Genomic adaptation of the picoeukaryote *Pelagomonas calceolata* to temperate iron-poor oceans revealed by a chromosome-scale genome sequence.

Nina Guérin[1,2], Marta Ciccarella[1], Elisa Flamant[1,2], Sophie Mangenot[1,2], Benjamin Istace[1], Benjamin Noel[1], Sarah Romac[3], Charles Bachy[3], Martin Gachenot[4], Eric Pelletier[1,2], Adriana Alberti[1,2,5], Corinne Cruaux[1], Patrick Wincker[1,2], Jean-Marc Aury[1], Quentin Carradec[1,2]*

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057, Evry, France

[2]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, R2022/Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016, Paris, France

[3]Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR7144, Place Georges Tessier, 29680 Roscoff, France

[4]Sorbonne Université, CNRS, FR2424, Station Biologique de Roscoff, 29680 Roscoff, France

[5]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

*Correspondence and requests for materials should be addressed to qcarrade@genoscope.cns.fr

Summary

Eukaryotic phytoplankton are key actors in marine ecosystems, they contribute to atmospheric CO_2 sequestration and supply organic matter to the trophic network. Among them, Pelagophytes (Stramenopiles) algae are a diverse class with coastal species causative of harmful algal blooms while others are cosmopolites and abundant in open ocean ecosystems. Despite their ecological importance, only a few genomic references exist limiting our capacity to identify them and study their adaptation mechanisms in a changing environment. Here, we report the complete chromosome-scale assembled genome sequence of *Pelagomonas calceolata*. We identified unusual large low-GC and gene-rich regions potentially hosting centromeres. These particular genomic structures could be explained by the absence of genes necessary for an important recombination pathway in this species. We identified a large repertoire of genes involved in inorganic nitrogen sensing and uptake as well as many genes replacing iron-required proteins potentially explaining its ecological success in oligotrophic waters. Finally, based on this high-quality assembly, we evaluated *P. calceolata* relative abundance in all oceans using environmental *Tara* datasets. Our results suggest that *P. calceolata* is one of the most abundant eukaryote species in the oceans with a relative abundance driven by the high temperature and iron-poor conditions. Collectively, these findings bring new insights into the biology and ecology of *P. calceolata* and lay the foundation for the analysis of the adaptation and acclimation strategy of this picophytoplankton.

Introduction

Marine phytoplankton account for more than 45% of net primary production on Earth and play an essential role in supplying organic matter to marine food webs[1]. They are key global actors in CO_2 uptake and provide gaseous oxygen to the atmosphere. A global decline of phytoplankton biomass has been reported over the past century (1% of chlorophyll-a concentration per year) leading to a decrease of net primary production in many oceanic regions[2]. This decline is probably a consequence of global ocean warming which drives water column stratification, reducing the nutrient supply to surface waters. Temperature-driven reductions in phytoplankton productivity in the tropics and temperate regions are likely to have cascading effects on higher trophic levels and ecosystem functioning[3].

Picophytoplankton are photosynthetic unicellular organisms <2 µm in cell diameter. They encompass the abundant cyanobacteria *Synechococcus* and *Prochlorococcus* and photosynthetic picoeukaryotes (PPEs) belonging to different phyla including Chlorophyta, Cryptophyta, Haptophyta and Stramenopiles[4]. PPEs are present in all oceans and the dominant primary producers in warm and oligotrophic regions. Ocean warming and expansion of oligotrophic regions in the next decades may extend the ecological niche of PPEs and a global shift from large photosynthetic organisms toward smaller primary producers is expected[3,5]. For example, sea ice melting in the Canadian Arctic Basin has been associated with an increase in the abundance of PPEs such as *Micromonas* at the expense of larger algae[6]. In the laboratory, this alga has the capacity to change its optimum temperature for growth in only a few hundred generations, which suggests that it will be less affected by global warming than many larger organisms[7]. In addition, the larger cell surface-area-to-volume ratio of PPEs compared to larger phytoplankton cells is advantageous in terms of resource acquisition used in growth in nutrient-limited environments[8,9].

