An INDEL genomic approach to explore population diversity of phytoplankton : Bathycoccus , a case study

Martine Devic, Cédric Mariac, Valérie Vergé, Philipe Schatt, Louis Dennu, Jean-Claude Lozano, François-Yves Bouget, François Sabot

An INDEL genomic approach to explore population diversity of phytoplankton : *Bathycoccus*, a case study

Martine Devic*[1], Cédric Mariac[2], Valérie Vergé[1], Philipe Schatt[1], Louis Dennu[1], Jean-Claude Lozano[1], François-Yves Bouget*[1] and François Sabot*[2]

Author affiliations :

[1] Laboratoire d'Océanographie Microbienne (LOMIC), CNRS/Sorbonne Université, UMR7621, Observatoire Océanologique, 66650 Banyuls/mer, France.

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

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

Running head : Markers of genetic diversity in *Bathycoccus*

Key words : Phytoplankton, *Bathycoccus*, intraspecies diversity, INDEL marker, bloom

Abstract

Although metabarcoding has generated large dataset on world-wide phytoplankton species diversity, little is known about the intraspecies diversity underlying adaptation to environmental niches. To gain insight into population diversity, a novel INDEL based method was developed on *Bathycoccus prasinos*. Oxford Nanopore Technology (ONT) sequencing was first used to characterise structural variants (SV) among the genomes of *Bathycoccus* sampled from geographically distinct regions in the world ocean. Markers derived from INDEL were validated by PCR and sequencing in the world-wide strains. These markers were then used to genotype 55 *Bathycoccus* strains isolated during the winter bloom 2018-2019 in the bay of Banyuls-sur-Mer. With five markers, eight Multi Loci Genotypes (MLG) were determined, two of which represented 53% and 29% of the isolates. Physiological studies confirmed that isolates are phenotypically different, cells isolated in February growing better at low temperature than those isolated in December and January. When tested directly on environmental samples, two diversity markers showed a similar allele frequency in sea water as in individual *Bathycoccus* strains isolated at the same period. We conclude that these markers constitute a resource to identify the most abundant variant alleles in a given bloom. A follow-up on three consecutive blooms revealed differences in allele abundance during the course of a bloom, particularly at initiation and between years. This INDEL-based genotyping constitutes a new methodological approach that may be used to assess the population structure and diversity of other species.

Introduction

Marine phytoplankton, including picoeukaryotic algae, is responsible for a large fraction of primary production (Li et al. 1983). In temperate regions, the abundance and diversity of the phytoplankton is often seasonal and occurs in bursts, as algal blooms. Per se, blooms have a large impact on global primary production and therefore the understanding of the genetic basis of phytoplankton adaptation to seasonal niches and the effects of ocean warming on phytoplankton blooms are of the utmost importance.

Meta-ribosomal barcoding on the nuclear or plastidial 18/16S rRNA gene has opened the access to massive data in time and space and has accelerated the study of phytoplankton species diversity, interspecies co-occurrence and potential biotrophic interactions in natural communities. However, metabarcoding approaches do not provide information on intraspecies genetic variations. Assessing only interspecies diversity, underestimates the diversity of the populations. Equally, a single isolate cannot represent the diversity of a population. Since natural selection acts on variation among individuals within populations, it is essential to incorporate both intra and interspecies trait variability into community ecology (Violle et al. 2012). Raffard et al. (2019) demonstrated that intraspecies variation has significant ecological effects across a large set of species, confirming a previous estimate based on a more restricted species set (Des Roches et al. 2018). Furthermore, it has been shown that diversity within species is rapidly decreasing, making them more homogenous and highlighting the need to preserve intraspecies variations (Des Roches et al. 2018) since intraspecies diversity reinforces the overall population stability in the face of environmental change.

In diatoms, intraspecies variation has been shown to play a key role in the responses of the species to several important environmental factors such as light, salinity, temperature and nutrients (Godhe et al. 2017). Modelling efforts indicate that this variation within species

extends bloom periods and likely provides sufficient variability in competitive interactions between species under variable conditions. The intraspecies variation most likely corresponds to optimal fitness in temporary microhabitats. This rich intraspecies genetic diversity allows for the possibility of local adaptation and for differentiation in important physiological characteristics that produces local populations that are exceptionally fit and competitive in their respective local habitat.

Several studies suggest that previously recognized cosmopolitan species are actually composed of multiple populations or even multiple species (Kashtan et al. 2014). Genotypes or species can either replace each other temporally (but with overlap) as in the case of the marine diatoms *Pseudo-nitzschia multistriata* (Tesson et al. 2014) and *Skeletonema costatum* (Gallagher, 1980) or co-exist sympatrically as in the freshwater *Asterionella formosa* (Van Den Wyngaert et al. 2015). Furthermore, the frontier between variant genotype and species is thin and insight will be gained by whole genome sequencing of a large number of strains. Read et al. (2013) documented a pan genome of the coccolithophore *Emiliania sp* revealing that what was previously considered a single species, is actually composed of multiple species. Similarly Bentif et al. (2023) with morphological and genomic surveys showed *Gephyrocapsa huxleyi* has evolved to comprise at least three distinct species.

Since an assembly of genotypically diverse individuals constitutes a population, methodological approaches have been developed in order to determine the genetic variation among individuals (reviewed for Diatoms in Rynearson et al. 2022). One of the major challenges is the difficulty to isolate individuals in sufficient number for classical diversity analyses.

Historically, the intraspecies markers corresponded to small nucleotide repeats such as microsatellite (Srivastava et al. 2019), chloroplastic (Wheeler et al. 2014) and mitochondrial (Galtier et al. 2009) genes and a few nuclear genes that were applied to several hundreds of

isolated individuals. Microsatellites have been described in some diatoms (Tesson et al. 2011) but not to date in Mamiellales. The development of PCR for Randomly Amplified Polymorphic DNA (RAPD, Lewis et al. 1997) and more recently Restriction-site-Associated DNA sequencing techniques (RADseq, Andrews et al. 2016) have allowed the discovery and genotyping of thousands of genetic markers for any given species at relatively low-cost. Some of these approaches have been used to analyse the diversity of populations during algal blooms (Rengefors et al. 2017). The recent dramatic increase of the number of sequenced genomes led to large-scale diversity studies with large sets of nuclear genes or whole genome comparisons. However most of these approaches required the isolation of a large number of individuals. As a consequence, intraspecies diversity has been poorly documented in the past in marine phytoplankton.

Widely distributed from the equator to arctic and antarctic poles with a marked seasonality in temperate and polar regions (Joli et al. 2017, Tragin et al. 2018, Lambert et al. 2019, Leconte et al. 2020), picoeukaryotes belonging to the order of Mamiellales (*Bathycoccus*, *Ostreococcus and Micromonas*) have a cosmopolitan presence illustrating a high capacity for adaptation to a wide range of contrasting environments. Novel, rapid and cheap sequencing technologies have given access to Mamiellales diversity by metagenomic approaches (Leconte et al. 2020, Da Silva et al. 2022, Richter et al. 2022) or metatranscriptomic approaches (Simmons et al. 2016), however to date, very little information is available on intraspecies diversity of Bathycoccaceae with the exception of *Ostreococcus tauri* (Blanc-Mathieu et al. 2017). Unlike *O. tauri* which is usually not detectable in publicly available metagenomes, *Bathycoccus* is the most cosmopolitan Mamiellophyceae (de Vargas et al. 2015). *Bathycoccus* can be divided in two species, the polar and temperate *Bathycoccus prasinos* type B1 genome (Moreau et al. 2012, Joli et al. 2017) and the tropical *Bathycoccus calidus* type B2 genome (Vannier et al. 2016, Limardo et al. 2017, Bachy et al. 2021). Thus

the cosmopolitan nature of *Bathycoccus* from poles to equator might be due to the combination of both B1 and B2 species.

In the bay of Banyuls, *Bathycoccus* and *Micromonas* bloom yearly from November to April, *Bathycoccus* being one of the most abundant species (Lambert et al. 2019). The highly reproducible yearly occurrence of *Bathycoccus* in the Banyuls bay during the last decade (Lambert et al. 2019) raises the question of the persistence of a *Bathycoccus* population adapted to the bay or of a variation of the population structure each year. In addition, since outside of the bloom period Mamiellales are virtually absent from the bay, is the *Bathycoccus* bloom initiated by an uptake of resident "resting cells" in the sediment or by a fresh input carried by North western Mediterranean currents along the Gulf of Lion? At present no resting stages that can act as inoculum of subsequent blooms have been described for *Bathycoccus*.

To assess the intraspecies diversity of *Bathycoccus* in the Bay of Banyuls, we combined an efficient method to isolate Mamiellales together with whole genome sequencing by Oxford Nanopore Technology (ONT) in order to identify Structural Variants (SV) in the *Bathycoccus* genome. Diversity markers designed from INDEL (insertion or deletion of bases in the genome of an organism) were used to genotype *Bathycoccus* strains and populations from environmental samples. This approach constitutes an unprecedented tool which could be potentially applied to a large variety of species.

Materials and Methods

Algal strains and culture conditions

World-wide *Bathycoccus* strains were obtained at the Roscoff Culture Collection (RCC) centre and renamed as a town or region to be reminiscent of their geographical origin: RCC4222 (named BANYULS), RCC5417 (BAFFIN), RCC1613 (OSLO), RCC685

(HELGOLAND), RCC1615 (DIEPPE), RCC1868 (ROSCOFF), RCC4752 (NAPLES) (Supplemental Table 1). RCC1105-REFERENCE from which the current *Bathycoccus prasinos* reference genome originated (Moreau et al. 2012) was lost and replaced by RCC4222-BANYULS. The strains were cultivated in 100 mL flasks in filtered artificial seawater (24.55 g/ L NaCl, 0.75 g/L KCl, 4.07 g/L $MgCl_2$ $6H_2O$, 1.47 g/L $CaCl_2$ $2H_2O$, 6.04 g/L $MgSO_4$ $7H_2O$, 0.21 g/L $NaHCO_3$, 0.138 g/L NaH_2PO_4 and 0.75 g/L $NaNO_3$) supplemented with trace metals and vitamins (Supplemental data 1). Cultures were maintained under constant gentle agitation in an orbital platform shaker (Heidoph shaker and mixer unimax 1010). Sunlight irradiation curves recreating realistic light regimes at a chosen latitude and period of the year were applied in temperature-controlled incubators (Panasomic MIR-154-PE).

