding to chromosome 19 in *Bathycoccus* species. The SOC chromosome size varied considerably among the 24 assemblies of *B. prasinos*, particularly in arctic strains, with sizes ranging from 48 to 230 kb. With the exception of RCC1868 from the English channel, for which major conserved and syntenic regions were identified with strains A8, G2 and A818 (up to 190 kb conserved with A8), little or no synteny was detected between SOCs **(Figure 6, Figure S9)**. This held true even between strains originating from the same geographic basin or isolated at the same time in both Banyuls and Baffin bays. Surprisingly, despite the low synteny, conserved sequences with a wide range of sizes (15,787 ± 22,706 bp) were identified, indicating extensive reorganization within each genome **(Figure 6)**.

Higher chromosome 19 sequence conservation was observed between strains isolated

from the same geographic region, at the same time, such as for the Baffin bay in 2021, or the Banyuls bay in 2018-2019, up to the point that several assemblies had their chromosome almost fully composed of shared sequences **(Figure S10)**. However, chromosome 19 of Mediterranean strains isolated earlier (such as RCC4752 isolated in 1986 in the Gulf of Naples or RCC4222 isolated in 2006 in the Banyuls bay) had fewer conserved sequences with chromosome 19 of other Mediterranean strains than with strains from other geographical regions. Apart from the strain C2, which showed no chromosome 19 conservation with arctic assemblies, all genomes had at least one conserved region with another strain **(Figure S10)**. Remarkably, all sequences of the shortest chromosomes 19 (H718, B218 and A827 strains) were detected in other strains, these conserved sequences were usually concentrated in a distal region of chromosome 19, as seen in strains from the Banyuls bay and the Baffin bay **(Figure S9 & S10)**.

6. Discussion

Worldwide biological and genomic resources for the study of latitudinal and seasonal adaptation

In this study, we isolated 183 Arctic strains to add to the 66 Mediterranean strains isolated from the Banyuls bay (Devic et al., 2023) and the available strains of the Roscoff Culture Collection, further extending the current *Bathycoccus sp.* biological resource. All arctic and mediterranean strains were genotyped and the main haplotypes were fully sequenced including 10 strains from the Baffin bay and 10 strains from the Banyuls bay. This biological resource, covering a latitudinal gradient from pole to equator, provides an unprecedented tool to study the *Bathycoccus sp.* diversity and the physiological responses underlying adaptation to latitudinal gradients and seasons. Phylogenomic analysis revealed that 3 strains isolated from the Banyuls bay corresponded to a yet undescribed *Bathycoccus* species, morphologically indistinguishable from *B. prasinos* **(Figure 3)**. This new species was named *Bathycoccus catiminus* (from the French expression “en catimini" meaning “hidden”) in reference to its unexpected discovery and its very low abundance in metagenomic datasets compared to known *Bathycoccus* species. Concurrently, sequencing of ITS2 rDNA regions in the Kara sea suggests that a fourth species is yet to be isolated (Belevich et al., 2021) **(Figure S2)**, highlighting the underestimated taxonomic diversity of the *Bathycoccus* genus. As reviewed by Yung et al. (2022), the description of new species following genome sequencing is frequent in *Mamiellophyceae*, due to the lack of polymorphism in cellular ultrastructure and in the sequence of rRNA 18S V4 and V9 regions, commonly used as distinguishing characters (Piganeau et al., 2011). While the taxonomic diversity of *Bathycoccus sp.* may not yet be fully described, the distinct geographic distribution of the two most abundant *Bathycoccus* species, *B. prasinos* and *B. calidus* has been well documented using metagenomic approaches (Vannier et al., 2016; Leconte et al., 2020). Here, we confirmed the distinct geographic distribution between *B. prasinos* and *B. calidus*, presenting a complementary coverage along the whole latitudinal gradient. *B. calidus*

is more abundant in equatorial regions, while *B. prasinos* is predominant at higher latitudes, including Arctic regions **(Figure S4)**. The distinct relative abundances and distribution patterns of the three *Bathycoccus* species suggest fitness differences and specific mechanisms of adaptations to latitude and temperature within this genus.

The exploitation of natural diversity has greatly contributed to the characterization of genomic features and to the exploration of adaptive mechanisms not only in *Arabidopsis* but also in other wild land plants and crops cultivars (Alonso-Blanco et al., 2016; Gabur et al., 2019 for review). However, the intraspecific diversity of eukaryotic phytoplankton remains widely unexplored, with only a few studies conducted in *Ostreococcus tauri* (Blanc-Mathieu et al., 2017), *Phaeodactylum tricornutum* (Rastogi et al., 2020) and *Emiliania Huxleyi* (Read et al., 2013; Bendif et al., 2023). In all these studies, reference-assisted assembly approaches based on short reads mapping to a single reference genome were used. This allows for high resolution of single nucleotide polymorphism and small structural variations, but ignores larger structural variations and population-specific sequences. However, larger modifications also play an important role in environmental adaptation, since they may contain additional genes or regulatory sequences. The application of *de novo* assembly methods based on long-read sequencing in *Bathycoccus sp.* enables the identification of such large structural variations and the resolution of complex genomic structures specific to this taxonomic class **(Figure 5 & 6).** While the current state of algal genomic data is characterized by assembled genomes of heterogeneous quality and completeness, restricting intraspecific comparisons to few often incomplete genomes (Hanschen and Starkenburg, 2020), we aimed to produce a resource for the study of intraspecific genetic diversity within the *Bathycoccus* genus. Thus, we report here 28 *de novo* assembled and annotated genomes of *B. prasinos* (24) and of its 2 sister species. To our knowledge, this is the largest genomic resource available for a single species of green algae (Hanschen and Starkenburg, 2020; Shi et al., 2021). The strains and associated genomic data presented in our study, together with the large metagenomic dataset available, provide a substantial resource for the functional description and the understanding of the fundamental processes underlying the adaptation and ecological success of phytoplanktonic species.

Low polymorphism within BOC haplotypes indicates strong selective pressure

We used the *B. prasinos* resource consisting of 24 genomes together with the existing large metagenomics dataset to conduct an analysis of outlier chromosomes. The big outlier chromosome (BOC), discovered in *O. tauri*, is a putative sexual chromosome that has been identified in all other *Mamiellophyceae* genomes (Derelle et al., 2006; Grimsley et al., 2015). While two haplotypes of this BOC were identified in *O. tauri*, only one haplotype was previously reported for the *Bathycoccus* genus (Moreau et al., 2012; Blanc-Mathieu et al., 2017). In this study, the 2 BOC haplotypes were identified in *B. prasinos* strains from both the Banyuls and the Baffin bays, suggesting that these haplotypes are not the result of populations segregation due to geographic isolation, but rather a defining feature of *B.*

prasinos genome, as it was previously reported for other *Mamiellophyceae* (Grimsley et al., 2015).

The BOC of *B. prasinos*, corresponding to chromosome 14, has been characterized by several distinct features located in an outlier region, including lower GC content, high number of introns, and increased expression levels (Derelle et al., 2006; Moreau et al., 2012). In *O. tauri*, previous reports indicated a higher linkage disequilibrium and reduced recombination events in the outlier region, leading to an accumulation of transposable elements and suggesting a putative mating function of this region (Blanc-Mathieu et al., 2017), thus following the model of the mating chromosome of *Chlamydomonas reinhardtii* (Ferris and Goodenough, 1994). However, no accumulation of transposable elements was detected in the *B. prasinos* reference genome and no loss of synteny was identified in strains of the same BOC haplotype **(Figure S6A)** (Moreau et al. 2012). Instead, a lower density of single nucleotide polymorphism and structural variations was observed in the outlier region of both *B. prasinos* BOC haplotypes, even between geographically distant strains **(Figure 5B, Figure S7)**. Lower than average polymorphism has also been reported in *O. tauri* outlier region and *C. reinhardtii* mating chromosome (Blanc-Mathieu et al., 2017; Hasan et al., 2019), in apparent contradiction with recombination suppression that would intensify mutation accumulation in sexual chromosomes. Gene conversion between mating types was proposed as a sequence homogenizing mechanism by De Hoff et al. (2013), but this hypothesis proved difficult to test with *O. tauri* genomic data due to the representation of one of the putative mating types by just a single strain. In *B. prasinos*, our analysis of shared genes between BOC haplotypes revealed a phylogenetic separation, suggesting a divergent evolution of chromosome 14 shared genes between haplotypes **(Figure S5)**. Strong mating-type-specific selective pressure linked to background selection, as suggested by Hasan et al. (2019, 2020), might be responsible for the reduced diversity and strong differentiation observed in this complex genomic context. Furthermore, this pressure may contribute to the preservation of increased gene expression and fragmented gene structure in the outlier region of the BOC. Although the fitness impact of this 400 to 500 kb haplotype-specific region still needs to be assessed, it is possible that the regional imbalances observed in metagenomic dataset in *B. prasinos* **(Figure 5C)** and *Ostreococcus lucimarinus* (Leconte et al., 2020) are attributed to either differential fitness, or seasonal dynamics among BOC haplotypes, or both. The cosmopolitan distribution of *B. prasinos*, along with its abundance in metagenomic dataset and the important biological resource available, offer a unique opportunity to monitor the spatio-temporal population dynamics of BOC haplotypes and to further study the role of BOCs in an environmental context (Vannier et al., 2016; Lambert et al., 2019).