Iron is one key compound required for the activity of the respiratory chain, photosynthesis and nitrogen fixation[9]. Because bioavailable iron is extremely low in more than one third of the surface ocean, small phytoplankton has developed several strategies to optimize iron uptake and reduce iron needs[10]. In diatoms, reductive and non-reductive iron uptake mechanisms involve many proteins including phytotransferrins, transmembrane ferric reductases, iron permeases and siderophore-binding proteins[11]. The iron needs can be modulated by variation of gene expression levels between iron-required proteins and their iron-free equivalent. These protein switches include genes involved in electron transfer (flavodoxin/ferredoxin and plastocyanin/cytochrome c6), in gluconeogenesis (fructose-bisphosphate aldolase type I or type II) and superoxide dismutases (Mn/Fe-SOD, Cu/Zn-SOD or Ni-SOD)[12–14].

PPE growth is also limited by nitrogen (N) availability in large portions of the world ocean[15]. Ammonium (NH_4^+), nitrate (NO_3^-) and nitrite (NO_2^-) are the main source of inorganic N for PPEs, however several studies have shown that dissolved organic N, like urea, can be metabolised in N-limited environments[16]. For example, several membrane-localized urea transporters in the diatom *Phaeodactylum tricornutum* are maximally expressed in nitrogen-limited conditions[17] and the harmful algal blooms of the pelagophyceae *Aureococcus anophagefferens* may be fuelled by urea[18].

Despite their large taxonomic distribution, most molecular studies on the ecological role of PPEs and their adaptation to the environment are restricted to a few species. PPEs are suspected to possess highly developed acclimation/adaptation capacities, but the underlying molecular mechanisms remain poorly characterized due to the lack of reference genomic data.

Among PPEs, *Pelagomonas calceolata* was the first described member of the heterokont class Pelagophyceae[19]. It has since been identified in many oceanic regions using its 18S rRNA sequence and chloroplastic genome[20–22]. Several studies have demonstrated the capacity of *P. calceolata* to adapt to different environmental conditions. In the laboratory, *P. calceolata* has been shown to exhibit a high degree of acclimation to light fluctuations with rapid activation of the photo-protective xanthophyll cycle and non-photochemical quenching[23]. In the Marquesas archipelago, *P. calceolata* is one of the most responsive species to iron fertilization with upregulation of genes involved in photosynthesis, amino acid synthesis and nitrogen assimilation. A global scale analysis of pelagophytes genes revealed that they are adapted to low iron conditions[13]. In the subtropical Pacific, *P. calceolata* expresses stress genes in surface samples and genes involved in nitrogen assimilation are overexpressed in the deep chlorophyll maximum[24]. A laboratory study suggests that *P. calceolata* also has the ability to increase the transcription levels of organic-nitrogenous compound cleavage enzymes (cathepsin, urease, arginase) under low nitrogen concentration[25]. Thus, gene expression appears to be controlled according to the nitrogen source and quantity. Taken together, this apparent adaptive plasticity may explain the presence of *P. calceolata* in many different oceanic environments, however, an exhaustive analysis of the genetic capacity of this species and the *in situ* characterization of its ecological niche is lacking.

Here we sequenced, assembled and annotated *Pelagomonas calceolata* nuclear genome, with a combination of long- and short-reads. We used this genome to examine its genomic structure and its gene content relatively to other unicellular phytoplankton. Specific analyses were performed to get new insights into its life cycle and its genetic capacity of nutrient uptake. Finally, we used this genome

to detect *P. calceolata* in environmental datasets of *Tara* expeditions across all oceans, to characterize its ecological niche and to identify the environmental conditions controlling its relative abundance.

Material and Methods

Pelagomonas culture

Pelagomonas calceolata RCC100 culture was grown in 12:12-h light:dark photoperiod in K medium with natural seawater base at 20°C. At the Roscoff Culture Collection, cells were kept at a light intensity of ~80 µmol photon m-2 s-1 and volume of culture was ramped up to 1 liter in mid-exponential growth phase before harvesting. RCC100 culture was not axenic and grown in the presence of undefined bacterial microbiota.

DNA extraction, library preparation and sequencing

We pelleted cells from 500 ml of culture by two successive centrifugations at 10,000 g for 15 minutes at 4°C. Genomic DNA was extracted using the NucleoSpin Plant II Mini kit according to the manufacturer's instructions (Macherey-Nagel, Hoerdt, France) with the following exception for the lysis step : 400µL of lysis buffer PL1 and 25µL of proteinase K 25mg/mL were added to strain pellets, and lysates were incubated at 55°C for 1 hour at 900 rpm. DNA quantity and integrity were respectively evaluated on a Qubit 2.0 spectrofluorometer (Invitrogen, Carlsbad, CA, USA) and a Nanodrop1000 spectrophotometer (Thermo Fisher Scientific, MA, USA).