Cell isolation

Surface water was collected at 3 meter depth at SOLA buoy in Banyuls bay, North Western Mediterranean Sea, France (42°31′N, 03°11′E) approximately every week from December 2018 to March 2019, November 2019 to March 2020 and October 2020 to April 2021. Two ml aliquots were used to determine the quantity and size of phytoplankton by flow cytometry. For the bloom 2018/2019, 50 ml were 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 4 culture flasks with 10 ml of filtrate each. The sea water was supplemented by vitamins, NaH2PO4, NaNO3 and metal traces at the same final concentration as artificial sea water (ASW; Supplemental data 1), antibiotics (Streptomycine sulfate and Penicillin at 50 µg/ml) were added to half of the cultures. The cultures were incubated under light and temperature conditions similar to those during sampling date for 3-4 weeks. The presence of picophytoplankton was analysed by a BD accuri C6 flow cytometer. In general, superior results were obtained without antibiotics. Cultures containing at least 90% of picophytoplankton with only residual

nanophytoplankton were used for plating on agarose. Colonies appearing after 10 days were hand-picked and further cultured in 2 ml ASW in deepwell plates (Nunc, Perkin Elmer, Hessen, Germany) for 10 days. Cells were cryopreserved at this stage. Circa 500 clones were cryopreserved. At the same time, DNA extraction and PCR were performed in order to identify *Bathycoccus* clones.

DNA extraction, genome sequencing, assembly and PCR amplification

For PCR analysis, total DNA was extracted from 4 ml *Bathycoccus* cell cultures according to the Plant DNA easy Qiagen protocol. For whole genome sequencing by Oxford Nanopore technology (ONT), DNA was extracted by a CTAB method from 100 ml culture principally based on Debladis et al. (2017). ONT libraries were barcoded using the Rapid Barcoding Sequencing (SQK-RBK004) and deposited on R9.4 flow cell. For environmental samples, 5 litres of seawater at SOLA 3 meter depth were passed through 3 microns and 0.8 micron pore filters. DNA from cells collected on the 0.8 micron filters were extracted using the Plant DNA easy Qiagen protocol with the addition of a proteinase K treatment in the AP1 buffer. PCR was performed using the Red Taq polymerase Master mix (VWR) with the required primers (Supplemental data 2) and corresponding DNA. For sequencing, the PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel reference 740609.50) and the filtrate was sent to GENEWIZ for Sanger sequencing.

Raw ONT Fast5 data were basecalled using Guppy 4.0.5 (https://nanoporetech.com) and the HAC model, and QC performed using NanoPlot 1.38.1 (De Coster et al. 2018). All reads with a QPHRED higher than 8 were retained and subjected to genome assembly using Flye 2.8 (Kolmogorov et al. 2019) under standard options. Raw assemblies were then polished with 3 turns of standard Racon (Vaser et al. 217) after mapping of raw reads on the previous sequence using minimap2 (-ax map-ont mode; Li et al. 2021). Final scaffolding was performed using Ragoo 1.1 (Alonge et al. 2019) upon the original *B. prasinos* reference

189 genome (GCA_002220235.1, Moreau et al. 2012). Final QC of assemblies was performed
190 using QUAST 5.0 (Mikheenko et al. 2018).

191 *Relative allelic abundance in environmental samples*

192 Amplifications were performed twice with a difference of 2 cycles in order to obtain clear
193 bands on ethidium bromide stained agarose gels for each sample. Similarly, the gels were
194 photographed after different exposure times in order to obtain a non-saturated image for each
195 sample. Relative abundance of each variant within the same DNA sample was performed
196 using ImageJ software Analyse Gel.

197 *Determination of Growth Rates*

198 Cells isolated during December 2018, January and February 2019 in Banyuls bay were used
199 in this experiment. For each culture condition the cell number was determined by flow
200 cytometry daily, for 9 days. The growth rate was determined as Ln(N)/dT, where N is the cell
201 concentration per ml and T the time (days). The maximal growth rate (μmax) was determined
202 according to Guyon et al. (2018) on a graph expressing the neperian logarithm of cell
203 concentration as a function of time of culture. Mmax corresponded to the slope of the linear
204 part of the growth curve (i.e., excluding the lag phase and the stationary phase).

205 $$\mu_{max} = Log(N_{fmax}) - Log(N_0)/Log(2) \times T$$

206 *Measurement of photosynthetic efficiencies*

207 Cultures were acclimated to the temperature and light rhythm and intensity for 7-10 days
208 before subculturing at 10^6 cells /ml in triplicates. After 3-4 days of growth, the photosynthetic
209 activities were recorded with PHYTO-PAM-II (Walz). After 20 min in obscurity, the samples
210 were transferred into the PHYTO-PAM and Fv/Fm, ETRmax and NPQ were measured.

211

212 ## Results

213 ***Search for intraspecies diversity markers***

With the aim to differentiate *Bathycoccus* isolates, we undertook a search for genetic determinants of diversity. Since only two Mediterranean strains were available in the Roscoff Culture Collection at the beginning of the project, we examined the world-wide diversity of *Bathycoccus* and selected the most geographically dispersed strains (Supplemental Table 1). The strains align along a latitude gradient from the Baffin bay (67°) to the Mediterranean sea (40°): RCC5417-AFFIN, RCC1613-OSLO, RCC685-HELGOLAND, RCC1615-DIEPPE, RCC1868-ROSCOFF, RCC4222-BANYULS and RCC4752-NAPLES. For simplification, the town close to the site of sampling instead of the RCC number will name strains further on. Oxford Nanopore Technology (ONT) was used to sequence the genome of the selected strains. After *de novo* assembly, each genome was compared to the reference genome (Moreau et al. 2012). There were some large chromosomal rearrangements but for the design of diversity markers, we only considered INDEL inferior or equal to 2 kb within regions mapping on the reference genome. The number and size of INDEL are detailed in Table 1. The goal was to identify INDEL instead of SNP (Single Nucleotide Polymorphism) that could be used to genotype the strains directly by PCR.

Validation of sequence variations in the genomes of world-wide Bathycoccus

Putative markers were selected on several criteria. The insertion should be found at the same or close location in the genome of at least three strains and of different sizes in at least two genomes, preferably three. In addition, the size of the amplified fragment should be between 200 bp and 2 kb (this size restriction reduced the mean number of insertions from 88 to 35, Table 1) and sufficiently different among the genomes of the strains to be unambiguously visualised on agarose gel after amplification with a single set of primers. The aim of this drastic selection was to identify the most divergent markers among the largest available genetic diversity of *Bathycoccus* with the expectation that some of this variation would be found in local communities of *Bathycoccus* in the Banyuls bay. Only five candidate markers

met these criteria and were experimentally tested. For two markers targeting variations of the number of amino acid repeats in open reading frames, primers positioned at proximity of the repeats did not produce a single amplimer and these two predictions from ONT sequencing could not be validated nor invalidated. Marker on chromosome 15 (the number of repeats in a zinc finger protein), marker chromosome 3 (variation in repeat number in a flavodoxin-like protein) and marker chromosome 1 (insertion and deletion into the promoter of yrdC gene) were validated (Table 2). To increase the number of markers, the striking insertion of 1.5 kb into the promoter of the clock gene TOC1 (TIMING OF CAB EXPRESSION 1) on chromosome 17 was included even though its diversity was below three (Table 2). The fifth marker was selected as marker of the Big outlier Chromosome (BOC) on chromosome 14, an atypical chromosomal structure found in Mamiellales (Moreau et al. 2012).

Typical PCR results for the 5 validated diversity markers are presented in Figure 1. The detailed description of each marker is provided as supplementary data 3.

Marker yrdC promoter: Bathy01g04300 encodes a yrdC domain-containing protein of unknown function. In *Escherichia coli*, yrdC binds preferentially to double-stranded RNA, consistent with a role of the protein in translation (Teplova et al. 2000). A diverse organisation was identified in the promoter region of Bathy01g04300 in comparison to the reference genome and this was visualised by PCR amplification (Figure 1, Table 2, Supplemental data 3A). This set of primers showed marked differences in its amplification success among the strains (specially in Baffin) indicating important nucleotide variations.

Marker TOC1 promoter: The intergenic region upstream Bathy17g01510 encoding an homolog of TOC1 involved in the control of circadian rhythm was different in some strains. A 2.2 kb insertion was identified in OSLO, DIEPPE and NAPLES, whilst the other strains were similar to BANYULS (Figure 1, Table 2, Supplemental data 3B).

Marker flavodoxin-like: Bathy03g02080 encodes a protein containing a flavodoxin-like domain, a flavin mononucleotide (FMN)-binding site and 6 imperfect repeats of 25 amino acids. In comparison to the reference, ONT sequencing revealed a deletion in BAFFIN, insertions of 147 bp in BANYULS, 74 bp in OSLO and 375 bp HELGOLAND, while ROSCOFF and NAPLES were unchanged. These predictions verified by PCR and sequencing confirm that substantial INDEL can also occur within coding regions (Figure 1, Table 2, Supplemental data 3C).

Marker Zinc Finger: Bathy15g02320 encodes a protein with Zinc Finger repeats (ZF) of a greater length in BAFFIN (9ZF) and HELGOLAND (7ZF) than in BANYULS (6ZF), differences that were tested by PCR (Figure 1, Table 2, Supplemental data 3D). Since the primers did not amplify a single fragment in BAFFIN and HELGOLAND, the corresponding region was obtained by the ONT data (Table2). Zinc finger C2H2 proteins are numerous (53 genes) and highly conserved in the *Bathycoccus* reference genome (Moreau et al. 2012).

Marker TIMa: Bathy14g30100 encodes a protein containing a TIM domain found in the protein Timeless involved in circadian rhythm control in Drosophila (Sehgal et al. 1995). This gene is located in the BOC region of chromosome 14. The ORF of TIMa was found conserved among the selected strains (Figure 1, Table 2). This fifth marker was non discriminating among the subset of world-wide *Bathycoccus* strains.

Accessory genes in Bathycoccus prasinos genome

The large insertions found in the promoters of yrdC and TOC1 were sequenced and analysed in detail (supplementary data 3A-B).

In yrdC promoter, a gene encoding a protein of unknown function possessing ANK repeats similar to the gene products of Bathy01g04610 (68% amino acid identity), Bathy01g04570 (67%) and Bathy11g02720 (57%) was inserted between an Evening Element-like (EEL) *cis* element and the start of Bathy01g04300 in BAFFIN, OSLO, HELGOLAND and NAPLES. In

order to gain information on this additional gene, the Ocean Gene Atlas (OGA) website (http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/) (Villar et al., 2018, Vernette et al. 2022) was interrogated with the additional protein and its three homologs. The only sequence retrieved from OGA shared significant similarity with the 4 proteins but was not identical. So the geographical distribution that we obtained is not the one corresponding to the accessory ANK gene.

In TOC1 promoter, the 2.2 kb insertion identified in OSLO, DIEPPE and NAPLES encodes a Methyltransferase-like protein (AdoMTase *METTL24* IPR026913) of 320 aa with 41.7% amino acids identity to the predicted *Ostreococcus lucimarinus* CCE9901 protein (XP_001422352). Searches were performed at high stringency in OGA (Expect threshold 1e-300) so that only the presence of the near-identical sequences was retrieved. TOC1 and AdoMTase sequences were not strictly co-occuring. At one station located near Chile (arrow Figure 2), TOC1 sequences were abundant but no AdoMTase was recorded which suggests that in this particular area the *Bathycoccus* genomes are almost devoid of AdoMTase. Surprisingly, it was possible to also find AdoMTase sequences not associated to *Bathycoccus* TOC1 sequences. Many AdoMTase hits were found near the equator at temperatures higher than 25°C, temperatures too high for *Bathycoccus prasinos* growth, indicating that this accessory gene can be found in other unknown microorganisms suggesting possible horizontal gene transfer. As a control, TOC1 and CCA1 (CIRCADIAN CLOCK ASSOCIATED 1, Bathy06g4380) sequences were fully co-occurrent (Supplemental Figure 1). Transcription of the gene encoding this AdoMTase was confirmed in the metaT database at OGA (data not shown).