SOC variable sizes and sequence conservation suggest a bipartite chromosome structure

The small outlier chromosome (SOC) is a hypervariable chromosomal structure common to *Mamiellophyceae* genomes, and characterized in *O. tauri* by a low synteny, a

high sequence diversity between strains due to extensive structural rearrangements and an important variation in chromosome size (Derelle et al., 2006, Blanc-Mathieu et al., 2017). In this study, we report SOC conserved and rearranged sequences between 24 *B. prasinos* strains. This genomic resource encompasses broad geographical and temporal scales, revealing significant sequence conservation between strains from separate oceanic basins, despite the overall lack of synteny. While several strains come from the same water samples in the Baffin and Mediterranean sea, others were isolated at different times over a period spanning from 1986 to 2020, offering the opportunity to study both the diversity and dynamics of SOC structure at different spatial and temporal scales.

Important size variations for the SOC were displayed in our dataset, especially between Arctic strains isolated during the DarkEdge cruise for which we identified the smallest SOCs **(Figure 6)**. Structural and transcriptional modifications of the SOC were shown by Yau et al. (2016, 2018) to be induced by prasinovirus infection and linked to increased spectrum of resistance, suggesting a role of the SOC in viral resistance mechanisms in *Mamiellophyceae*. Moreover, switches between resistant and susceptible phenotypes within isogenic cultures were associated with SOC rearrangements in *O. tauri* and *O. mediterraneus*, and proposed as a coexistence mechanism in microalgae-virus interactions (Yau et al., 2020). As shown by Blanc-Mathieu et al. (2017), smaller SOC sizes are associated with decreased resistance spectrum. Thus, the size variations might reflect the immune history of the corresponding strains, selecting a more compact chromosome 19 for better local fitness and specific resistance at the expense of broader resistance range. However, these could also be the sign of continuous rearrangement at the time of sampling due to ongoing viral infection.

The smallest SOCs may also provide valuable insight into the structure of *B. prasinos* chromosome 19, as the sequences found in these reduced chromosomes tend to be conserved in nearly all strains and located at a distal position. This separates the SOC into a more conserved portion and a more variable one, with the smallest chromosomes potentially representing a minimal chromosomal content. This observation aligns with the bipartite pattern of SOC transcription reported for *O. tauri* by Yau et al. (2016), in which one half of the chromosome is expressed in susceptible strains while the other half is expressed by resistant strains. Therefore suggesting a common regulation of both transcription and structural rearrangement in this region to minimize loss of fitness in favor of genomic diversity for rapid genomic adaptation. Yau et al. (2016) proposed several theories to explain these observations, including the activation of retrotransposons and epigenetic modifications in response to biotic stress. To comprehensively test these theories, an in-depth characterization of SOC structural variations, including gene-content and structure, is necessary. As such, the extensive dataset described here represents a valuable resource for understanding the intricate interactions between *Mamiellophyceae* and associated viruses.

Conclusion

The biological and associated genomic resources of *Bathycoccus* presented in this study cover an unprecedented range of latitudes for a eukaryotic phytoplankton species. Genotyping and high quality *de novo* whole-genome sequencing of newly isolated strains have led to the identification of a third cryptic species named *Bathycoccus catiminus* which, unlike other *Bathycoccus* species, is present at very low levels in metagenomic datasets. Comparative analysis focused on *B. prasinos* strains revealed a previously undescribed BOC haplotype and highlighted the low polymorphism of the outlier region in this putative sexual chromosome. Meanwhile, further description of the SOC hypervariability uncovered significant sequence conservation and unveiled a bipartite structure of the SOC.

The *Bathycoccus* genomic resources paves the way for multi-scale comparative analyses at the genus level and for the construction of a first species-specific pangenome for *B. prasinos.* This genomic resource, combined with metagenomic studies, physiological analysis of Bathycoccus natural variants, and recently developed genetic engineering tools for *B. prasinos* (Faktorova et al., 2020), offer new opportunities to study the molecular basis of adaptation underlying the ecological success of *B. prasinos.*

7. Accession numbers

8. Acknowledgements

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/ - http://www.southgreen.fr. We thank Marie-Hélène Forget, as DarEdge coordinator, as well as all the other participants of this expedition and the members of the Amundsen Vessel for their help in collecting water samples during the DarkEdge Cruise. We thank the Roscoff Culture Collection for providing access to collected strains. We thank Marie-Line Escande and the platform BioPIC at the Oceanological Observatory of Banyuls-sur-Mer for electron microscopy service. This work and the salaries of LD and JR were funded by the french Agence Nationale de la recherche (ANR Clima-Clock, ANR-20-CE20-0024 to FYB, AF and OJ).

9. Author contributions

MD, FS, FYB and LD designed the study. MD, AF and the Dark Edge genomics sampling team designed and carried out the sampling. MD, JCL, VV and CM isolated, genotyped and produced the sequencing data. LD, JR and FS performed bioinformatic analyses. LD, MD, FS and FYB wrote the manuscript. MD and OJ provided critical revisions of the manuscript.

10. References

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., Soyk, S. (2021) Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. https://doi.org/10.1101/2021.11.18.469135

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M., Cao, J., Chae, E., Dezwaan, T.M., Ding, W., Ecker, J.R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D.G., Hancock, A.M., Henz, S.R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R.A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T.P., Mott, R., Muliyati, N.W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P.Yu., Picó, F.X., Platzer, A., Rabanal, F.A., Rodriguez, A., Rowan, B.A., Salomé, P.A., Schmid, K.J., Schmitz, R.J., Seren, Ü., Sperone, F.G., Sudkamp, M., Svardal, H., Tanzer, M.M., Todd, D., Volchenboum, S.L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., Zhou, X. (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166, 481–491. https://doi.org/10.1016/j.cell.2016.05.063

Bachy, C., Yung, C.C.M., Needham, D.M., Gazitúa, M.C., Roux, S., Limardo, A.J., Choi, C.J., Jorgens, D.M., Sullivan, M.B., Worden, A.Z. (2021) Viruses infecting a warm water picoeukaryote shed light on spatial co-occurrence dynamics of marine viruses and their hosts. *ISME J*, 15, 3129–3147. https://doi.org/10.1038/s41396-021-00989-9

Belevich, T.A., Milyutina, I.A., Abyzova, G.A., Troitsky, A.V. (2021) The pico-sized Mamiellophyceae and a novel *Bathycoccus* clade from the summer plankton of Russian Arctic Seas and adjacent waters. *FEMS Microbiology Ecology*, 97, fiaa251. https://doi.org/10.1093/femsec/fiaa251

Bendif, E.M., Probert, I., Archontikis, O.A., Young, J.R., Beaufort, L., Rickaby, R.E., Filatov, D. (2023) Rapid diversification underlying the global dominance of a cosmopolitan phytoplankton. *ISME J*, 17, 630–640. https://doi.org/10.1038/s41396-023-01365-5

Benites, L.F., Poulton, N.J., Labadie, K., Sieracki, M.E., Grimsley, N., Piganeau, G. (2019) Single cell ecogenomics reveals mating types of individual cells and ssDNA viral infections in the smallest photosynthetic eukaryotes. *Philosophical Transactions of the Royal Society B*, 374, 20190089. https://doi.org/10.1098/rstb.2019.0089