For Illumina sequencing, DNA (1.5µg) was sonicated using a Covaris E220 sonicator (Covaris, Woburn, MA, USA). Fragments were end-repaired, 3'-adenylated and Illumina adapters (Bioo Scientific, Austin, TX, USA) were then added using the Kapa Hyper Prep Kit (KapaBiosystems, Wilmington, MA, USA). Ligation products were purified with AMPure XP beads (Beckmann Coulter Genomics, Danvers, MA, USA). The library was then quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems), and library profile was assessed using a High Sensitivity DNA kit on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was sequenced on an Illumina NovaSeq instrument (Illumina, San Diego, CA, USA) using 150 base-length read chemistry in a paired-end mode.

For Oxford Nanopore Technologies (ONT) sequencing, the library was prepared using the 1D Native barcoding genomic DNA (with EXP-NBD104 and SQK-LSK109). Genomic DNA fragments (1 µg) were repaired and 3'-adenylated with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Adapters with barcode provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) were then ligated using the NEB Blunt/TA Ligase Master Mix (NEB). After purification with AMPure XP beads (Beckmann

Coulter, Brea, CA, USA), the sequencing adapters (ONT) were added using the NEBNext Quick T4 DNA ligase (NEB). The library was purified with AMPure XP beads (Beckmann Coulter), then mixed with the Sequencing Buffer (ONT) and the Loading Bead (ONT) and loaded on a MinION R9.4.1 flow cell. Reads were basecalled using Guppy 3.1.5.

RNA extraction, library preparation and sequencing

When the cell concentration reached 10 million cell/mL in mid-exponential growth phase, 160 mL of culture were collected by three successive filtrations on 1.2 µm polycarbonate filters of 47mm to avoid prokaryotic contamination. To preserve cells and RNA integrity, we kept filtration time and depression below 10 min and 20 mmHg, respectively. Then filters were stored in 15 mL Falcon tubes with 3 mL of Trizol (Invitrogen, Carlsbad, CA, USA), mixed and flash-frozen in liquid nitrogen for further processing. RNA were extracted by incubation at 65°C for 15 min, followed by a chloroform extraction. Aqueous phase was purified using Purelink RNA Isolation kit (Ambion Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. DNA contamination was removed by digestion using the TURBO DNA-free™ Kit (Ambion Invitrogen) according to the manufacturer's DNase treatment protocol. After two rounds of 30 min incubation at 37 °C, the efficiency of DNase treatment was assessed by PCR. Quantity and quality of extracted RNA were analyzed with RNA-specific fluorimetric quantitation on a Qubit 2.0 Fluorometer using Qubit RNA HS Assay (Invitrogen). The qualities of total RNA were checked by capillary electrophoresis on an Agilent Bioanalyzer, using the RNA 6,000 Nano LabChip kit (Agilent Technologies, Santa Clara, CA).

RNA-Seq library preparations were carried out from 1 µg total RNA using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA), which allows mRNA strand orientation. Briefly, poly(A)+ RNAs were selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. After second strand synthesis, double-stranded cDNA was 3'-adenylated and ligated to Illumina adapters. Ligation products were PCR-amplified following the manufacturer's recommendations. Finally, ready-to-sequence Illumina library was quantified by qPCR using the KAPA Library Quantification Kit for Illumina libraries (KapaBiosystems, Wilmington, MA, USA), and evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was sequenced using 101 bp paired end reads chemistry on a HiSeq2000 Illumina sequencer. Low-quality nucleotides ($Q < 20$) from both ends of the reads were discarded. Illumina sequencing adapters and primer sequences were removed and reads shorter than 30 nucleotides after trimming were discarded. These trimming and cleaning steps were achieved using in-house-designed software based on the FastX package (https://www.genoscope.cns.fr/externe/fastxtend/). The last step identifies and discards read pairs that are mapped to the phage phiX genome, using SOAP aligner[26] and the Enterobacteria phage PhiX174 reference sequence (GenBank: NC_001422.1). This processing,

described in Alberti et al, resulted in high-quality data[27]. Moreover, ribosomal RNA-like reads were excluded using SortMeRNA[28].