In conclusion, two large INDEL comprised additional or duplicated genes and these should be considered as structural rearrangements rather than simple INDEL. These accessory genes could potentially be beneficial for adaptation.

313 ***Isolation of Bathycoccus during the 2018/2019 winter bloom in Banyuls bay***

314 Surface water was collected weekly from December 3rd to March 19th. During this period, the
315 sea temperature rose to 15.87°C in December and did not descend below 10.68°C in February
316 (Figure 3). According to the 10-year study at the same location (SOLA Buoy in Banyuls bay)
317 for 2007-2015 (Lambert et al. 2019) and 2015-2017, this period can be considered as an
318 average climatic year in term of temperature (Lambert et al. 2021). The presence and
319 abundance of pico- and nano-phytoplankton <3 μm were determined by flow cytometry
320 (Figure 3). Whilst cyanobacteria were the most abundant at all time with a peak at the end of
321 February, picophytoplankton was the second most abundant category with a first peak in
322 December and a second in February (Supplemental Figure 2). At each sampling date,
323 collected seawater was also filtered through 1.2 μm pore-size and transferred to culture flasks.
324 This pore size allows the passage of most *Bathycoccus* cells (which size is estimated at 1,5
325 μm) and most importantly eliminate the potential larger predators.
326 After a period of acclimation of two weeks in supplemented sea water, cells were isolated by
327 plating on agarose. Light green-yellow coloured colonies were picked and sub-cultured. In
328 order to accelerate the identification of *Bathycoccus* among the isolated cells, amplifications
329 of a fragment of the *LOV-HK* (Bathy10g02360) gene were performed. These primers were
330 specific to the *Bathycoccus prasinos* genome and did not amplify the homologous gene in
331 *Ostreococcus* or *Micromonas* nor in any other species. The identity of these clones was
332 further confirmed by ribotyping (amplification of a 2kb ribosomal DNA fragment followed
333 by sequencing). In total, 55 *Bathycoccus prasinos* isolates were recovered at nine sampling
334 dates (Supplemental Table2).

335 ***Identification of dominant Bathycoccus Multi Loci Genotypes in Banyuls bay in 2018/2019***

336 The five diversity markers were used in a combination of PCR and sequencing in order to
337 distinguish the different isolates of *Bathycoccus* sampled during the 2018/2019 winter bloom.

For the *yrdC* promoter, the inserted gene encoding a protein with ANK repeats was not detected in the Banyuls samples. In one isolate only (F1), the yrdC promoter was near identical to the one in RCC4222-BANYULS (99% identity in 332 bp, Supplemental data 3A). In 54 other isolates, a fragment of 200 bp similar in size to that in ROSCOFF was amplified and sequenced (100% identity, Supplemental data 3A). The insertion present in the *TOC1* promoter was more prevalent in the Banyuls isolates (91%) than in the world-wide strains (43%) (Table 2 and 3). A high degree of similarity (98.6%) was observed over the entire insertion whilst the core promoter of *TOC1* containing the essential Evening Element-Like (EEL) *cis* element had a minimum of 71.6% nucleotide identity in 400 bp. A phylogenetic tree was constructed with these sequences (Supplemental data 3B). The 2 kb inserted sequence was not included in the study but its occurrence is indicated between brackets for each isolate. Most of the Mediterranean isolates were found in the main clade which divided further into two subclades Ia and Ib. Subclade Ia contained all the Mediterranean reference type and HELGOLAND, whereas subclade Ib most of the Mediterranean 2kb type and DIEPPE. OSLO and NAPLES lie in a separate clade II. ROSCOFF and BAFFIN contained the most divergent sequences in clade III. As a consequence of the 2kb insertion, the *TOC1* promoter activity is most likely not abolished since the crucial EEL *cis* element is still present but the analysis of the core promoter sequences indicates that its activity could differ between the clades. Thus the *TOC1* promoter could constitute a functional marker as well as a diversity marker.

The gene encoding a flavodoxin-like protein was more diverse in size than were the intergenic regions (Supplemental data 3C, Table 2, 3). The maximum size difference increases to 245 amino acids between F1 (February Banyuls isolate) and BAFFIN. The function related to this sequence variation is unknown. It forms a coiled-coil structure of several alpha-helices. Coiled-coil domains have been identified in enzymes where they function as molecular

spacers positioning catalytic activities. Thus the variable length of the repeats could influence the activity of the flavodoxin-like protein.

The genome of most Banyuls isolates encoded a Zinc finger C2H2 protein with 6 ZF motifs (Bathy15g02320, Supplemental data 3D, Table 3). Only three isolates from February sampling were similar to HELGOLAND with an additional ZF motif.

Amplification of a fragment of 530 bp of TIM (TIMa) was observed in 66% (approximately two thirds) of the isolates (Table 3). For the isolates for which there was no amplification, we designed an additional primer in order to amplify a variant of TIM, TIMb. The alignment of the two predicted variant proteins TIMa in RCCB4222-BANYULS and TIMb in A8 isolate showed that one third of the protein is well conserved while two thirds were more variable (Supplemental data 3E). A phylogenetic tree confirmed this dichotomy (Figure 4A). OGA metagenomic database was interrogated with TIMa and TIMb and each retrieved a single hit, respectively OGATIMa and OGATIMb confirming the existence of the 2 isoforms of TIM world-wide. OGATIMb has a marked abundance in high latitudes and cold temperatures in the Northern hemisphere while OGATIMa is more widely distributed (Figure 4B).

Based on the results described above, the Banyuls isolates were classified in eight MLG (Table 3). MLG 1 and MLG 2 represent respectively 53% and 29% of the population. Since the number of isolates from each sampling date was not identical (Supplemental Table 2), this percentage may not be entirely representative. However, the presence of MLG 1 and 2 in five out of nine independent samplings rules out a bias due to experimental cloning during isolation (Table 3). We can thus confidently state that these two MLG were dominant in Banyuls bay during the 2018/2019 bloom. No isolate with a MLG identical to RCC4222-BANYULS was found.

Determination of major allelic variants in environmental samples: a three year follow-up

The identification of dominant MLG during the bloom 2018/2019 raises the question of their yearly or occasional prevalence in Banyuls bay. Since isolating strains is highly time consuming, an alternative approach was developed to estimate local diversity. Five litres of seawater were sampled once a week and filtered between 3 and 0.8 µm. DNA extracted from 0.8 µm filters was used as template for PCR analysis using our set of diversity markers. Samplings were performed during 3 successive blooms from 2018 to 2021 (Figure 3, Supplemental Figure 2).

Variations in the yrdC and TOC1 promoters and TIM ORF sequences were analysed on these environmental DNA samples. For markers Flavodoxin-like and zinc finger, despite the high specificity of the primers on DNA of individual *Bathycoccus* (Table3), they could not be used on complex environmental DNA samples due to the presence of high background.

The initiation of the bloom was analysed during two consecutive years (Figure 5). In 2019, the presence of *Bathycoccus* was detected in the third week of November with a clear predominance of the 200 bp allele of yrdC promoter and the presence of both TIMa and TIMb. In 2020, *Bathycoccus* was detected in October, was barely detected or absent in November and reappeared in December, consistent with the decrease of abundance of picophytoplankton by flow cytometry (Supplemental Figure 2). In 2020, the allelic ratios of 200/400 bp of the yrdC promoter were clearly different from those in 2019 and TIMb was not detected. We conclude that the onsets of the bloom were different both their chronology and their population diversity.

The diversity of populations was also assessed during the bloom (Figure 5, Figure 3). The abundance of alleles of yrdC promoter and TIM were clearly different in November-December compared to February.

In summary, the study of three successive blooms showed changes in the diversity of *Bathycoccus* populations within and between blooms.

Physiological characteristic of isolated Bathycoccus strains

Since abundance of alleles showed that November-December and February populations are different and because temperature and light intensity are significantly different in December (15.9°C, 9h15 light, maximum intensity 540 µE/m^2/s) and February (11.8°C, 10h30 light, maximum intensity 830 µE/m^2/s), the physiological parameters of the December (D, 03/12/2018), January (J, 28/01/2019) and February (F1 F2, 25/02/2019; F3-F4, 26/02/2019) isolates were determined. Cells were grown at 13°C or 16°C under December or February illumination. In the 4 conditions, F1-F4 isolates grew better than D and J cells (Figure 6), although the difference was reduced at 16°C.

Photosynthetic parameters were determined by PAM fluorometry. For most isolates, no significant differences in indicators of photosynthesis parameters were observed with the exception of F1 with the unique MLG 7 (Supplemental Figure 4).

Together, the results showed a clear difference in the growth curve of the *Bathycoccus* isolated in December and February mainly due their capacity of adaptation to low temperature.

Discussion

Contribution of ONT sequencing to the identification of Bathycoccus *INDEL markers*

Bathycoccus has a small nuclear genome of approximately 15 Mb distributed among 19 chromosomes and has only been found in a haploid phase (Moreau et al. 2012). This organisation makes it suitable for Oxford Nanopore Technology Rapid barcoding libraries and sequencing. On a single flow cell, it was possible to obtain sufficient coverage for the genome of up to 4 strains. The main difficulty was to obtain good quality genomic DNA for each strain, a criterion particularly important when pooling barcoded libraries. With the exception of RCC1615-DIEPPE, all ONT data were superior to 10 times the size of the *Bathycoccus* genome (Table 1). When comparing ONT data from RCC4222-BANYULS to

437 the clonal RCC1105 reference genome, 19 INDEL were found. Two main reasons can be
438 proposed to explain this variation. Firstly, RCC14222-BANYULS is not strictly identical to
439 RCC1105. Secondly, ONT is particularly suitable to identify structural variations previously
440 undetected by conventional sequencing method (Michael et al. 2018, Mantere et al. 2019).
441 Therefore, most of the INDEL identified between RCC1105 and RCC4222-BANYULS
442 genomes could result from the use of different sequencing techniques. To design diversity
443 markers, INDEL in intergenic regions (promoters) as well as ORF encoding repeat amino
444 acids sequences were selected. Essentially all predictions were accurate and validated by PCR
445 and sequencing.