Blanc-Mathieu, R., Krasovec, M., Hebrard, M., Yau, S., Desgranges, E., Martin, J., Schackwitz, W., Kuo, A., Salin, G., Donnadieu, C., Desdevises, Y., Sanchez-Ferandin, S., Moreau, H., Rivals, E., Grigoriev, I.V., Grimsley, N., Eyre-Walker, A., Piganeau, G. (2017) Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci. Adv.*, 3, e1700239. https://doi.org/10.1126/sciadv.1700239

Botebol, H., Lesuisse, E., Šuták, R., Six, C., Lozano, J.-C., Schatt, P., Vergé, V., Kirilovsky, A., Morrissey, J., Léger, T., Camadro, J.-M., Gueneugues, A., Bowler, C., Blain, S., Bouget, F.-Y. (2015) Central role for ferritin in the day/night regulation of iron homeostasis in marine phytoplankton. *Proc. Natl. Acad. Sci. U.S.A.*, 112, 14652–14657. https://doi.org/10.1073/pnas.1506074112

Brůna, T., Lomsadze, A., Borodovsky, M. (2023) GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data. https://doi.org/10.1101/2023.01.13.524024

Buchfink, B., Xie, C., Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12, 59–60. https://doi.org/10.1038/nmeth.3176

Cabanettes, F., Klopp, C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, 6, e4958. https://doi.org/10.7717/peerj.4958

Campbell, M.S., Holt, C., Moore, B., Yandell, M. (2014) Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, 39. https://doi.org/10.1002/0471250953.bi0411s48.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., Yandell, M. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, 18, 188–196. https://doi.org/10.1101/gr.6743907

Chernomor, O., von Haeseler, A., Minh, B.Q. (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol*, 65, 997–1008. https://doi.org/10.1093/sysbio/syw037

Chrétiennot-Dinet, M.-J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J., Machado, M.C. (1995) A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia*, 34, 285–292. https://doi.org/10.2216/i0031-8884-34-4-285.1

Corellou, F., Schwartz, C., Motta, J.-P., Djouani-Tahri, E.B., Sanchez, F., Bouget, F.-Y. (2009) Clocks in the Green Lineage: Comparative Functional Analysis of the Circadian Architecture of the Picoeukaryote *Ostreococcus. The Plant Cell*, 21, 3436–3449. https://doi.org/10.1105/tpc.109.068825

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H. (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10, giab008. https://doi.org/10.1093/gigascience/giab008

De Barros Dantas, L.L., Eldridge, B.M., Dorling, J., Dekeya, R., Lynch, D.A., Dodd, A.N. (2023) Circadian regulation of metabolism across photosynthetic organisms. *The Plant Journal*, 16405. https://doi.org/10.1111/tpj.16405

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., Van Broeckhoven, C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

De Hoff, P.L., Ferris, P.J., Olson, B.J.S.C., Miyagi, A., Sa Geng, Umen, J.G. (2013)

Species and population level molecular profiling reveals cryptic recombination and emergent asymmetry in the dimorphic mating locus of C. reinhardtii. *PLOS Genetics*, 9, e1003724. https://doi.org/10.1371/journal.pgen.1003724

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E.G., Sardet, C., Sullivan, M.B., Velayoudon, D. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348, 1261605. https://doi.org/10.1126/science.1261605

Delmont, T.O., Gaia, M., Hinsinger, D.D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A.M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., de Vargas, C., Bowler, C., Karsenti, E., Pelletier, E., Wincker, P., Jaillon, O., Sunagawa, S., Acinas, S.G., Bork, P., Karsenti, E., Bowler, C., Sardet, C., Stemmann, L., de Vargas, C., Wincker, P., Lescot, M., Babin, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Jaillon, O., Kandels, S., Iudicone, D., Ogata, H., Pesant, S., Sullivan, M.B., Not, F., Lee, K.-B., Boss, E., Cochrane, G., Follows, M., Poulton, N., Raes, J., Sieracki, M., Speich, S. (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, 2, 100123. https://doi.org/10.1016/j.xgen.2022.100123

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J*, 5, 1095–1107. https://doi.org/10.1038/ismej.2010.209

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piégu, B., Ball, S.G., Ral, J.-P., Bouget, F.-Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., Moreau, H. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U.S.A.*, 103, 11647–11652. https://doi.org/10.1073/pnas.0604795103

Devic, M., Mariac, C., Vergé, V., Schatt, P., Dennu, L., Lozano, J.-C., Bouget, F.-Y., Sabot, F. (2023) Population dynamics of the cosmopolitan eukaryotic picophytoplankton Bathycoccus during seasonal blooms in the bay of Banyuls sur Mer (North Western Mediterranean sea). https://doi.org/10.1101/2023.02.09.527951

Eikrem, W., Throndsen, J. (1990) The ultrastructure of Bathycoccus gen. nov. and B. prasinos sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia*, 29, 344–350. https://doi.org/10.2216/i0031-8884-

29-3-344.1

Faktorová, D., Nisbet, R.E.R., Fernández Robledo, J.A., Casacuberta, E., Sudek, L., Allen, A.E., Ares, M., Aresté, C., Balestreri, C., Barbrook, A.C., Beardslee, P., Bender, S., Booth, D.S., Bouget, F.-Y., Bowler, C., Breglia, S.A., Brownlee, C., Burger, G., Cerutti, H., Cesaroni, R., Chiurillo, M.A., Clemente, T., Coles, D.B., Collier, J.L., Cooney, E.C., Coyne, K., Docampo, R., Dupont, C.L., Edgcomb, V., Einarsson, E., Elustondo, P.A., Federici, F., Freire-Beneitez, V., Freyria, N.J., Fukuda, K., García, P.A., Girguis, P.R., Gomaa, F., Gornik, S.G., Guo, J., Hampl, V., Hanawa, Y., Haro-Contreras, E.R., Hehenberger, E., Highfield, A., Hirakawa, Y., Hopes, A., Howe, C.J., Hu, I., Ibañez, J., Irwin, N.A.T., Ishii, Y., Janowicz, N.E., Jones, A.C., Kachale, A., Fujimura-Kamada, K., Kaur, B., Kaye, J.Z., Kazana, E., Keeling, P.J., King, N., Klobutcher, L.A., Lander, N., Lassadi, I., Li, Z., Lin, S., Lozano, J.-C., Luan, F., Maruyama, S., Matute, T., Miceli, C., Minagawa, J., Moosburner, M., Najle, S.R., Nanjappa, D., Nimmo, I.C., Noble, L., Novák Vanclová, A.M.G., Nowacki, M., Nuñez, I., Pain, A., Piersanti, A., Pucciarelli, S., Pyrih, J., Rest, J.S., Rius, M., Robertson, D., Ruaud, A., Ruiz-Trillo, I., Sigg, M.A., Silver, P.A., Slamovits, C.H., Jason Smith, G., Sprecher, B.N., Stern, R., Swart, E.C., Tsaousis, A.D., Tsypin, L., Turkewitz, A., Turnšek, J., Valach, M., Vergé, V., Von Dassow, P., Von Der Haar, T., Waller, R.F., Wang, L., Wen, X., Wheeler, G., Woods, A., Zhang, H., Mock, T., Worden, A.Z., Lukeš, J. (2020) Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nat Methods*, 17, 481–494. https://doi.org/10.1038/s41592-020-0796-x

Ferris, J., Goodenough, W. (1994) The Mating-Type Locus of Chlamydomonas reinhardtii Contains Highly Rearranged DNA Sequences. *Cell*, 76, 1135–1145. https://doi.org/10.1016/0092-8674(94)90389-1

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gabriel, L., Brůna, T., Hoff, K.J., Ebel, M., Lomsadze, A., Borodovsky, M., Stanke, M. (2023) BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. https://doi.org/10.1101/2023.06.10.544449

Gabur, I., Chawla, H.S., Snowdon, R.J., Parkin, I.A.P. (2019) Connecting genome structural variation with complex traits in crop plants. *Theor Appl Genet*, 132, 733–750. https://doi.org/10.1007/s00122-018-3233-0

Goel, M., Schneeberger, K. (2022) plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, 38, 2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Goel, M., Sun, H., Jiao, W.-B., Schneeberger, K. (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*, 20, 277. https://doi.org/10.1186/s13059-019-1911-0