Long-read based genome assembly

Raw nanopore reads were used for genome assembly. Taxonomic assignation was performed using Centrifuge 1.0.3 to detect potential contamination. Genome size and heterozygosity rate were estimated using Genomescope[29] and Illumina short-reads. For the genome assembly, we generated three sets of ONT reads: all the reads, 30x genome coverage with the longest reads and 30x genome coverage of the highest-scored reads estimated by Filtlong tool (https://github.com/rrwick/Filtlong). We then applied four different assemblers, Smartdenovo, Redbean, Flye and Ra on these three sets of reads (Table S1)[30–32]. After the assembly phase, we selected the best assembly (Flye with all reads) based on the cumulative size and contiguity. The assembler output was polished three times using Racon with Nanopore reads, and two times with Hapo-G and Illumina reads[33,34]. Gene completeness of the assembly was estimated using the single-copy orthologous gene analysis from BUSCO v5 with the stramenopile dataset version 10 containing 100 genes[35].

Repeat masking and GC analyses

Repetitive regions on the genome were masked using Tandem Repeat Finder tool[36], Dust tool to detect low complexity regions[37] and RepeatMasker[38] to identify interspersed repeats based on homology search among the Stramenopiles clade and other low complexity sequences . The positions of detected repeats were merged and hard-masked on the genome, amounting to a total of 8% of its length. *Ab initio* identification of repeat family sequences was performed using RepeatScout[39]. The algorithm first calculates the frequency of all k-mers in the genome, then removes low-complexity regions and tandem repeats. In > 80% of the cases, repeat families identified using *ab initio* approaches do not overlap with repetitive regions identified by homology search. GC content along the genome was calculated with Bedtools nuc version 2.29.2[40] and the coverage over a non-overlapping window of 2 Kb with Mosdepth version 0.2.8[41].

Transcriptome assembly

RNA sequencing reads from *P. calceolata* RCC100 were assembled using Velvet 1.2.07 and Oases 0.2.08 with a k-mer size of 63 bp[42,43]. Reads were mapped back to the contigs with BWA-mem[44] and only consistent paired-end reads were kept. Uncovered regions were detected and used to identify chimeric contigs. In addition, open reading frames (ORF) and domains were searched using respectively TransDecoder (http://transdecoder.sourceforge.net) and CDDsearch[45]. Contig extremities without predicted ORFs nor functional domains were removed. Lastly, we used the read strand information to correctly orient correctly RNA contigs. We completed the RNA contigs dataset with the

two transcriptome assemblies of the RCC100 strain of *P. calceolata* from the Marine Eukaryotes Transcriptomes database (METdb) (http://metdb.sb-roscoff.fr/metdb/)[46].

Gene prediction

Nuclear gene prediction was performed using 23,696 Pelagomonadales proteins (mainly *Aureococcus anophagefferens*) downloaded from the NCBI website. Proteins were aligned on the genome in a two-steps strategy. First, BLAT (version 36 with default parameters) was used to rapidly localize corresponding putative regions of these proteins on the genome. The best match and the matches with a score greater than or equal to 90% of the best match score were retained. Then, the regions with BLAT alignments were masked and we aligned the same set of proteins using BLAST, which is able to identify more divergent matches. Second, alignments were refined using Genewise (version 2.2.0 default parameters, except the -splice model option to detect non-canonical splicing sites), which is more accurate for detecting intron/exon boundaries. Alignments were kept if more than 50% of the length of the protein is aligned on the genome. Additionally, the transcriptome assemblies of *P. calceolata* RCC969, RCC2362, RCC706 and RCC981 included in the METdb were translated into proteins and aligned to the genome using BLAT, a BLAT score > 50 % filter, and alignments refined with Genewise as previously described.

We selected alignments from the newly generated transcriptome assembly and the two assemblies available in METdb belonging to the *P. calceolata* CCMP1214 strain to build a training set for AUGUSTUS *ab initio* gene predictor. Only gene models with complete coding DNA sequences were retained for training and 1,000 genes were taken apart for testing AUGUSTUS accuracy. A first training produced exon and intron parameters for *P. calceolata* species. Parameters were optimized using successive steps of training and testing. We calculated the accuracy of gene prediction by running AUGUSTUS on the test set. At the exon level, AUGUSTUS performed well in terms of sensitivity (0.619) and specificity (0.669). We thus run AUGUSTUS on the masked genome based on trained parameters. The *ab initio* prediction and all the transcriptomic and protein alignments were combined using Gmove which is an easy-to-use predictor with no need for a pre-calibration step[47]. Briefly, putative exons and introns, extracted from predictions and alignments, were used to build a graph, where nodes and edges represent exons and introns respectively. From this graph, Gmove extracts all paths and searches open reading frames (ORFs) which are consistent with the protein evidence. We trimmed untranslated transcribed regions that overlap coding part of a neighbour gene and renamed the genes following the standard nomenclature. Mono-exonic genes models encoding proteins of less than 200 amino acids without significant protein match (1,006 genes) were excluded. Chloroplastic and mitochondrial genes (contig 7 and 8) were predicted using previously published annotations for

P.calceolata[22,48]. Following this pipeline, we predicted 16,667 genes with 1.45 exons per gene on average.