446 ***Identification of Bathycoccus local diversity using INDEL markers***

447 The 55 freshly isolated Mediterranean *Bathycoccus* were genotyped using five INDEL
448 markers that were based on ONT sequencing of strains originating from contrasted
449 geographic locations between arctic and temperate regions. The presence of a marker type
450 was quite different between the world-wide strains and the Banyuls strains (Table 2 and 3).
451 For example, the insertion of a gene encoding ANK repeats was found in four out of seven
452 world strains but was absent from Banyuls strains, whereas the insertion of the AdoMTase
453 sequence into the TOC1 promoter was much more prevalent in Banyuls strains. In general,
454 the genomes of Banyuls *Bathycoccus* isolates possess common allelic variants (65-98%,
455 Table 3) representing the dominant MLG 1 and 2. Remarkably, these three dominant
456 INDEL/rearrangements (200 bp yrdC, 2.2 Kb TOC1, 1.2 Kb Flavodoxin-like) were not found
457 in RCC4222-BANYULS that was isolated in the Banyuls bay in 2006. In addition, none of
458 the 2018-2019 isolates share the same five marker types with BANYULS (2006) or NAPLES
459 that was isolated in 1986 (Table 3). Thus, it is clear that RCC4222-BANYULS isolated in
460 Banyuls bay in 2006 was certainly not abundant and probably not present during the bloom

2018/2019. Finally, we did not observe any obvious consistent pattern of occurrence between specific INDEL markers and the geographic origin of the world-wide strains (Table 2 and 3).

The observation of eight MLG in *Bathycoccus* is probably an underestimation. As an example in diatoms (although not directly comparable since microsatellite markers have a higher mutation rate than other region of the genome), more than 600 individuals have been genotyped using microsatellite markers, it was estimated that the blooming population was comprised of at least 2400 different genotypes (Rynearson and Armbrust 2005). With additional markers, the number of MLG would probably increase. However, the dominance of a few MLG among *Bathycoccus* isolates is probably representative, despite being based on a small number of isolates and loci.

The existence of major MLG highlights an apparent paradox: how can blooms be diverse, given that the best genotype should prevail? Blooms are predicted to quickly become dominated by a few particularly well-adapted genotypes (De Meester 1996). Nevertheless, most studies describing genetic diversity of blooming phytoplankton populations report high intraspecies variation (Rynearson and Armbrust 2005, Alpermann et al. 2009, Lebret et al. 2012, Dia et al. 2014). Indeed, our results revealed such diversity by the detection of six minor MLG in addition to the two major ones. Remarkably, the different growth rates observed between December/January and February 2018/2019 isolates correlate with fluctuations in allelic frequencies in population at onset of a bloom and during the course of a bloom as determined on seawater samples, suggesting that best seasonal MLG may become dominant at specific times of the year. Similarly, temporal succession of two genetically distinct sub-populations was observed during the bloom of the haploid *Alexandrium dinoflagellate* in Gulf of Maine (Erdner et al., 2011).

Structural variants as markers to follow intraspecies diversity in environmental samples

485 We aimed to develop a rapid and cost effective alternative which did not rely on isolated
486 individuals since it is very challenging and time consuming to isolate marine microalgae from
487 complex microbial communities. Compared to short read sequencings, the recent ONT and
488 PACBIO sequencing technology provided information on structural variants, in particular on
489 relatively large INDEL and on repeated/low complexity sequences, with some of which
490 previously overlooked (Wellenreuther et al. 2019). This knowledge was particularly useful to
491 develop a novel type of marker for assessing intraspecies diversity that will rely only on PCR
492 amplification of variable size fragment without the need of sequencing. Furthermore, with
493 INDEL based on ONT sequencing we are reaching a higher level of population structure
494 compared to microsatellite or SNP markers. Structural variants are expected to be less neutral
495 and more stable than microsatellite markers (Mérot et al. 2020). Microsatellite markers can
496 change in clonal strains of *P. multistriata* in the laboratory over several months (Ruggiero et
497 al. 2018), while our structural markers are still identical in the reference genome published in
498 2012 and its clonal strain RCC4222-BANYULS sequenced in 2018. Thus INDEL markers
499 can identify large subpopulations rather than small groups of individuals. Our studies on three
500 successive blooms in Banyuls showed that INDEL markers are capable of determining the
501 dominant allelic variants and their perennial occurrence. Similarly, INDEL markers could be
502 used as query on metagenomic databases for a wider analysis of *Bathycoccus* populations.

503 Variations in allele frequencies were observed for three consecutive years, raising the
504 question of the nature of the highly reproducible yearly occurrence of *Bathycoccus* in the bay
505 of Banyuls. Seasonal blooms may result either from re-activation of "dormant/survivor" cells
506 from the water column (whose genetic fingerprint will determine the genetic profile of the
507 next bloom) or by yearly *de novo* seeding by cells carried by the north Mediterranean current
508 along the gulf of Lion. At first glance, our preliminary results are in favour of the introduction
509 of a new population rather than "resurrection" of cells from the previous bloom since allelic

510 frequencies are distinct between the end of a bloom and the onset of the next (Figure 5). By
511 monitoring the temporal population structures of the dinoflagellate *Alexandrium minutum* in
512 two estuaries in France, Dia et al. (2014) showed that interannual genetic differentiation was
513 greater than intra-bloom differentiation. Alternation of genotypes/populations has also been
514 observed with diatoms in the dominance of one of the two sympatric populations of
515 *Pseudonitzschia multistriata* which could be due either to environmental factors favouring
516 one population over the other or intrinsic factors coupled to the obligate sexual life cycle of *P.*
517 *multistriata* (D'Alelio et al. 2010). Thus the observed fluctuations in allele frequencies could
518 equally be the result of new inoculum from currents or sexual reproduction. Even though
519 sexual reproduction has not been demonstrated in *Bathycoccus*, there is genomic evidence
520 that it may occur (Benites et al. 2021). Sexual recombination generates new combinations of
521 alleles, whereas clonality favours the spread of the fittest genotype through the entire
522 population (Dia et al. 2014). Erdner et al. (2011) propose for *A. fundyense* that mitosis is the
523 primary mode of multiplication during blooms whereas mating is triggered presumably in
524 response to unfavourable conditions at the end of blooms, with vegetative cells not
525 overwintering in the water column. In Banyuls bay, the abundance during the bloom is
526 followed by severe bottlenecks in which *Bathycoccus* are hardly detected in the water column
527 (Lambert et al. 2019).
528 Knowing that (1) *Bathycoccus* blooms are followed by severe bottlenecks between one bloom
529 and the next, (2) allelic frequencies were different at the end of one bloom and at the onset of
530 the next and (3) structural markers are very stable in mitotic dividing cells, the hypothesis of
531 rare vegetative cells remaining in the water column between the blooms is unlikely except if
532 those remaining cells were produced by sexual reproduction. The alternative hypothesis of
533 new strains brought by current is equally probable.
534 ***Structural variants versus functional variants***

Our principal interest was to identify markers of intraspecies diversity in order to follow the dynamics of *Bathycoccus* population during annual blooms in the bay of Banyuls. However structural variants are probably not neutral unlike microsatellite markers. Most of the selected markers could also represent functionally significant variants such as an additional gene function or a modified promoter activity or protein function, that could correspond to an adaptation to contrasted intra- and inter-annual variations in environmental parameters in the Banyuls Bay (Lambert et al., 2021).

Remarkably in just five world-wide INDEL, we discovered two additional genes in the genome of *Bathycoccus prasinos* and a particular protein structure. The additional ANK repeat encoding gene in chromosome 1 has probably arisen by gene duplication or gene loss since it belongs to a multigenic family. The origin of the AdoMTase could be the result of Horizontal Gene Transfer. The Flavodoxin-like protein has an organisation specific to *Bathycoccus* with a coiled-coil domain of variable size with similarity to Eukaryotes parasites and toxic bacteria proteins and a flavodoxin domain found in the 7 other *Bathycoccus* flavodoxin-like proteins. *Bathycoccus* culture strains do not possess a flavodoxin per se while it was found in uncultured *Bathycoccus* (Pierella Karlusich et al. 2015). This peculiar flavodoxin-like protein could represent a case of neofunctionalisation in *Bathycoccus*. The core promoter of the central circadian clock *TOC1* gene has a conserved evening element like box (EEL box) that has been experimentally demonstrated as essential in the central oscillator of *Ostreococcus tauri* (Corellou et al. 2009). Although the EEL box is found in all the strains sequenced, the distance between the *cis* element and the initiation codon is variable. In addition, the phylogenetic tree of the core promoter sequences clearly discriminated BAFFIN and ROSCOFF and to a lesser extent, OSLO and NAPLES (Suppl. data 3B). An insertion of the AdoMTase was found about 100 bp upstream the EEL box. This insertion could potentially modify the promoter activity and ultimately the expression pattern of TOC1. Such

a natural variation of promoter length modulates the photoperiodic response of FLOWERING LOCUS T by differentially spacing two interdependent regulatory regions (Liu et al. 2014). Although the presence of the AdoMTase was not correlated with the latitude or the temperature in the Ocean Gene Atlas (Figure 2), it could still be associated with a seasonal niche. Less information is available for the promoter and function of the yrdC gene. The rearrangements are more drastic, especially with the displacement of the EEL box by insertion or deletion, and could lead to the inactivation of the promoter. The most striking feature concerns the TIM protein where only one third of the protein is conserved between TIMa and TIMb. Due to its position in the BOC of chromosome 14 putatively involved in mating, this raises the question of the mating types of cells with genome containing a TIMa or a TIMb.

Taken together, these data suggest that some INDEL markers identified in this study may be potentially involved in adaptation to changing environmental conditions.

Conclusions and perspectives

In this paper we described the conception and construction of a new type of diversity marker based on genomic structural variants. After validation on freshly isolated individuals, INDEL markers were used *in situ* on environmental samples. This pioneer study on *Bathycoccus* diversity in the bay of Banyuls now paves the way to an in depth analysis of multiple markers present in more than a decade of bimonthly sampled metagenomic data at a discrete location (Lambert et al. 2019; Lambert et al. 2021). The sequencing of the whole genomes of the different MLG, together with the assessment of their physiological performances will bring additional information contributing to the local diversity of *Bathycoccus* and provide insight into their seasonal pattern of abundance. In addition, the INDEL markers represent an essential tool for grasping the maximum diversity of newly isolated *Bathycoccus* and to identify putative molecular mechanisms involved in adaptation to environmental niches of this cosmopolitan genus.

Acknowledgments

We are grateful to the captain and the crew of the RV 'Nereis II' for their help in acquiring the samples. Additional ONT were performed with the help of Christel Llauro and Marie Mirouze LGDP. The authors acknowledge the ISO 9001 certified IRD itrop 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). The work was financed by an internal LOMIC Microprojet and the ANR Climaclock 2020-2024. We thank Thomas Roscoe for critically reading the manuscript.

References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20(1):224. doi: 10.1186/s13059-019-1829-6.

Alpermann TJ, Beszteri B, John U, Tillmann U, Cembella AD. (2009) Implications of life-history transitions on the population genetic structure of the toxigenic marine dinoflagellate Alexandrium tamarense. *Molecular Ecology,* 18(10):2122-33. doi: 10.1111/j.1365-294X.2009.04165.x.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genet*ics, 17(2):81-92. doi: 10.1038/nrg.2015.28.