Gotoh, O. (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Research*, 36, 2630–2638. https://doi.org/10.1093/nar/gkn105

Grimsley, N., Yau, S., Piganeau, G., Moreau, H. (2015) Typical Features of Genomes in the Mamiellophyceae. *Marine Protists*, 107–127. https://doi.org/10.1007/978-4-431-55130-0_6

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hanschen, E.R., Starkenburg, S.R. (2020) The state of algal genome quality and diversity. *Algal Research*, 50, 101968. https://doi.org/10.1016/j.algal.2020.101968

Hasan, A.R., Duggal, J.K., Ness, R.W. (2019) Consequences of recombination for the evolution of the mating type locus in *Chlamydomonas reinhardtii*. *New Phytol*, 224, 1339–1348. https://doi.org/10.1111/nph.16003

Hasan, A.R., Ness, R.W. (2020) Recombination Rate Variation and Infrequent Sex Influence Genetic Diversity in Chlamydomonas reinhardtii. *Genome Biology and Evolution*, 12, 370–380. https://doi.org/10.1093/gbe/evaa057

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, 35, 518–522. https://doi.org/10.1093/molbev/msx281

Joli, N., Monier, A., Logares, R., Lovejoy, C. (2017) Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome. *ISME J*, 11, 1372–1385. https://doi.org/10.1038/ismej.2017.7

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*, 14, 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30, 772–780. https://doi.org/10.1093/molbev/mst010

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K.J., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B.D., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R.P., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D.H., Marchetti, A., McManus, G.B., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A.,

Murray, S.A., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T.A., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E.C., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., William H. Wilson, Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, 12, 1–6. https://doi.org/10.1371/journal.pbio.1001889

Kim, D., Langmead, B., Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12, 357–360. https://doi.org/10.1038/nmeth.3317

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, 5, 59. https://doi.org/10.1186/1471-2105-5-59

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biology*, 5, R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., Galand, P.E. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J*, 13, 388–401. https://doi.org/10.1038/s41396-018-0281-z

Leconte, J., Benites, L.F., Vannier, T., Wincker, P., Piganeau, G., Jaillon, O. (2020) Genome Resolved Biogeography of Mamiellales. *Genes*, 11, 66. https://doi.org/10.3390/genes11010066

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., De Clerck, O. (2012) Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, 31, 1–46. https://doi.org/10.1080/07352689.2011.615705

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997

Li, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnol. Oceanogr.*, 39, 169–175. https://doi.org/10.4319/lo.1994.39.1.0169

Lomsadze, A., Burns, P.D., Borodovsky, M. (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, 42, e119. https://doi.org/10.1093/nar/gku557

Lozano, J.-C., Schatt, P., Botebol, H., Vergé, V., Lesuisse, E., Blain, S., Carré, I.A., Bouget, F.-Y. (2014) Efficient gene targeting and removal of foreign DNA by homologous recombination in the picoeukaryote *Ostreococcus*. *Plant J*, 78, 1073–1083. https://doi.org/10.1111/tpj.12530

Lynch, M., Gabriel, W., Wood, A.M. (1991) Adaptive and demographic responses of plankton populations to environmental change. *Limnol. Oceanogr.*, 36, 1301–1312. https://doi.org/10.4319/lo.1991.36.7.1301

Manni, M., Berkeley, M.R., Seppey, M., Zdobnov, E.M. (2021) BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, 1, e323. https://doi.org/10.1002/cpz1.323

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M.F., Piganeau, G., Rouzé, P., Da Silva, C., Wincker, P., Van de Peer, Y., Vandepoele, K. (2012) Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol*, 13, R74. https://doi.org/10.1186/gb-2012-13-8-r74

Moulager, M., Corellou, F., Vergé, V., Escande, M.-L., Bouget, F.-Y. (2010) Integration of Light Signals by the Retinoblastoma Pathway in the Control of S Phase Entry in the Picophytoplanktonic Cell Ostreococcus. *PLoS Genet*, 6, e1000957. https://doi.org/10.1371/journal.pgen.1000957

O'Neill, J.S., Van Ooijen, G., Dixon, L.E., Troein, C., Corellou, F., Bouget, F.-Y., Reddy, A.B., Millar, A.J. (2011) Circadian rhythms persist without transcription in a eukaryote. *Nature*, 469, 554–558. https://doi.org/10.1038/nature09654

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., Schacherer, J. (2018) Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, 556, 339–344. https://doi.org/10.1038/s41586-018-0030-5

Piganeau, G., Eyre-Walker, A., Grimsley, N., Moreau, H. (2011) How and Why DNA

Barcodes Underestimate the Diversity of Microbial Eukaryotes. *PLoS ONE*, 6, e16342. https://doi.org/10.1371/journal.pone.0016342

Rastogi, A., Vieira, F.R.J., Deton-Cabanillas, A.-F., Veluchamy, A., Cantrel, C., Wang, G., Vanormelingen, P., Bowler, C., Piganeau, G., Hu, H., Tirichine, L. (2020) A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom Phaeodactylum tricornutum. *ISME J*, 14, 347–363. https://doi.org/10.1038/s41396-019-0528-3

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J.P., Monier, A., Salamov, A., Young, J.R., Aguilar, M., Claverie, J.-M., Frickenhaus, S., Gonzalez, K., Herman, E.K., Lin, Y.-C., Napier, J.A., Ogata, H., Sarno, A.F., Shmutz, J., Schroeder, D.C., de Vargas, C., Frédéric Verret, Verret, F., von Dassow, P., Valentin, K., Van de Peer, Y., Wheeler, G.L., Dacks, J.B., Delwiche, C.F., Dyhrman, S.T., Glöckner, G., John, U., Richards, T.A., Worden, A.Z., Zhang, X., Grigoriev, I.V., Allen, A.E., Bidle, K.D., Borodovsky, M., Bowler, C., Brownlee, C., Cock, J.M., Eliáš, M., Gladyshev, V.N., Groth, M., Guda, C., Hadaegh, A.R., Iglesias-Rodriguez, M.D., Jenkins, J., Jones, B.M., Lawson, T., Leese, F., Lindquist, E., Lobanov, A., Lomsadze, A., Malik, S.-B., Marsh, M.E., Mackinder, L.C.M., Mock, T., Mueller-Roeber, B., Pagarete, A., Parker, M.S., Probert, I., Ian Probert, Quesneville, H., Raines, C.A., Rensing, S.A., Riaño-Pachón, D.M., Richier, S., Sophie Richier, Rokitta, S.D., Yoshihiro Shiraiwa, Shiraiwa, Y., Soanes, D.M., van der Giezen, M., Wahlund, T.M., Williams, B.A.P., Wilson, W., Wolfe, G., Wurch, L.L. (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature*, 499, 209–213. https://doi.org/10.1038/nature12221

Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*, 21, 245. https://doi.org/10.1186/s13059-020-02134-9

Shi, C., Liu, X., Han, K., Peng, L., Li, L., Ge, Q., Fan, G. (2021) A database and comprehensive analysis of the algae genomes. https://doi.org/10.1101/2021.10.30.466624

Shumate, A., Wong, B., Pertea, G., Pertea, M. (2022) Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol*, 18, e1009730. https://doi.org/10.1371/journal.pcbi.1009730

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, 24, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Suyama, M., Torrents, D., Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34, W609–W612. https://doi.org/10.1093/nar/gkl315

The 3,000 rice genomes project (2014) The 3,000 rice genomes project. *GigaSci*, 3, 7. https://doi.org/10.1186/2047-217X-3-7

The 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., De Jong, H., De Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., Van Leeuwen, H., Van Ham, R.C.H.J., Fito, L., Guignier, L., Sevilla, M., Ellul, P., Ganko, E., Kapur, A., Reclus, E., De Geus, B., Van De Geest, H., Te Lintel Hekkert, B., Van Haarst, J., Smits, L., Koops, A., Sanchez-Perez, G., Van Heusden, A.W., Visser, R., Quan, Z., Min, J., Liao, L., Wang, X., Wang, G., Yue, Z., Yang, X., Xu, N., Schranz, E., Smets, E., Vos, R., Rauwerda, J., Ursem, R., Schuit, C., Kerns, M., Van Den Berg, J., Vriezen, W., Janssen, A., Datema, E., Jahrman, T., Moquet, F., Bonnet, J., Peters, S. (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J*, 80, 136–148. https://doi.org/10.1111/tpj.12616