The presence of introner elements (IE) in *P. calceolata* was investigated by overlapping the position of the introns in transcriptome alignments with the positions of repeat families detected *ab initio* with RepeatScout[39]. Intron overlapping repeats over more than 90% of their length were identified as putative IEs.

Functional analysis

Predicted gene models of *P. calceolata* nuclear genome (contig 1 to contig 6) were annotated for protein function using InterProScan v5.41-78.0[49]. A protein alignment against the NR database (01-12-2021 version) was performed with diamond v0.9.24[50]. The best protein match with a functional annotation and an e-value < 10^{-5} was retained. KEGG Orthologues (KO) were identified with the HMM search tool KofamKoala v1.3.0 and KO annotations with an e-value <10^{-5} and a score above the HMM threshold were retained[51]. Finally, Gene Ontology (GO) terms and Enzyme commission (EC) numbers were recovered from the interproscan and KO analysis respectively. Previously published chloroplastic and mitochondrial gene names and functions were reported on the corresponding genes[22,48]. All gene functional annotations of *P. calceolata* are available in Table S2.

In order to compare the functional annotation of *P. calceolata* with other small free-living photosynthetic eukaryote, we applied the same analysis on the predicted proteins available for the following species: *Aureococcus anophagefferens*, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Nannochloropsis oceanica*, *Bathycoccus prasinos*, *Micromonas pusilla*, *Ostreococcus lucimarinus* and *Emiliania huxleyi* (references are indicated in Table 1).

We defined a list of 23 meiosis-specific genes using three previous studies[52–54]. KO annotations and Interproscan domains were used to recover these genes in the *P. calceolata* genome.

Transmembrane regions in NIT-domain containing proteins were identified with TMHMM v 2.0[55]. *P. calceolata* ISIP proteins were identified by aligning diatom ISIPs on *P. calceolata* predicted proteins with blastp (p-value < 1e-5).

Mapping and filtering of environmental metagenomic reads.

We used metagenomics datasets of *Tara* Oceans and *Tara Polar Circle* expeditions to detect the *P. calceolata* genome in the oceans. Datasets from water samples collected on the photic zone: surface (SUR) and deep-chlorophyll maximum (DCM) were analysed. Size-fractionated water samples containing pico- and nano-eukaryotes (organisms <5µm in cell diameter) were selected: 0.2-3 µm (100 samples), 0.8-2000 µm (119 samples) and 0.8-5 µm (148 samples)[27]. Metagenomic reads were aligned on the *Pelagomonas calceolata* genome with BWA-mem version 0.7.15 with default parameters[58].

Alignments with at least 90% of identity over 80% of read length were retained for further analysis. In the case of several possible best matches, a random one was picked. In order to remove putative PCR duplicates, multiple read pairs aligned at the exact same position on *P. calceolata* genome were removed with samtools rmdup version 1.10.2[44].

Relative abundance calculation.

The relative abundance of *P. calceolata* was calculated from both metabarcoding data and metagenomic data. For the metabarcoding abundance, we used the 18SV9 rRNA OTU table published in 2021 and available here https://zenodo.org/record/3768510#.YEX2S9zjJaQ[59]. Bacterial and archaea OTUs were removed for this analysis. For the metagenomic abundance, we divided the total number of reads aligned on the *P. calceolata* genome by the total number of sequenced reads for each sample.