Bachy C, Yung CCM, Needham DM, Gazitúa MC, Roux S, Limardo AJ, Choi CJ, Jorgens DM, Sullivan MB, Worden AZ. (2021) Viruses infecting a warm water picoeukaryote shed light on spatial co-occurrence dynamics of marine viruses and their hosts. *The ISME Journal*, doi: 10.1038/s41396-021-00989-9.

609 Bendif EM, Probert I, Archontikis OA, Young JR, Beaufort L, Rickaby RE, Filatov D. (2023)
610 Rapid diversification underlying the global dominance of a cosmopolitan phytoplankton.
611 ISME J. 17(4):630-640. doi: 10.1038/s41396-023-01365-5.
612 Benites LF, Bucchini F, Sanchez-Brosseau S, Grimsley N, Vandepoele K, Piganeau G. (2021)
613 Evolutionary Genomics of Sex-Related Chromosomes at the Base of the Green Lineage.
614 *Genome Biology Evolution*, 13(10):evab216. doi: 10.1093/gbe/evab216.
615 Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, Schackwitz W,
616 Kuo A, Salin G, Donnadieu C, Desdevises Y, Sanchez-Ferandin S, Moreau H, Rivals E,
617 Grigoriev IV, Grimsley N, Eyre-Walker A, Piganeau G. (2017) Population genomics of
618 picophytoplankton unveils novel chromosome hypervariability. *Science Advances*,
619 3(7):e1700239. doi: 10.1126/sciadv.1700239.
620 Corellou F, Schwartz C, Motta JP, Djouani-Tahri el B, Sanchez F, Bouget FY. (2009) Clocks
621 in the green lineage: comparative functional analysis of the circadian architecture of the
622 picoeukaryote *Ostreococcus*. *The Plant Cell*, 21(11):3436-49. doi: 10.1105/tpc.109.068825.
623 D'Alelio D, Ribera d'Alcala M, Dubroca L, Sarno D, Zingone A, Montresor M (2010) The
624 time for sex: A biennial life cycle in a marine planktonic diatom *Limnol.ogy Oceanography*,
625 55(1) 106–114.
626 Da Silva O, Ayata SD, Ser-Giacomi E, Leconte J, Pelletier E, Fauvelot C, Madoui MA, Guidi
627 L, Lombard F, Bittner L (2022) Genomic differentiation of three pico-phytoplankton species
628 in the Mediterranean Sea. *Environmental Microbiology*, Aug 16. doi: 10.1111/1462-
629 2920.16171.
630 Debladis E, Llauro C, Carpentier MC, Mirouze M, Panaud O. (2017) Detection of active
631 transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing
632 technology. *BMC Genomics*, 18(1):537. doi: 10.1186/s12864-017-3753-z.

633 De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. (2018) NanoPack:
634 visualizing and processing long-read sequencing data. *Bioinformatics* 34(15):2666-2669. doi:
635 10.1093/bioinformatics/bty149.

636 De Meester L. (1996) Evolutionary potential and local genetic differentiation in a
637 phenotypically plastic trait of a cyclical ârthenogen Daphnia Magna. *Evolution*, 50(3):1293-
638 1298. doi: 10.1111/j.1558-5646.1996.tb02369.x.

639 Des Roches S, Post DM, Turley NE, Bailey JK, Hendry AP, Kinnison MT, Schweitzer JA,
640 Palkovacs EP (2018) The ecological importance of intraspecific variation. *Nature Ecology*
641 *Evolution*, 2(1):57-64. doi: 10.1038/s41559-017-0402-5.

642 Dia A, Guillou L, Mauger S, Bigeard E, Marie D, Valero M, Destombe C. (2014)
643 Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic
644 dinoflagellate Alexandrium minutum. *Molecular Ecology,* 23(3):549-60. doi:
645 10.1111/mec.12617.

646 Erdner DL, Richlen M, McCauley LA, Anderson DM. (2011) Diversity and dynamics of a
647 widespread bloom of the toxic dinoflagellate Alexandrium fundyense *PLoS One*,
648 6(7):e22965. doi: 10.1371/journal.pone.0022965.

649 Gallagher JC (1980) Population genetics of Skeletonema costatum (Bacillariophyceae) in
650 Narragansett bay. *Journal of Phycology*, 16, 464-474 doi.org/10.1111/j.1529-
651 8817.1980.tb03061.x

652 Galtier N, Nabholz B, Glémin S, Hurst GD. (2009) Mitochondrial DNA as a marker of
653 molecular diversity: a reappraisal. *Molecular Ecology,* 18(22):4541-50. doi: 10.1111/j.1365-
654 294X.2009.04380.x.

655 Godhe A, Rynearson T. (2017) The role of intraspecific variation in the ecological and
656 evolutionary success of diatoms in changing environments. *Philosophical Transaction of the*

Royal Society London B Biological Sciences, 5;372(1728):20160399. doi: 10.1098/rstb.2016.0399.

Guyon JB, Vergé V, Schatt P, Lozano JC, Liennard M, Bouget FY. (2018) Comparative Analysis of Culture Conditions for the Optimization of Carotenoid Production in Several Strains of the Picoeukaryote *Ostreococcus*. *Marine Drugs*, 16(3):76. doi: 10.3390/md16030076.

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. *The ISME Journal*, 11(6):1372-1385. doi: 10.1038/ismej.2017.7.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*, 344(6182):416-20. doi: 10.1126/science.1248575.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 37(5):540-546. doi: 10.1038/s41587-019-0072-8.

Lambert S, Tragin M, Lozano JC, Ghiglione JF, Vaulot D, Bouget FY, Galand PE. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, 13(2):388-401. doi: 10.1038/s41396-018-0281-z.

Lambert S, Lozano JC, Bouget FY, Galand PE. (2021) Seasonal marine microorganisms change neighbours under contrasting environmental conditions. *Environmental Microbiology*, 23(5):2592-2604. doi: 10.1111/1462-2920.15482. Epub 2021

Lebret K, Kritzberg ES, Figueroa R, Rengefors K. Environ Microbiol. (2012) Genetic diversity within and genetic differentiation between blooms of a microalgal species. *Environmental Microbiology*, 14(9):2395-404. doi: 10.1111/j.1462-2920.2012.02769.x.

682 Leconte J, Benites LF, Vannier T, Wincker P, Piganeau G, Jaillon O. (2020) Genome
683 Resolved Biogeography of Mamiellales. *Genes (Basel)*, 11(1):66. doi:
684 10.3390/genes11010066.
685 Lewis R.J., Jensen S.I., DeNicola D.M., Miller VI, Hoagland K and Ernst SG (1997) Genetic
686 variation in the diatom*Fragilaria capucina* (*Fragilariaceae*) along a latitudinal gradient
687 across North America. *Pl Syst Evol* **204:** 99–108 . https://doi.org/10.1007/BF00982534

688 Li H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics*
689 37(23):4572-4. doi: 10.1093/bioinformatics/btab705.

690 Li WK, Rao DV, Harrison WG, Smith JC, Cullen JJ, Irwin B, Platt T. (1983) Autotrophic
691 picoplankton in the tropical ocean. *Science*, **219**: 292-5.

692 Limardo AJ, Sudek S, Choi CJ, Poirier C, Rii YM, Blum M, Roth R, Goodenough U, Church
693 MJ, Worden AZ. (2017) Quantitative biogeography of picoprasinophytes establishes ecotype
694 distributions and significant contributions to marine phytoplankton. *Environmental*
695 *Microbiology*, 19(8):3219-3234. doi: 10.1111/1462-2920.13812.
696 Liu L, Adrian J, Pankin A, Hu J, Dong X, von Korff M, Turck F. (2014) Induced and natural
697 variation of promoter length modulates the photoperiodic response of FLOWERING LOCUS
698 T. *Nature Communications*, 5:4558. doi: 10.1038/ncomms5558.
699 Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics.
700 (2019) *Frontiers in Genetics*, 10:426. doi: 10.3389/fgene.2019.00426.
701 Mérot C, Oomen RA, Tigano A, Wellenreuther M. (2020) A Roadmap for Understanding the
702 Evolutionary Significance of Structural Genomic Variation. Trends Ecol Evol. 35(7):561-
703 572. doi: 10.1016/j.tree.2020.03.002.

704 Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker
705 JR. (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore
706 flow cell. *Nature Communications*, 9(1):541. doi: 10.1038/s41467-018-03016-2.

707 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. (2018) Versatile genome
708 assembly evaluation with QUAST-LG. *Bioinformatics* 34(13):i142-i150. doi:
709 10.1093/bioinformatics/bty266.

710 Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, Van Bel M, Poulain
711 J, Katinka M, Hohmann-Marriott MF, Piganeau G, Rouzé P, Da Silva C, Wincker P, Van de
712 Peer Y, Vandepoele K. (2012) Gene functionalities and genome structure in *Bathycoccus*
713 prasinos reflect cellular specializations at the base of the green lineage. *Genome Biology*,
714 13(8):R74. doi: 10.1186/gb-2012-13-8-r74.

715 Pierella Karlusich JJ, Ceccoli RD, Graña M, Romero H, Carrillo N. (2015) Environmental
716 selection pressures related to iron utilization are involved in the loss of the flavodoxin gene
717 from the plant genome. *Genome Biology and Evolution*, 7(3):750-67. doi:
718 10.1093/gbe/evv031.

719 Raffard A, Santoul F, Cucherousset J, Blanchet S. (2019) The community and ecosystem
720 consequences of intraspecific diversity: a meta-analysis. *Biological Reviews of the Cambridge*
721 *Philosophical Society*, 94(2):648-661. doi: 10.1111/brv.12472.

722 Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A,
723 Salamov A, Young J, Aguilar M, Claverie JM, Frickenhaus S, Gonzalez K, Herman EK, Lin
724 YC, Napier J, Ogata H, Sarno AF, Shmutz J, Schroeder D, de Vargas C, Verret F, von
725 Dassow P, Valentin K, Van de Peer Y, Wheeler G; Emiliania huxleyi Annotation Consortium,
726 Dacks JB, Delwiche CF, Dyhrman ST, Glöckner G, John U, Richards T, Worden AZ, Zhang
727 X, Grigoriev IV. Pan genome of the phytoplankton Emiliania underpins its global
728 distribution. (2013) *Nature*, 499(7457):209-13. doi: 10.1038/nature12221.