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.-M., de Vargas, C., Sieracki, M.E., Iudicone, D., Vaulot, D., Wincker, P., Jaillon, O. (2016) Survey of the green picoalga Bathycoccus genomes in the global ocean. *Scientific Reports*, 6, 37900–37900. https://doi.org/10.1038/srep37900

Vaulot, D., Lepère, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., Moreau, H., Vandepoele, K., Ulloa, O., Gavory, F., Piganeau, G. (2012) Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLoS ONE*, 7, e39648. https://doi.org/10.1371/journal.pone.0039648

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Worden, A.Z., Nolan, J.K., Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol. Oceanogr.*, 49, 168–179. https://doi.org/10.4319/lo.2004.49.1.0168

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., Zhang, Y. (2020) TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, 9, giaa094. https://doi.org/10.1093/gigascience/giaa094

Yau, S., Caravello, G., Fonvieille, N., Desgranges, É., Moreau, H., Moreau, H., Grimsley, N. (2018) Rapidity of Genomic Adaptations to Prasinovirus Infection in a Marine Microalga. *Viruses*, 10, 441. https://doi.org/10.3390/v10080441

Yau, S., Hemon, C., Derelle, E., Moreau, H., Piganeau, G., Grimsley, N. (2016) A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri. *PLoS Pathog*, 12, e1005965. https://doi.org/10.1371/journal.ppat.1005965

Yau, S., Krasovec, M., Benites, L.F., Rombauts, S., Groussin, M., Vancaester, E., Aury, J.-M., Derelle, E., Desdevises, Y., Escande, M.-L., Grimsley, N., Guy, J., Moreau, H., Sanchez-Brosseau, S., Van de Peer, Y., Vandepoele, K., Gourbiere, S., Piganeau, G. (2020) Virus-host coexistence in phytoplankton through the genomic lens. *Sci. Adv.*, 6,

eaay2587. https://doi.org/10.1126/sciadv.aay2587

Yung, C.C.M., Elvira Rey Redondo, Frederic Sanchez, Sheree, Y., Piganeau, G. (2022) Diversity and Evolution of Mamiellophyceae: Early-Diverging Phytoplanktonic Green Algae Containing Many Cosmopolitan Species. *Journal of Marine Science and Engineering*, 10, 240–240. https://doi.org/10.3390/jmse10020240

11. Tables and figures

Table 1 : Metadata of sequenced strains.

	Species	Latitude (North)	Longitude (East)	Depth (m)	Sampling site	Sampling date	Sample source
ARCTIC OCEAN							
A818	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
B218	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
B518	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
C218	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
E318	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
H718	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
D119	*B. prasinos*	78.12	-74.36	0	Baffin bay DE310 †	16/10/2021	DarkEdge
H44	*B. prasinos*	76.03	-77.23	0	Baffin bay DE110	12/10/2021	DarkEdge
A727	*B. prasinos*	75.95	-85.58	20	Baffin bay DE410	21/10/2021	DarkEdge
A827	*B. prasinos*	75.95	-85.58	20	Baffin bay DE410	21/10/2021	DarkEdge
RCC5417	*B. prasinos*	67.48	-63.78	0	Baffin Island	01/06/2016	RCC ‡
ATLANTIC OCEAN							
RCC1613	*B. prasinos*	57.57	8.67	35	North Sea	26/07/2007	RCC ‡
RCC685	*B. prasinos*	54.18	45,176	0	North Sea	16/05/2001	RCC ‡
RCC1615	*B. prasinos*	50.2	0.32	10	English Channel	08/07/2007	RCC ‡
RCC1868	*B. prasinos*	48.75	-3.95	0	English Channel	05/02/2009	RCC ‡
MEDITERRANEAN SEA							
RCC4222	*B. prasinos*	42.48	3.53	3	Banyuls bay	01/01/2006	RCC ‡
G11	*B. prasinos*	42.48	3.53	3	Banyuls bay	03/12/2018	OOB §
C2	*B. prasinos*	42.48	3.53	3	Banyuls bay	28/01/2019	OOB §
G2	*B. prasinos*	42.48	3.53	3	Banyuls bay	11/02/2019	OOB §
E2	*B. prasinos*	42.48	3.53	3	Banyuls bay	25/02/2019	OOB §
A8	*B. prasinos*	42.48	3.53	3	Banyuls bay	26/02/2019	OOB §
B8	*B. prasinos*	42.48	3.53	3	Banyuls bay	26/02/2019	OOB §
A1	*B. prasinos*	42.48	3.53	3	Banyuls bay	19/03/2019	OOB §
G8	*Bathycoccus sp.*	42.48	3.53	3	Banyuls bay	03/12/2018	OOB §
C3	*Bathycoccus sp.*	42.48	3.53	3	Banyuls bay	12/12/2018	OOB §
G5	*Bathycoccus sp.*	42.48	3.53	3	Banyuls bay	12/12/2018	OOB §
RCC4752	*B. prasinos*	40.48	14.14	100	Gulf of Naples	17/04/1986	RCC ‡
INDIAN OCEAN							
RCC716	*B. calidus*	-14.48	113.45	70	Indian Ocean	11/06/2003	RCC ‡

† Ice station
‡ Roscoff Culture Collection
§ Oceanological Observatory of Banyuls-sur-Mer

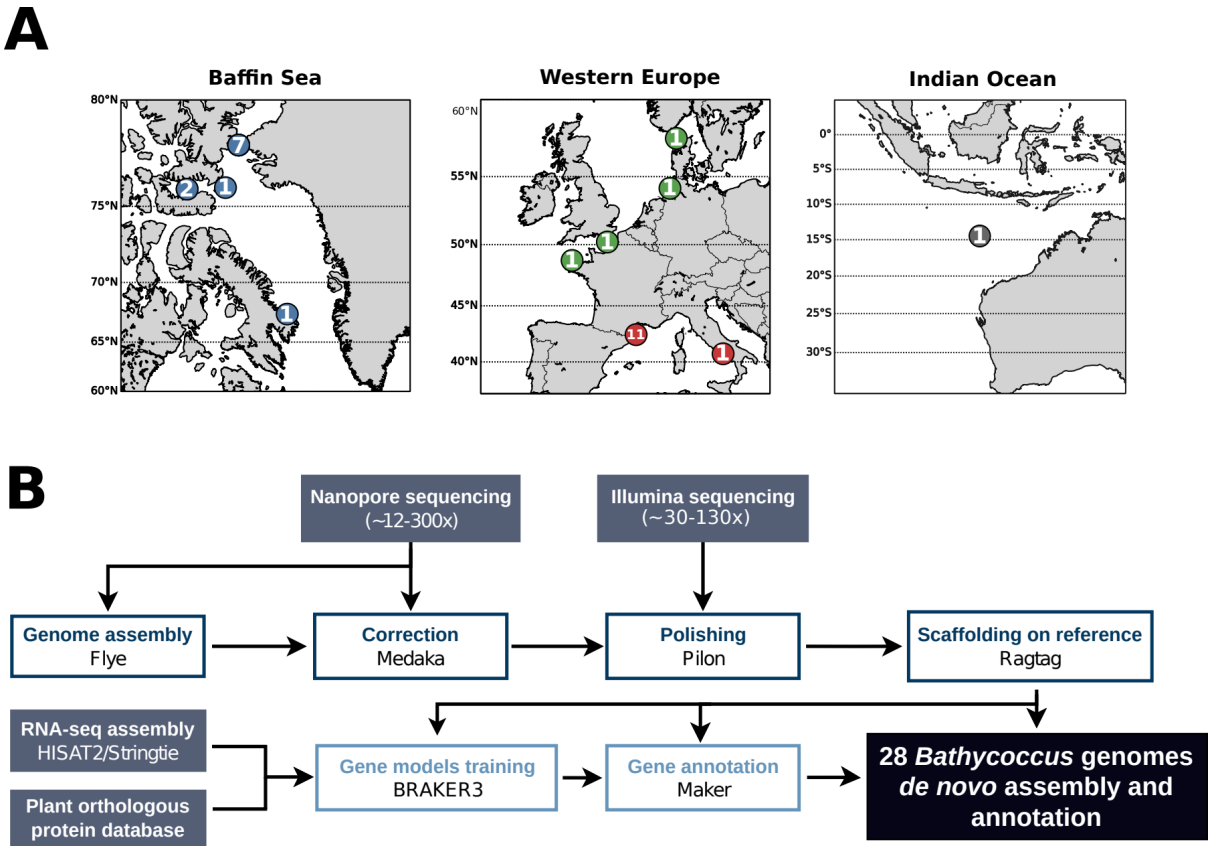

Fig 1 : Sampling, sequencing, genome assembly and annotation strategy of *Bathycoccus*. (A) Geographical distribution of the sampling sites for the selected *Bathycoccus* strains. The number of strains sampled at each site is indicated. Colors correspond to the oceanic basin of origin: blue, Arctic Ocean; green, Atlantic Ocean; red, Mediterranean Sea; grey, Indian Ocean. (B) Summary of the sequencing, assembly and annotation strategy.