Analysis of environmental parameters

The environmental parameters measured during the expedition are available in the Pangaea database (https://www.pangaea.de/) and are described in[60]. Iron concentrations are annual means derived from PISCES2 model[61] and described in[12]. Ammonium concentrations at the date and location of sampling are derived from the MITgcm Darwin model and available at https://doi.pangaea.de/10.1594/PANGAEA.875577[62]. Environmental parameters for each sample are available in Table S3. Pearson's correlation between the relative abundance of *P. calceolata* and all environmental parameters were calculated with the *cor* function in R and the GGally package version 2.1.0. The 8 parameters with a correlation above 0.2 or below -0.2 were selected for downstream analysis and include the temperature (°C, *in situ*), the salinity (mg/m^3, *in situ*), the oxygen (µmol/kg, *in situ*), the ammonium (µmol/L, Darwin model), the nitrate (µmol/L, *in situ*), the iron (nmol/L, PISCES2 model), the chlorophyll-A concentration (mg/m^2, *in situ* calculated by vertical integration of fluorescence profiles 0-200 m) and distance to the coast (km). The principal component analysis (PCA) was performed with the R package FactoMineR version 2.4. We used a Generalized Additive Model (GAM) for its ability to fit non-linear and non-monotonic functions and for its low sensitivity to extreme values to model the relative abundance of *P. calceolata* as a function of iron concentration and temperature[63]. This function is implemented in the mgcv R package version 1.8-33. All figures were generated with R version 4.0.3 and ggplot2 package version 3.3.2.

Results

A compact genome revealed by a chromosome-scale assembly.

To investigate its gene repertoire and its distribution across the oceans we sequenced and assembled the genome of *P. calceolata* using ONT long-reads and Illumina short-reads. Using k-mers distribution

of short reads, the genome was estimated to be homozygous with a size of 31 Mb (Figure S1). The ONT long-reads were assembled using Flye into 6 nuclear contigs for a total of 32.4 Mb, 1 plastid circular contig (90 Kb) and 1 mitochondrial circular contig (39 Kb) (Figure 1 and Figure S2A,B). The gene completeness was estimated to 94.0% using BUSCO[35]. Two large and highly similar duplicated regions (>99% of identity) were detected at the extremity of contig 1 and 5 (393 Kb) and at the extremities of contig 3 and 6 (192 Kb). The vertical read coverage of these two regions is similar to other genomic regions suggesting that they are duplicated in all cells of this *P. calceolata* culture (Figure 1 and Figure S2C). In addition, 150 Kb at one extremity of contig 4 present a higher vertical coverage suggesting that this region is also duplicated in *P. calceolata* genome (Figure S2). Interestingly, duplicate regions at the end of sequences have already been observed in the chlorophyte *Ostreococcus tauri* which has been maintained for several years in culture[64]. These observations suggest that culture could affect not only the sequence of a given organism's genome but also its structure. (TTAGGG)n telomeric repeats were detected at both ends of contigs 2, 3, 4 and 6 indicating that these 4 contigs represent complete chromosomes (Figure S2). For contig 1 and contig 5, telomeric sequences are present at only one extremity, the other extremity ending in the duplicated region. This result suggests that the six contigs correspond to six chromosomes of *P. calceolata*.

A total of 16,667 genes was predicted on *P. calceolata* genome with an average of 1.45 intron per gene (Table 1). The distribution of intron lengths reveals a peak at around 210 bp (Figure S3A), which is the characteristic length of Introner Elements (IE) described in *A. anophagefferens*[65]. 956 putative IE were then identified (see method). Both canonical (GT) and non-canonical (GC) donor splicing sites were present at the ends of putative IE, while the acceptor site was canonical (AG) in all cases (Table S4). The logo representation of the putative IE sequences reveals the presence of GT and GC donor sites of the 5' ending, AG acceptor site at the 3' ending, and conserved TIR at the flanking regions (Figure S3B).

9, 583 *P. calceolata* predicted proteins (58%) are homologous with at least one Stramenopile gene and among them, 3,005 (18%) are shared only with the Pelagophyceae *A. anophagefferens* (Figure S4). A conserved functional domain (Pfam, KO or InterProScan) was found in 11,240 (67%) proteins. 4,822 proteins (33%) have no known functional domain.

Phylum/Class	Species	Genome size (Mb)	Number of chromosomes	Predicted genes	GC%	Cell size	References
Pelagophyceae	*Pelagomonas calceolata*	32.4	6	16,613	63.6	2 µm	This study
Pelagophyceae	*Aureococcus anophagefferens*	56.0	Unkown	11,520	67.4	2 µm	[66]
Eustigmatophyceae	*Nannochloropsis oceanica*	29.3	32	7,730	54.0	3 µm	[67]
Diatom	*Phaeodactylum tricornutum*	27.0	33	10,402	48.8	11 µm	[68]
Diatom	*Thalassiosira pseudonana*	34.5	24	11,776	46.9	5 µm	[69]

Chlorophyta	*Micromonas pusilla*	21.9	17	10,575	65	≤2 µm	[70]
Chlorophyta	*Ostreococcus lucimarinus*	13.2	21	7,651	60	1.3 µm	[71]
Chlorophyta	*Bathycoccus prasinos*	15.1	19	7,847	48	1-2 µm	[72]
Haptophyta	*Emiliania huxleyi*	141.7	Unkown	38,549	64.5	4–5 µm	[73]

Table 1: Nuclear genome characteristics of several unicellular photosynthetic algae.