729 Rengefors K, Kremp A, Reusch TBH and Wood AM (2017) Genetic diversity and evolution
730 ineukaryotic phytoplankton: revelations from population genetic studies. *Journal of Plankton*
731 *Research*, (2017) 39(2): 165– 179.
732 Richter DJ, Watteaux R, Vannier T, Leconte J, Frémont P, Reygondeau G, Maillet N, Henry
733 N, Benoit G, Da Silva O, Delmont TO, Fernàndez-Guerra A, Suweis S, Narci R, Berney C,
734 Eveillard D, Gavory F, Guidi L, Labadie K, Mahieu E, Poulain J, Romac S, Roux S, Dimier
735 C, Kandels S, Picheral M, Searson S; Tara Oceans Coordinators, Pesant S, Aury JM, Brum
736 JR, Lemaitre C, Pelletier E, Bork P, Sunagawa S, Lombard F, Karp-Boss L, Bowler C,
737 Sullivan MB, Karsenti E, Mariadassou M, Probert I, Peterlongo P, Wincker P, de Vargas C,
738 Ribera d'Alcalà M, Iudicone D, Jaillon O. (2022) Genomic evidence for global ocean
739 plankton biogeography shaped by large-scale current systems. *Elife*, 11:e78129. doi:
740 10.7554/eLife.78129.
741 Ruggiero MV, D'Alelio D, Ferrante MI, Santoro M, Vitale L, Procaccini G, Montresor M.
742 (2018) Clonal expansion behind a marine diatom bloom. *The ISME Journal*, 2018
743 Feb;12(2):463-472. doi: 10.1038/ismej.2017.181.
744 Rynearson TA, Armbrust EV. (2005) Maintenance of clonal diversity during a spring bloom
745 of the centric diatom Ditylum brightwellii. *Molecular Ecology*, 14(6):1631-40. doi:
746 10.1111/j.1365-294X.2005.02526.x.
747 Rynearson TA, Bishop IW, Collins S (2022) The population Genetics and Evoluionary
748 Potential of Diatoms in The Molecular Life of Diatoms (pp.29-57) eds Falciatore and Mock,
749 Springer.
750 Sehgal A, Rothenfluh-Hilfiker A, Hunter-Ensor M, Chen Y, Myers MP, Young MW (1995).
751 "Rhythmic expression of timeless: a basis for promoting circadian cycles in period gene
752 autoregulation". *Science*, 270 (5237): 808–10. doi:10.1126/science.270.5237.808.

753 Simmons MP, Sudek S, Monier A, Limardo AJ, Jimenez V, Perle CR, Elrod VA, Pennington
754 JT, Worden AZ. (2016) Abundance and Biogeography of Picoprasinophyte Ecotypes and
755 Other Phytoplankton in the Eastern North Pacific Ocean. *Applied Environmental*
756 *Microbiology*, 82(6):1693-1705. doi: 10.1128/AEM.02730-15.
757 Srivastava S, Avvaru AK, Sowpati DT, Mishra RK. (2019) Patterns of microsatellite
758 distribution across eukaryotic genomes. *BMC Genomics*, 20(1):153. doi: 10.1186/s12864-
759 019-5516-5.
760 Teplova M, Tereshko V, Sanishvili R, Joachimiak A, Bushueva T, Anderson WF, Egli M.
761 (2000) The structure of the yrdC gene product from Escherichia coli reveals a new fold and
762 suggests a role in RNA binding. *Protein Science*, 9(12):2557-66. doi: 10.1110/ps.9.12.2557.
763 Tesson SV, Borra M, Kooistra WH, Procaccini G. (2011) Microsatellite primers in the
764 planktonic diatom Pseudo-nitzschia multistriata (Bacillariophyceae). *American Journal of*
765 *Botany*, 98(2):e33-5.
766 Tesson SV, Montresor M, Procaccini G, Kooistra WH. (2014) Temporal changes in
767 population structure of a marine planktonic diatom. *PLoS One*, 9(12):e114984. doi:
768 10.1371/journal.pone.0114984.
769 Tragin M, Zingone A, Vaulot D (2018) Comparison of coastal phytoplankton composition
770 estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic
771 groups and especially Chlorophyta. *Environmental Microbiology*, 20(2):506-520.
772 Van den Wyngaert S, Möst M, Freimann R, Ibelings BW, Spaak P. (2015) Hidden diversity
773 in the freshwater planktonic diatom Asterionella formosa. *Molecular Ecology*, 24(12):2955-
774 72. doi: 10.1111/mec.13218.
775 Vannier T, Leconte J, Seeleuthner Y, Mondy S, Pelletier E, Aury JM, de Vargas C, Sieracki
776 M, Iudicone D, Vaulot D, Wincker P, Jaillon O. . (2016) Survey of the green picoalga

Bathycoccus genomes in the global ocean*Scientific Reports*, 6:37900. doi: 10.1038/srep37900.

Vaser R, Sović I, Nagarajan N, Šikić M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27(5):737-746. doi: 10.1101/gr.214270.116.

Vernette C, Lecubin J, Sánchez P; Tara Oceans Coordinators, Sunagawa S, Delmont TO, Acinas SG, Pelletier E, Hingamp P, Lescot M. (2022) The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Research*, 50(W1):W516-26. doi: 10.1093/nar/gkac420

.Villar E, Vannier T, Vernette C, Lescot M, Cuenca M, Alexandre A, Bachelerie P, Rosnet T, Pelletier E, Sunagawa S, Hingamp P (2018) The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Research*, doi: 10.1093/nar/gky376.

Violle C, Enquist BJ, McGill BJ, Jiang L, Albert CH, Hulshof C, Jung V, Messier J. (2012) The return of the variance: intraspecific variability in community ecology. *Trends in Ecology and Evol.ution*, 27(4):244-52. doi: 10.1016/j.tree.2011.11.014.

Wellenreuther M, Mérot C, Berdan E, Bernatchez L (2019) Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28(6):1203-1209. doi: 10.1111/mec.15066.

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE. (2014) A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. *Applications in Plant Sciences*, 2(12):apps.1400059. doi: 10.3732/apps.1400059.

Data accessibility

Strains isolated in Banyuls have been sent to Roscoff Collection Centre and will be publicly available after curation by the Centre. For whole genome sequences, basecalled reads and

RaGOO outputs of the 7 RCC strains used in this study are available at Zenodo https://doi.org/10.5281/zenodo.7594933

Author contribution

MD, FYB and FS conceived the work and acquired funding. MD extracted high molecular weight genomic DNA. ONT sequencing was performed by CM and MD and sequence analysis by LD and FS. Diversity markers were designed and validated by MD. PS analysed seawater samplings by flow cytometry. MD and VV isolated Banyuls strains during the winter bloom, genotyped by MD and JCL. MD determined diversity of seawater. MD and FYB wrote the article and all authors participated in critical revisions and approved the final version for submission.

Table 1. Number and sizes of INDEL

Strain	Name	Insertion		Insert 0.2-2kb	deletion		Coverage
		number	size range	number	number	size range	
RCC5417	BAFFIN	101	36-15979 bp	36	80	50-19025bp	x9
RCC1613	OSLO	149	51-18972 bp	37	67	50-13730 bp	x239
RCC685	HELGOLAND	116	51-16349 bp	60	69	51-19026 bp	x52
RCC1615	DIEPPE	62	72-4620 bp	23	69	73-13730 bp	x4
RCC1868	ROSCOFF	49	50-5509 bp	22	43	51-13732 bp	x19
RCC4222	BANYULS	14	76-3011 bp	11	5	51-3999 bp	x30
RCC4752	NAPLES	53	50-15411 bp	34	60	50-13730 bp	x14
mean*		88		35			
*without RCC4222							
INDEL within aligment							

812

Table 2. Distribution of diversity markers in world-wide strains

RCC	Name	yrcD prom chr1	TOC1 prom chr17	Flavodoxin chr3	Zinc finger chr15	TIMa chr14
primers		MDB33+MDB34	MDB7+MDB9	MDB40+MDB41	MDB68+MDB69	MDB57+MDB58
RCC4222	BANYULS	400 bp	700 bp	800 bp	730 bp	530bp
RCC5417	BAFFIN	1400 bp	700 bp	600 bp	900 bp	530bp
RCC1613	OSLO	1400 bp	2.2 kb	1000 bp	730 bp	530bp
RCC685	HELGOLAND	1400 bp	700 bp	1200 bp	820 bp	530bp
RCC1615	DIEPPE	400 bp	2.2 kb	1400 bp	730 bp	530bp
RCC1868	ROSCOFF	200 bp	700 bp	800 bp	730 bp	530bp
RCC4752	NAPLES	1400 bp	2.2 kb	1200 bp	730 bp	530bp
		3 size variants	2 size variants	5 size variants	3 size variants	1 size variant
		14.28% (200 bp)	42.86% (2.2 kb)	28.57% (1.2 kb)	71.5% (730 bp)	100% (530 bp)

814

Table 3. Multi Loci Genotypes (MLG) of Banyuls isolates

Strains	**INDEL marker**							
	Chr1	Chr17	Chr3	Chr15	chr14			
	yrcD prom	TOC1 prom	flavodoxin	zinc finger	TIMa			
4222-BANYULS	400bp	700bp	800bp	730 bp	530bp			
4752-NAPLES	1.4kb	2.2kb	1200bp	730 bp	530bp			
						Numer of isolates	Occurance in samplings	Percentage of isolates
Banyuls 18/19								
MLG 1	200bp	2.2 kb	1200bp	730 bp	530bp	29	5/9 samplings	53%
MLG 2	200bp	2.2kb	1200bp	730 bp	0	16	5/9 samplings	29%
MLG 3	200bp	2.2 kb	800bp	730 bp	530bp	4	3/9 samplings	8%
MLG 4	200bp	700 bp	1200bp	730 bp	530bp	2	2/9 samplings	4%
MLG 5	200bp	2.2kb	1200bp	820 bp	0	1	1/9 samplings	2%
MLG 6	200bp	700 bp	1200bp	820 bp	0	1	1/9 samplings	2%
MLG 7	400bp	700 bp	1400bp	820 bp	530bp	1	1/9 samplings	2%
MLG 8	200bp	2.2kb	1600bp	730 bp	530bp	1	1/9 samplings	2%
Mediterranean								
Size number	2*	2	4	2	2			
Frequency	98% (200bp)	91% (2.2 kb)	89% (1.2kb)	95% (730 bp)	66% (530 bp)			
World-wide								
Size number	3	2	4	3	1			
Frequency	14.28% (200 bp)	42.86% (2.2 kb)	14.28% (1.2 kb)	71.5% (730 bp)	100% (530 bp)			
* no insertion of ANK gene tested with MDB33+MDB35								
similar to RCC4222-BANYULS								
similar to RCC 4752-NAPLES								

815

Figure 1. Diversity markers among world-wide strains

Samples are arranged on a latitudinal gradient from Baffin to Naples. Arrows indicate the size of the various fragments. A: Size variation in intergenic region in the promoters of yrdC on chromosome 1 and TOC1 on chromosome 17. B: size variation in nucleotidic sequences encoding amino acid repeats in a Flavodoxin-like gene on chromosome 3 and a zinc-finger protein on chromosome 15. C: detection of the nucleotidic sequence encoding the specific C terminal region of TIMa in world-wide strains on the left and in seven *Bathycoccus* isolates from Banyuls during winter 2018/2019 (B1-B7) on the right.

Figure 2. Geographical distribution of TOC1 and AdoMTase proteins

The protein sequences of TOC1 and AdoMTase were used as a query at high stringency in OGA. Their abundance was expressed as percent of total reads. Under these conditions, a single hit was found and its presence and abundance are represented on the world map at surface water (SRF) and deep chlorophyll maximum layer (DCM) and in relation to latitude and temperature. The arrows point to a station at the Chilean coast where TOC1 is present but not AdoMTase.