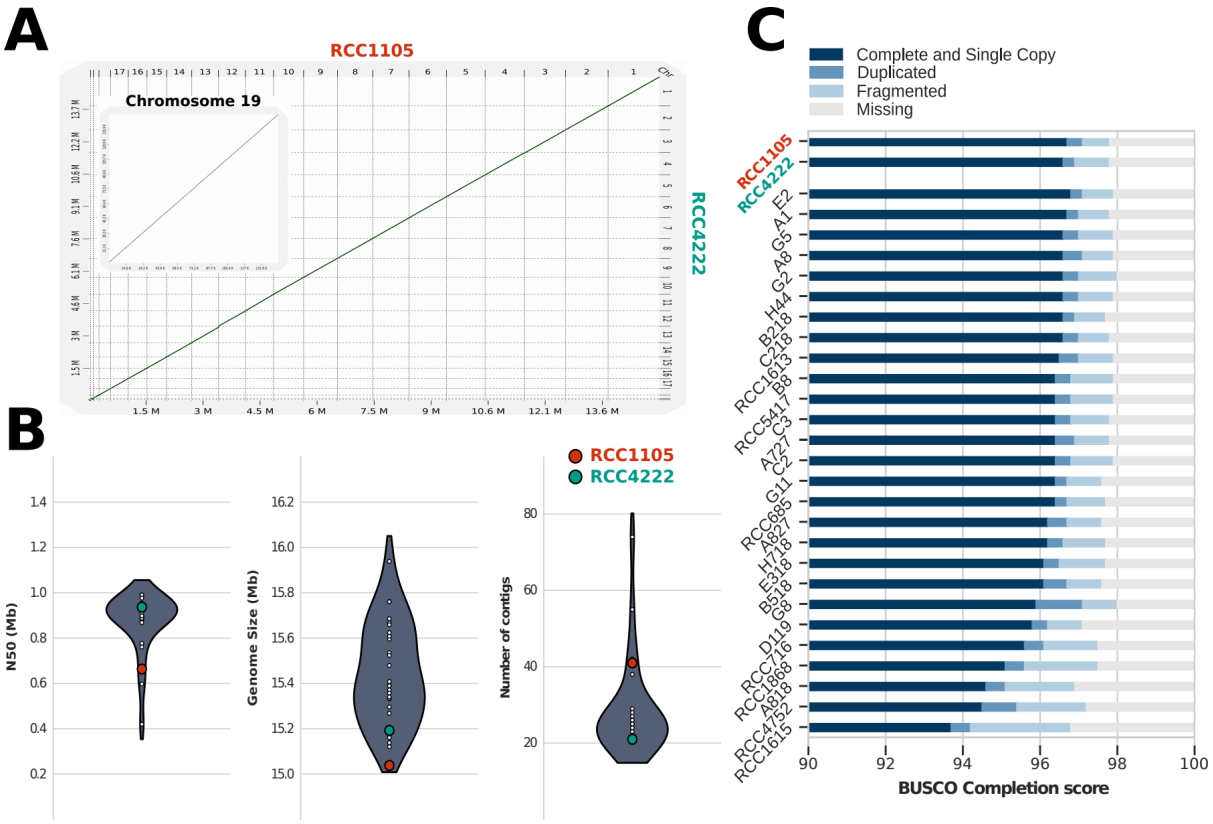

Fig 2 : Genome assemblies statistics and quality assessment. (A) Dot-plot of the reference genome RCC1105 *vs* the *de novo* assembled genome of the clonal strain RCC4222. (B) Distribution of principal assembly statistics for the 28 *de novo* assembled *Bathycoccus* genome. The N50 corresponds to the sequence length of the shortest contig at 50% of the total assembly length. (C) BUSCO completion score of genome assemblies using the *chlorophyta_odb10* library.

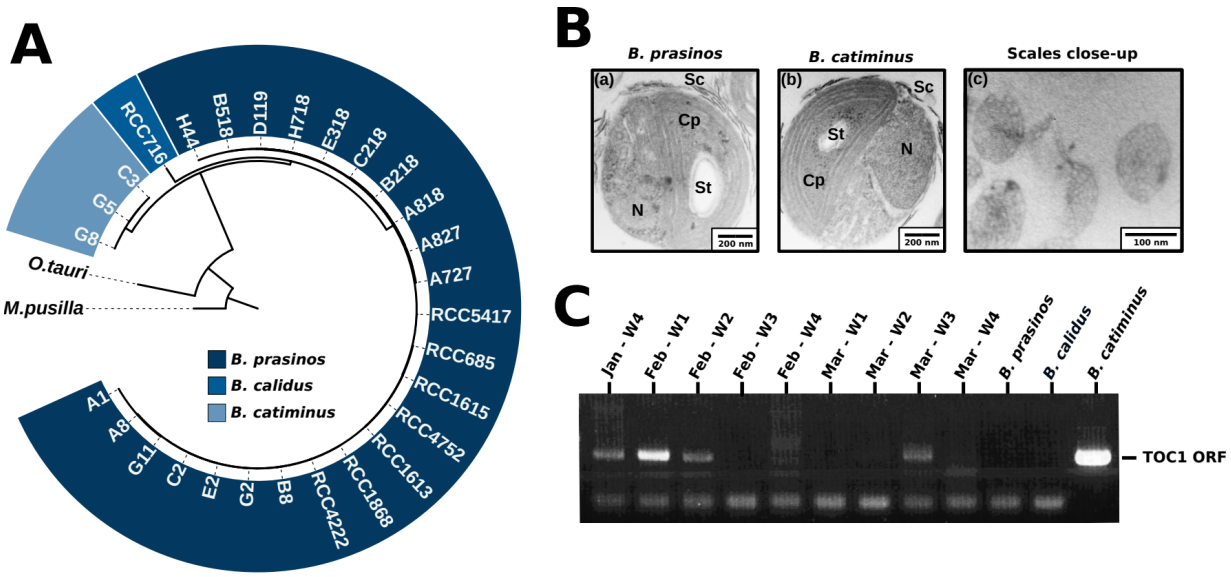

Fig 3 : Phylogenomic of the *Bathycoccus* genus and *B. catiminus* characterization (A) Maximum-likelihood phylogenetic tree of *Bathycoccus* strains based on 1,201 shared genes. Species separation is derived from the branch length and marked by colors. *O. tauri* and *M. pusilla* are used as outgroups. Only bootstrap values higher than 90% are displayed. (B) Transmission electron microscopy images of (a) *B. prasinos*, (b) *B. catiminus* isolated from the Banyuls bay, and (c) a close up of detached spider web scales from *B. catiminus*. N: nucleus, Cp: chloroplast, St: starch granule, Sc: scales. Bar = 100 nm. (C) Presence of *B. catiminus* in Banyuls bay seawater samples from 2019 identified by specific primers designed from non-conserved sequences of TOC1 ORF specific to *B. catiminus*.