GC content distribution along *P. calceolata* chromosomes

A remarkable feature in *P. calceolata* genome is the distribution of GC content along *P. calceolata* chromosomes (Figure 1). While the average GC content of the nuclear genome is 63%, one large region in each contig (259 Kb in average) is 52% GC. These unique large troughs in GC content in each chromosome suggest that these regions may encompass centromeres. We did not observe accumulation of repeated elements nor transposons in these low-GC regions; however, the gene structures differ from other genomic regions (Figure 1 and Table S5). The slight decrease of gene density observed in Figure 1 is explained a longer gene size (average of 3 Kb compared to 1.9 Kb in other regions). In addition, the 453 genes in low-GC regions contain more introns (2.56 introns per gene compared to 0.49) and introns are shorter (120 compared to 214 bases). We also noticed a higher proportion of intergenic regions, 18% in low-GC regions and 11% in other chromosomic regions (Table S5). Interestingly, three repeats of more than 500 bases are present in several low-GC regions (R_13 in contigs 2, 3 and 6; R_25 in all contigs except contig 1 and R_80 in contigs 1 to 4) (Figure 1). No homology was found between these sequences and known repeat elements.

Gene function analysis of low-GC regions reveals an enrichment of genes involved in specific cellular mechanisms (Table S6). Thirteen genes involved in DNA replication including the Anaphase-promoting complex subunit 4, the sister chromatid cohesion protein Dcc1 and 3 Mini-chromosome maintenance genes. Twenty-five genes are involved in microtubules synthesis and microtubules-binding motor proteins (9 genes carrying dynein domains, 6 genes carrying kinesin motor domains and 4 tubulin genes). These genes indicate that the low-GC regions contain many genes required for *P. calceolata* cellular division. Finally, 18 genes are involved in transcription including 3 genes encoding RNA Pol II rpb2 subunits and 7 genes encoding transcription factors suggesting an important role of these chromosomic regions for the regulation of gene expression.

Sex related genes in *P. calceolata*

Because low-GC regions could be related to meiotic recombination (see discussion), we looked for genes involved in sexual reproduction in *P. calceolata* genome. Among 20 genes specifically involved in meiosis, 16 homologs are present in *P. calceolata* genome (Table 2). These genes include the double-strand DNA breaks (DSB) initiator SPO11; RAD50, RAD52 and MRE11 to bind DSBs; HOP2, MND1, DMC1 and RAD51 to ensure pairing and invade the homologous strand; MSH2, MSH3 PMS1 and MSH6 genes

involved in the synthesis-dependent strand annealing pathway and MUS81 necessary for non-interfering (class II) crossing over. Interestingly, MSH4 and MSH5 genes are absent in *P. calceolata* genome. Indeed, these genes necessary to perform the interfering (class I) recombination pathway through Double Holliday Junctions are present in most eukaryotic lineages. The large low-GC regions could be a consequence of the absence of the MSH4/5 genes as suggested in yeasts[74] (see discussion). There are also no homologs of ZIP, HOP1 and RED1 genes in *P. calceolata* genome. These genes are known to be involved in homologous pairing of chromatids and construction of the synaptonemal complex in animals, plants and fungi but are absent in several phylum like diatoms[75] and ciliates[52]. Taken together, the genetic content of *P .calceolata* strongly suggests that this species is capable of meiosis.