Figure 3. Abundance of phytoplankton during three successive blooms

Seawater at SOLA buoy (Banyuls) was sampled at a depth of 3 meter. After passage on 3 μm filter, the flow through was analysed by flow cytometry. At each sampling time, the phytoplankton was categorised and quantified in function of cell size (pico- and nano-phytoplankton, cyano-bacteria) with indication of the seawater temperature. The main peak of picophytoplankton abundance was in December 3th 2018 and February 19th 2019, January 7th and February 11th 2020, January 27th and March 2nd 2021. The most striking difference between the three years was the sudden abundance in nanophytoplankton in March 2021.

Figure 4. Distribution and abundance of TIM variant proteins in MetaG database

839 A. Phylogenetic tree of TIM proteins from world-wide and Banyuls isolates presenting two
840 main clades each containing one TARA OGA hit, OGATIMa or OGATIMb. OGATIMa was
841 retrieved after a query with the TIM protein from RCC4222 and OGATIMb with the variant
842 TIM protein in the A8 Banyuls isolate.
843 B. Geographical presence of TIM variants in OGA. The presence and abundance (percent of
844 total reads) of each variant is represented in SRF and DCM samples with the temperature and
845 latitude parameters.
846 **Figure 5**. Diversity markers in seawater
847 Results the PCR amplification of marker yrdC (chr1) and TIM (chr14) from natural sea water.
848 For clarity, only a subset of the analysis is presented in this Figure, the complete dataset is
849 presented in Supplemental Figure 3 and Supplemental Table 3.
850 Seawater was filtered in autumn from end of November in 2019 and from October 2020.
851 The time of sampling is indicated as week of the month (e.g. Oct-01 = 1st week of October) to
852 facilitate the comparison between years. The relative abundance of the 2 allelic variants of
853 yrdC (200 bp, a deletion or 400 bp, the reference type) and TIMa and b (absence or presence)
854 were recorded. ND: not determined.
855 **Figure 6.** Growth curves and rates of Banyuls isolates
856 The growth curves and rates of the December (D), January (J) and February (F1-F4)
857 *Bathycoccus* isolates were determined under 4 different conditions by sampling every day for
858 9 days. Cell concentration was determined by flow cytometry and is expressed as 10^6
859 cells/ml.
860 **Table 1.** Number and sizes of INDEL
861 **Table 2.** Distribution of diversity markers in world-wide strains
862 **Table 3.** Multi Loci Genotypes of Banyuls isolates
863 **Supplemental Table1.** Strains used in this study

Supplemental Table 2. Isolation of *Bathycoccus* strains during 2018/ 2019 winter bloom in the Banyuls bay

Supplemental Table 3. Relative abundance of diversity markers in sea water

Supplemental data 1. Sequences and Alignments

Supplemental data 2. Primers used in this study

Supplemental Figure 1. Geographical distribution of TOC1 and CCA1 (Bathy05g02420) proteins

Supplemental Figure 2. Abundance of picophytoplankton during winter blooms

Supplemental Figure 3. Images of sea water PCR amplification for yrdC, TOC1 and TIM used for Supplemental Table 3

Supplemental Figure 4. Photosynthesis parameters of Mediterranean *Bathycoccus* strains