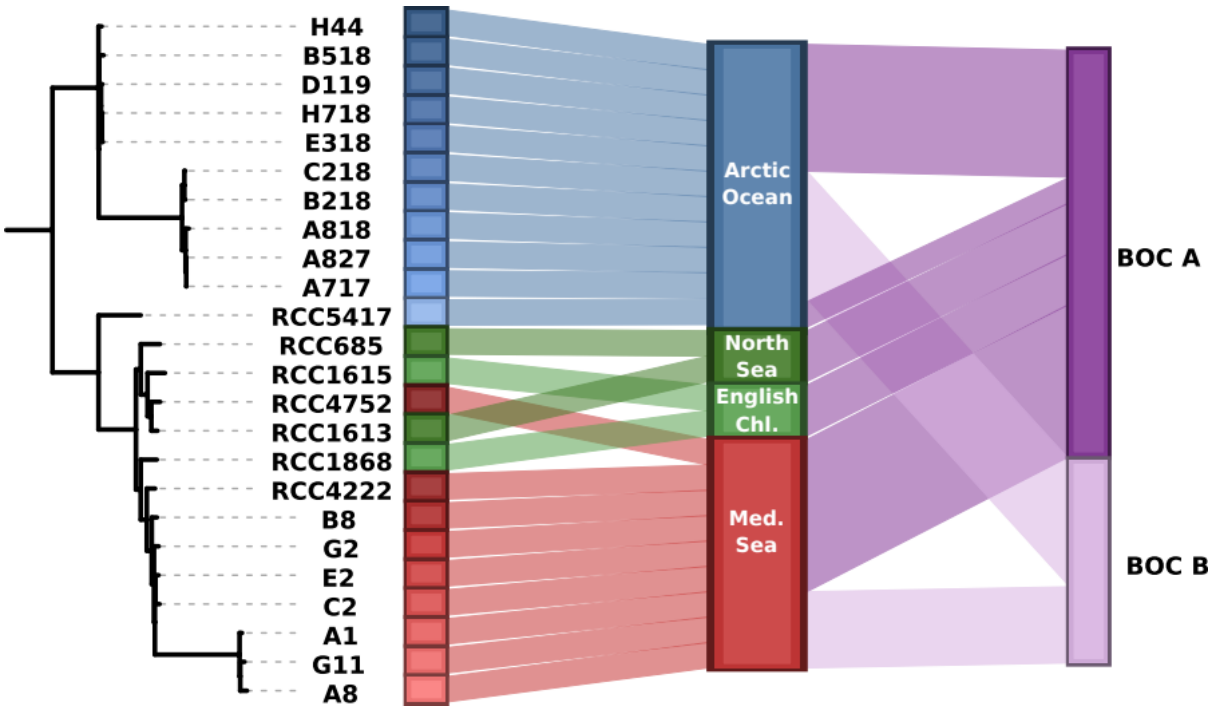

Fig 4 : Phylogenomic of *B. prasinos* linked to strain origin and Chromosome 14 BOC haplotypes. Maximum-likelihood phylogenetic tree of *B. prasinos* strains based on 1,201 shared genes. Only bootstrap values higher than 90% are displayed. BOC haplotypes separation was deducted from maximum-likelihood phylogeny based on 69 shared genes located on chromosome 14 **(Figure S5)** and confirmed by complete chromosome multiple sequence alignments **(Figure S6)**.

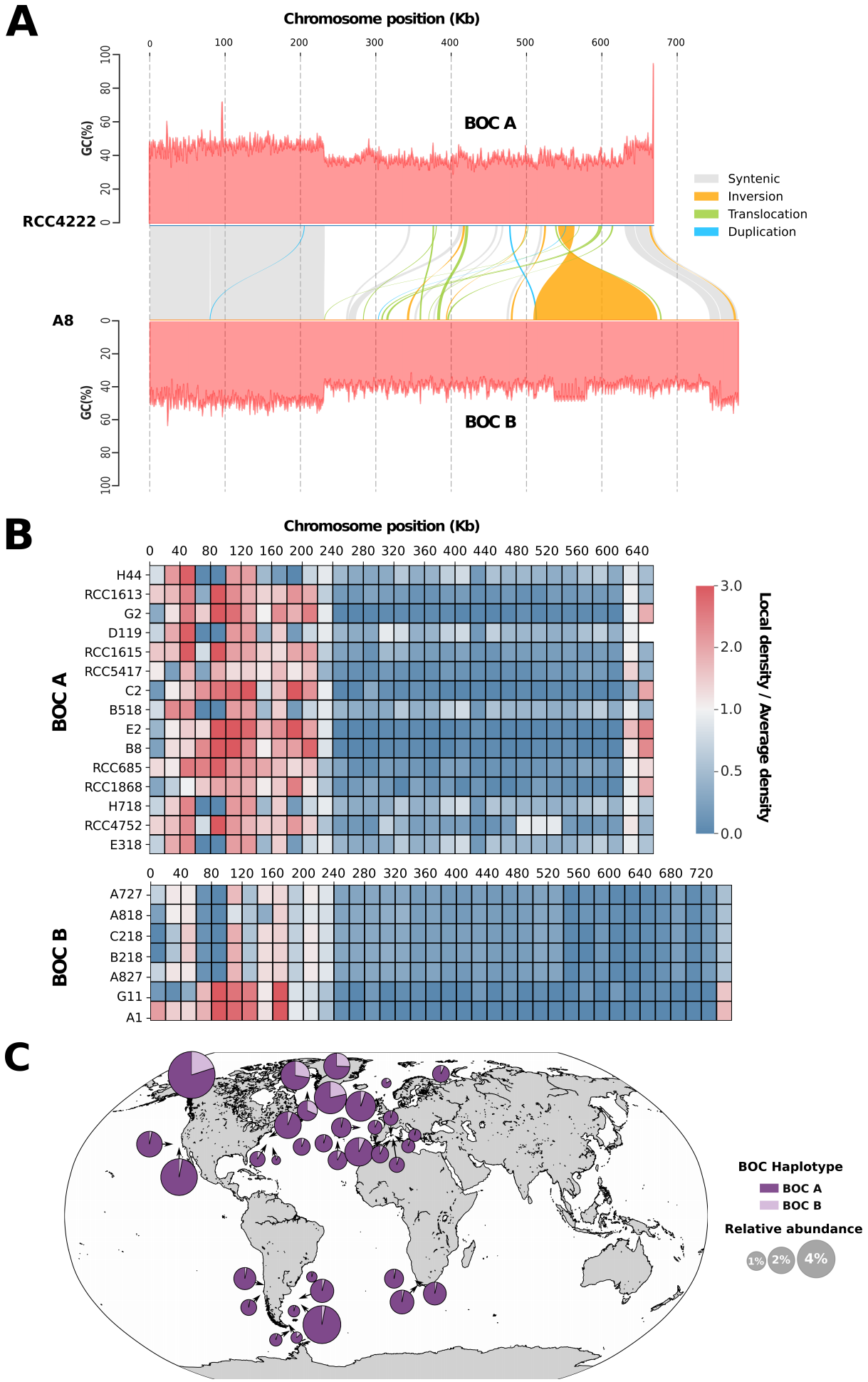

Fig 5 : Genomic diversity and haplotypes biogeography of *Bathycoccus prasinos* big outlier chromosome. (A) Alignment of Chromosome 14 from *B. prasinos* strains RCC4222 and A8. Colors indicate syntenic regions and structural rearrangements. Y axis displays the GC content of the corresponding sequences. (B) Divergences between local SNP density and genome wide SNP density along a 20 kb sliding window. SNP calling was computed for each BOC haplotype against strain RCC4222 for BOC A and strain A8 for BOC B. (C) BOC haplotypes ratio based on read mapping from TARA ocean metagenomic datasets in stations with *B. prasinos* relative abundance > 0.1%.