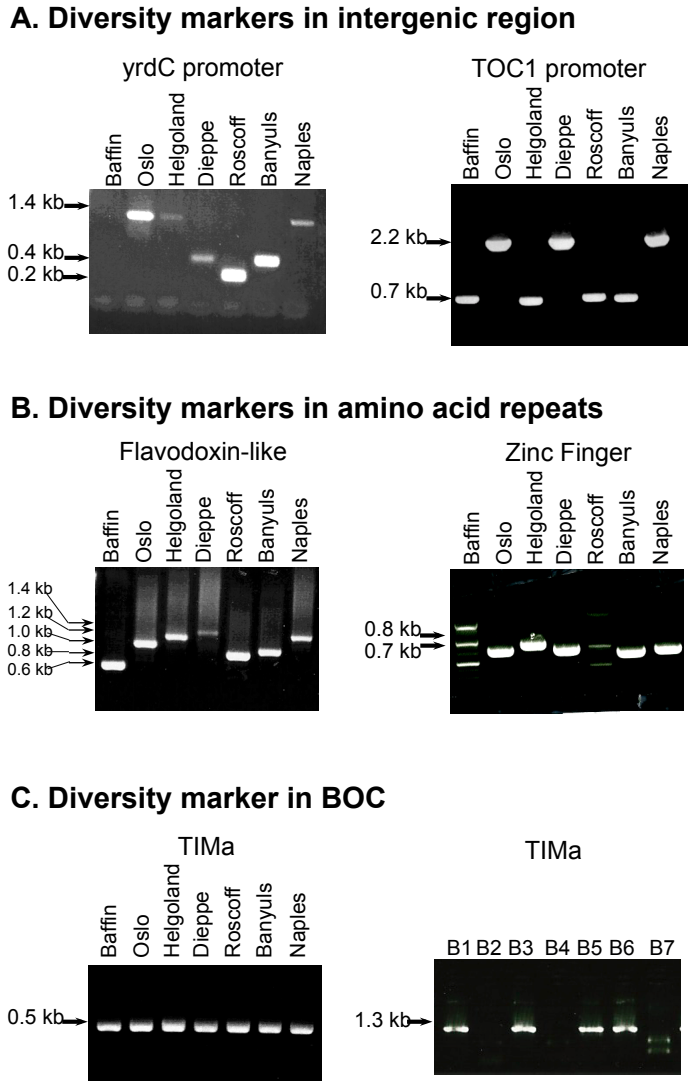


Figure 1. Diversity markers world-wide strains

875

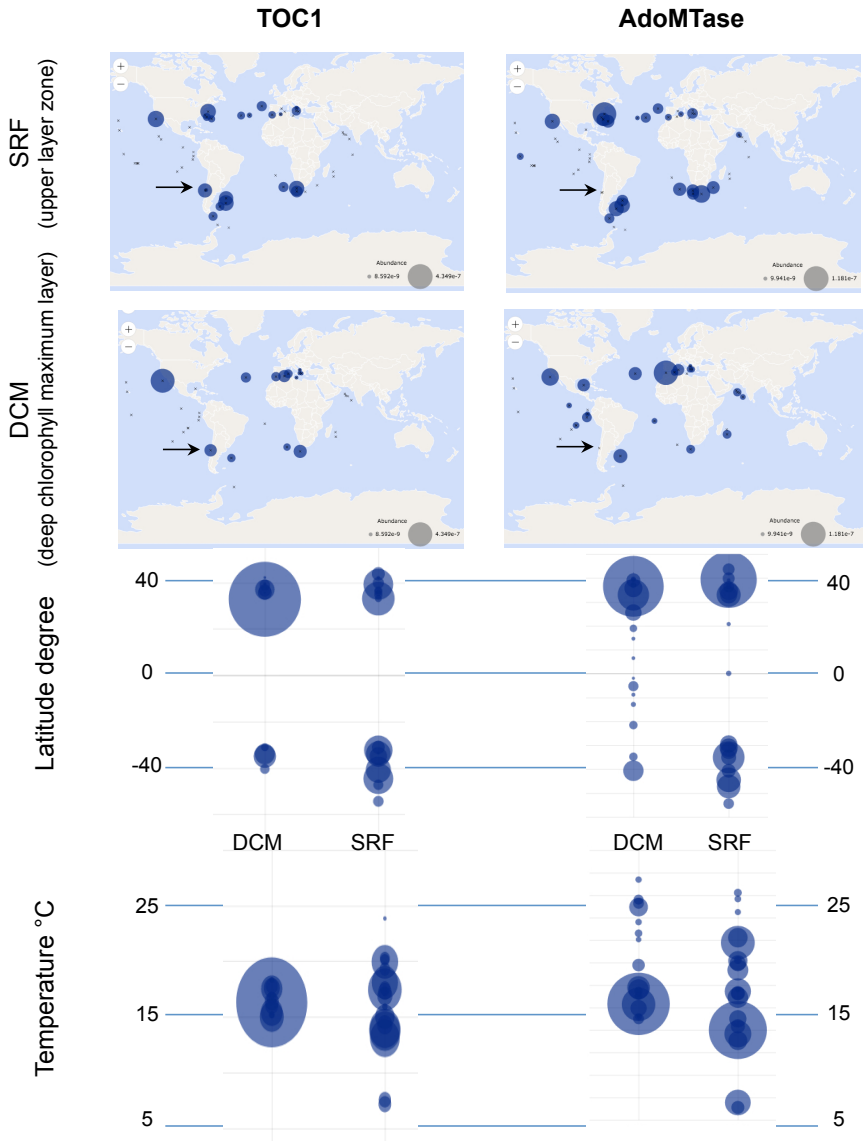


Figure 2 : Geographical distribution of TOC1 and AdoMTase proteins

876

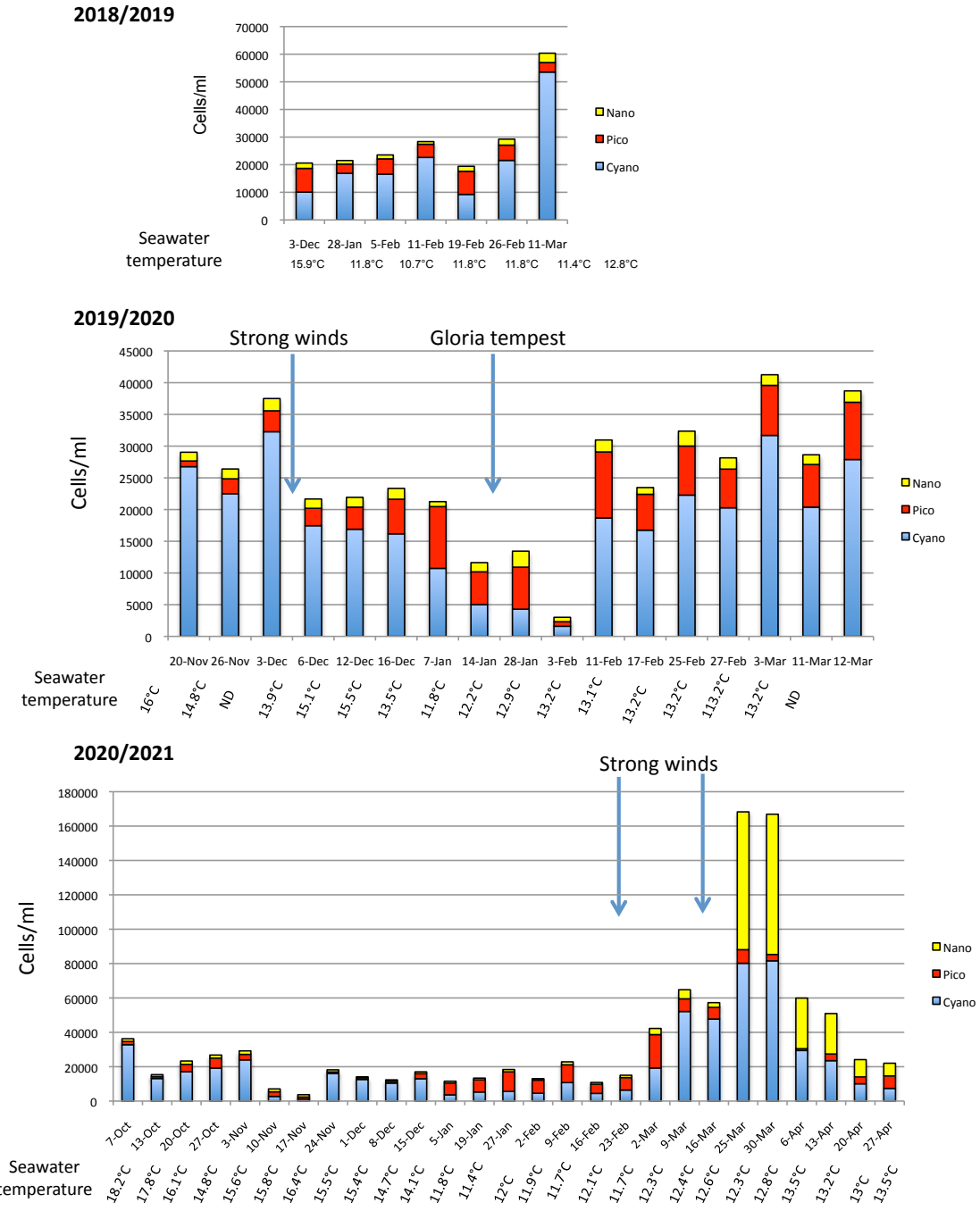


Figure 3. Abundance of phytoplankton during three successive blooms

877

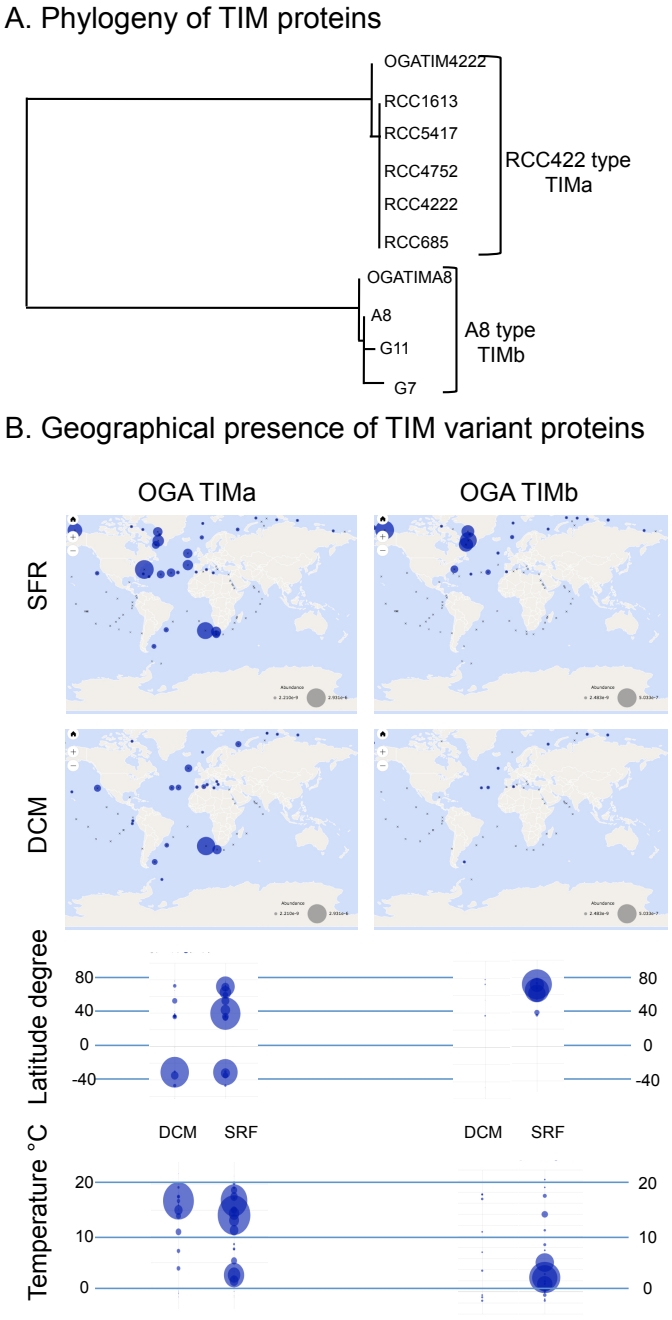
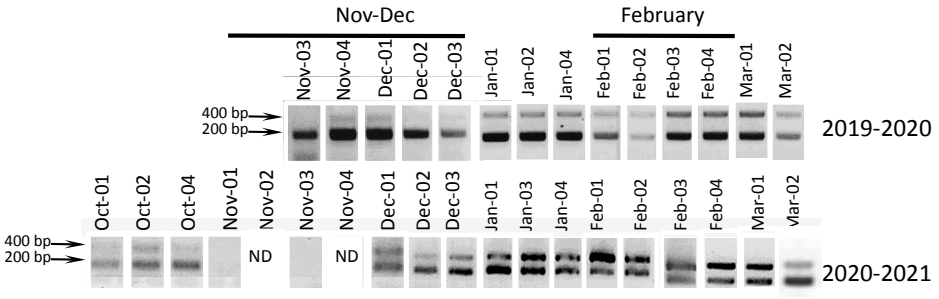


Figure 4. Distribution and abundance of TIM variant proteins in MetaG database

878

Marker chromosome 1 : yrdC promoter



Marker chromosome 14 : TIM

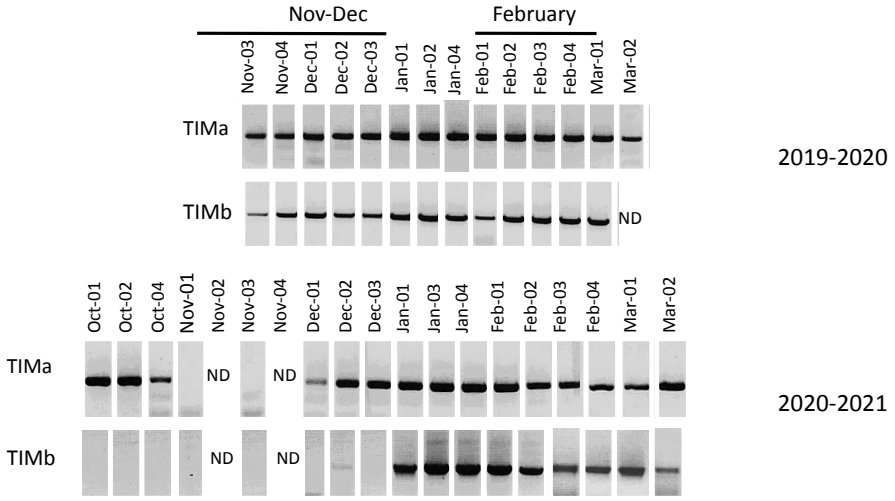


Figure 5. Diversity markers in seawater

879

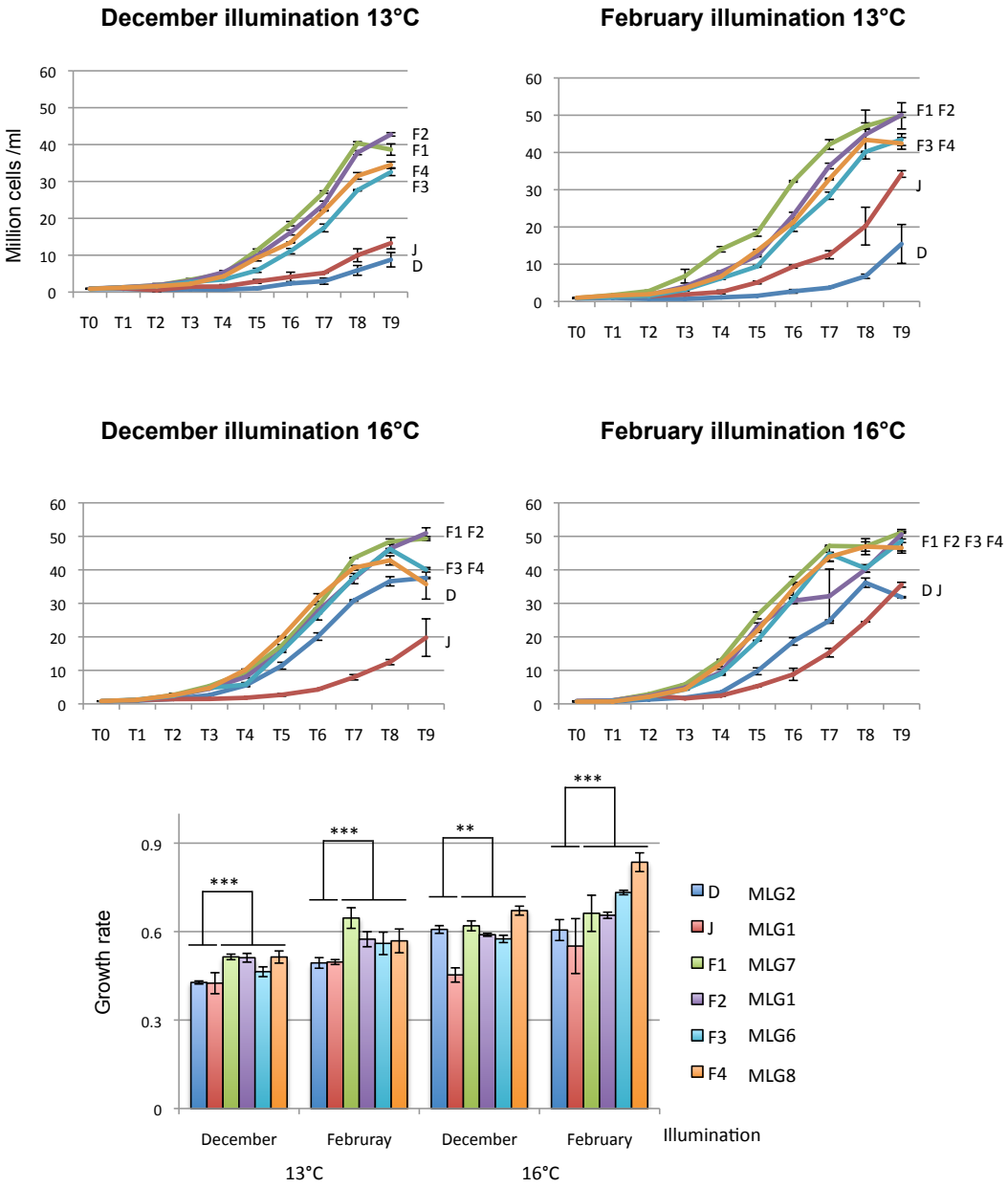


Figure 6 : Growth curves and rates of Banyuls isolates

880