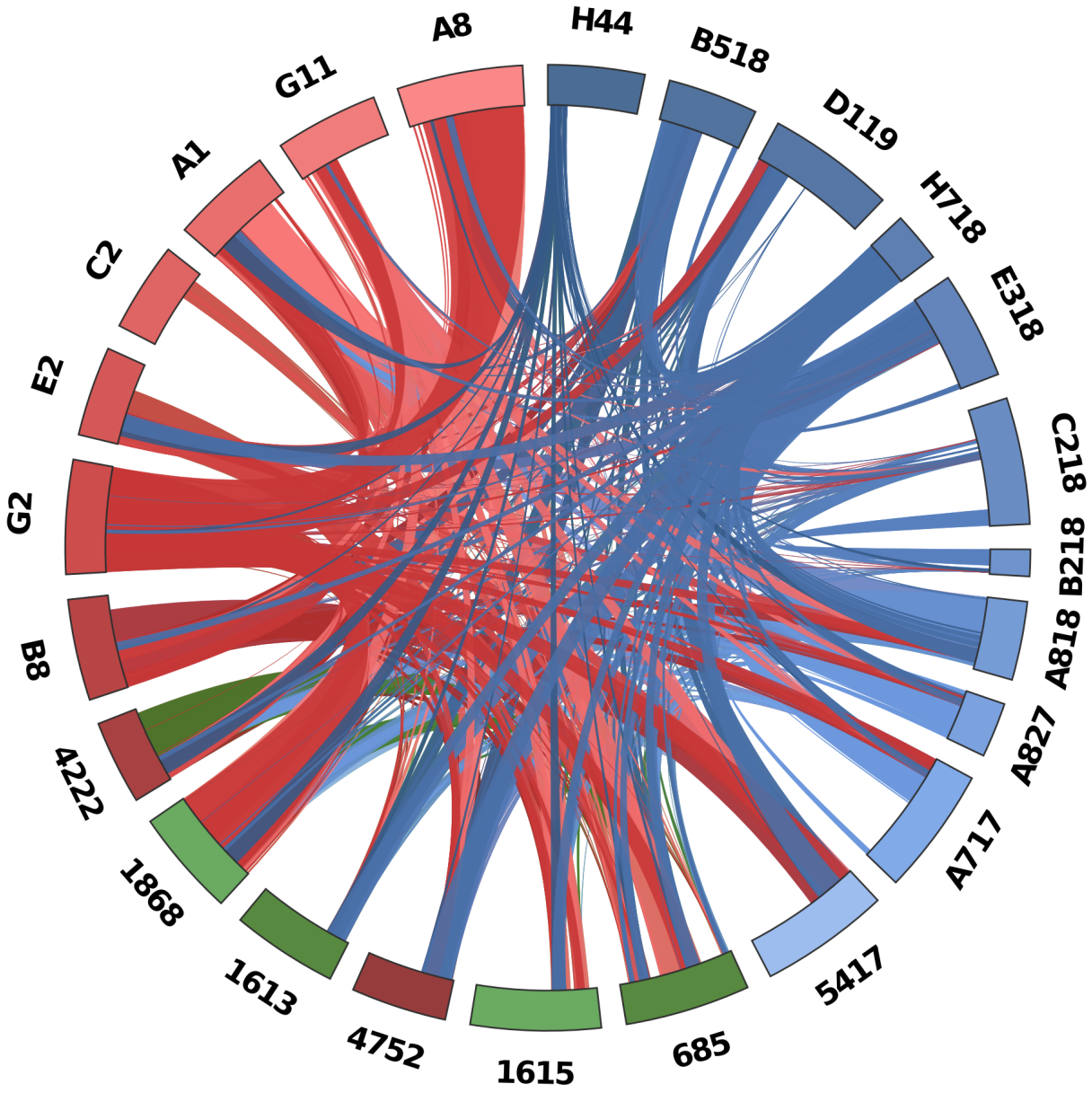

Fig 6 : Chromosome 19 SOC shared sequences between *B. prasinos* strains. Chord plot representation of chromosome 19 shared sequences between all 24 *B. prasinos* sequenced strains. Chromosome color corresponds to basin and geographical distribution of strains as shown in Fig. 4. Only shared sequences longer than 500 bp are displayed.

12. Supporting informations

Data S1: Quast output for all assembled genome against the reference genome

Appendix S1: RagTag was used to orient and scaffold contigs by mapping them onto a reference genome. All contigs that were not mapped on the reference are scaffolded together in a chimeric scaffold named Chr0. This scaffold contains sequences of contaminating DNA, including extracted DNA from bacteria associated in culture with *Bathycoccus*. It could also contain *Bathycoccus* assembled contigs that are not in the reference genome, in other words, strain specific big structural variations, however with a mean N50 of 920 kb it is very unlikely that any contig would contain structural variations without any flanking sequences allowing it to be anchored on the chromosomes of the reference genome. Dotplot comparisons and BlastN for sequences samples of Chr0 were computed across all assemblies to check that no *Bathycoccus* sequences were erroneously discarded. In the context of this study, this approach provided a quick and efficient way to eliminate contaminant sequences in our assemblies.

Table S1 : Multi loci genotypes identified in isolated strains from water samples collected in the Baffin bay during the DarkEdge cruise.

Table S2 : Prevalence of annotated repeat types in assembled genomes

Table S3: Statistics of annotated coding sequences in assembled genomes

Figure S1 : Identification and separation of contamination sequences from *Bathycoccus* sequences using Ragtag. (A) Dotplot of the reference genome of strain RCC1105 against the *de novo* assembled genome of the clonal strain RCC4222 including chromosome 0 of unaligned sequences identified by RagTag. (B) BlastN examples of Chromosome 0 sequence content.

Fig S2 : Maximum-likelihood phylogeny and alignment of available Bathycoccus ITS2 rDNA region. Only bootstrap values higher than 90% are displayed. ITS2 sequences were obtained from *B. prasinos* RCC1105 reference strain, *B. calidus* RCC716 strain and *Bathycoccus sp.* G8 strain respectively. *Bathycoccus* Kara environmental sequence was obtained from Belevich et al. (2021).

Fig S3 : Structural annotation pipeline of repeat and coding sequences. Repeat annotation was performed using RepeatMasker and a non redundant repeat library extracted from all 28 *Bathycoccus* genomes assemblies. Gene annotation was performed using RNAseq sequence from strain RCC4752 and protein sequences database as evidence to train

Genemark-ETP, Augustus and SNAP models. Models output was integrated using Maker.

Fig S4 : Biogeography of the *Bathycoccus* genus. Geographical distribution of *Bathycoccus* species based on read mapping from TARA ocean metagenomic datasets. For each species, only stations with horizontal genome coverage higher than 10% and 4 reads of minimal depth are displayed. Relative abundance corresponds to the percentage of reads from the dataset mapped to the genome.

Fig S5 : Phylogenomic of chromosome 14 conserved genes. Maximum-likelihood phylogenetic tree of *Bathycoccus* strains based on 69 conserved genes located on *B. prasinos* chromosome 14. Species separation is derived from complet phylogeny (Fig. 3A) and marked by colors. *O. tauri* and *M. pusilla* are used as outgroups. Only bootstrap values higher than 90% are displayed.

Fig S6 : BOC haplotypes shared sequences and ratio in isolated strains. (A) Chord plot representation of chromosome 14 shared sequences from strains RCC4222 (BOC A) and A8 (BOC B) on all strains. Chromosome color corresponds to the strain BOC haplotype: black, BOC A; grey, BOC B. Only shared sequences longer than 500 bp are displayed. (B) Pie chart of BOC haplotype ratio in isolated strains from the Banyuls bay (top, n = 55) and from the Baffin sea (bottom, n = 183)

Fig S7 : Structural variation (SV) density of Chromosome 14 in *Bathycoccus prasinos*. Divergences between local SV density and genome wide SV density along a 20kb sliding window. SV calling was computed for strains with identical BOC haplotypes against strain RCC4222 for BOC A and strain A8 for BOC B.

Fig S8 : Genomic diversity of Chromosome 1 in *Bathycoccus prasinos*. (A) Divergences between local SNP density and genome wide SNP density and (B) between local SV density and genome wide SV density along a 20kb sliding window. Variant calling was computed against strain RCC4222.

Fig S9 : Chromosome 19 pairwise alignments between *B. prasinos* strains. Chromosome 19 alignment and structural variation visualization between *B. prasinos* assemblies. Strain order follows phylogenetic order shown in Fig. 4. Strain C2 was not displayable due to the lack of shared sequences with phylogenetically close strains. Only structural variations longer than 500 bp are displayed.

Fig S10 : Chromosome 19 shared regions between *B. prasinos* strains from the same basin. Chord plot representation of chromosome 19 shared sequences between all 24 *B. prasinos* sequenced strains. Chromosome color corresponds to basin and geographical distribution of strains as shown in Fig. 4. Only shared sequences longer than 500 bp are displayed.

The Dark Edge genomics sampling team

Marcel Babin[5], Chris Bowler[6]

[5] Département de Biologie, Québec Océan and Takuvik Joint International Laboratory (UMI 3376), Université Laval (Canada)–CNRS (France), Université Laval, Quebec, QC, Canada.

[6] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, 75005, France.