Gene	KEGG or IPR domains	p-value	Gene name	Function
Pca_1p15030	K03348	0*	APC1	Anaphase-promoting complex subunit 1
Absent			HOP1	Synaptonemal complex protein; binds DSBs and oligomerizes during meiotic prophase I
Pca_1p11130	IPR010776	2.00E-09	HOP2	HOP2 and MND1 form a heterodimeric complex that interacts with RAD51 and DMC1 and promotes interhomolog meiotic recombination and reduces synapsis and recombination of nonhomologous chromosome
Pca_1p31050		7.90E-06		
Pca_6p12570	IPR040453	5.00E-20	MND1	
Pca_3p28280	K15271	5.50E-193	MER3	ATP-dependent DNA helicase
Pca_4p24660		3.70E-174		
Pca_5p17290	K08734	4.9E-210*	MLH1	Mismatch repair and promotion of meiotic crossing over ; Forms heterodimers with Mlh2, Mlh3, and Pms1
Pca_4p08690	K08739	7.50E-100	MLH3	Forms a heterodimer with Mlh1; Mismatch repair and promotion of meiotic crossing over
Pca_3p17860	K10865	1.9E-198*	MRE11	3'-5' dsDNA exonuclease and ssDNA endonuclease; forms complex with Rad50 and Xrs2/Nbs1
Pca_1p04040	K10866	9.4E-277*	RAD50	ATPase, DNA binding protein; holds broken DNA ends together while Mre11 trims
Pca_1p20950	K08735	3.5E-190*	MSH2	Dna mismatch repair protein ; Forms a heterodimer with Msh3 or Msh6
Pca_1p10430		4.00E-152		
Pca_2p23180	K08736	1.50E-150	MSH3	Dna mismatch repair proteinn ; Forms a heterodimer with Msh2
Absent			MSH4	The meiosis-specific MutS homologs, MSH4 and MSH5, function as a heterodimer and have specialized roles in meiotic recombination and Holliday junction resolution
Absent			MSH5	
Pca_1p10290	K08737	0*	MSH6	Forms a heterodimer with Msh2; binds base mismatches
Pca_4p14280		4.70E-50		
Pca_4p08690	K10864	3.70E-94	PMS1	Forms heterodimer with Mlh1 for repair of heteroduplex DNA; interacts with Msh2/Msh3
Pca_6p04320	K10872	8.80E-148	DMC1	Promotes interhomolog recombination; Homolog of strand exchange protein Rad51;
Pca_6p04320	K04482	2E-163*	RAD51	With Dmc1, catalyzes homologous DNA pairing and strand exchange
Pca_3p24140	K10873	3.30E-13	RAD52	Binds DSBs and initiates assembly of meiotic recombination complexes
Pca_3p27080	K12780	1.00E-29	REC8	Meiotic recombination protein REC8
Absent			RED1	Reductional division protein 1
Pca_2p25160	K10878	2.7E-142*	SPO11	Transesterase; creates DNA double-strand breaks (DSBs) in meiosis I
Absent			ZIP1	Synaptonemal complex protein ZIP1
Pca_1p11000	K08991	7E-40*	MUS81	crossover junction endonuclease

Table 2: Putative meiotic genes identified in *P. calceolata* genome. Kegg Orthology p-values with a star are above the HMM threshold defined by KoFamKoala tool.

Genes involved in nitrogen uptake, storage and recycling.

P. calceolata could be an important player in nitrogen (N) cycle in oceanic ecosystems[24], therefore we explored the genomic capacities of *P. calceolata* to assimilate and use nitrogen-containing compounds. We systematically identified and counted genes involved in the nitrogen cycle in *P. calceolata* compared to seven other photosynthetic pico- and nano- algae (Figure 2 and Table S7).

The uptake of nitrogen-containing inorganic compounds is supported by 13 genes of *P. calceolata*. Among them, 8 genes encode nitrate/nitrite or formate/nitrite transporters, which is on average higher than in other algae. In contrast, only 5 genes encode ammonium transporters that is low compared to other species suggesting that nitrite and/or nitrate is the main external source of inorganic nitrogen for *P. calceolata*. The number of enzymes incorporating ammonium into organic compounds (GS/GOGAT pathway) are higher in *P. calceolata* than in other species: 5 glutamine synthetase and 4 glutamate synthase are present in the *P. calceolata* genome. On the contrary, a lower number of genes supports the reduction of nitrate into nitrite then ammonium (2 nitrite reductases and 1 nitrate reductase) (Table S7).

We identified 3 genes carrying the nitrate and nitrite sensing (NIT) domain (IPR013587) in *P. calceolata* genome. Among the 7 other genomes studied, this protein domain is only found in *A. anophagefferens* (1 gene) and *E. huxleyi* (3 genes). One *P. calceolata* protein carry a NIT domain surrounded by 2 transmembrane domains suggesting a capacity of external Nitrate/Nitrite sensing while the 2 other NIT genes carry a protein-kinase domain (IPR000719) suggesting a phosphorylation-based signal transduction depending of intracellular nitrate or nitrite concentration (Figure S5).

We then identified genes involved in nitrogen recycling from organic compounds which are important in several species in case of inorganic nitrogen deprivation. One arginase gene and one cyanase gene were detected in *P. calceolata* genome but no gene encoding formamidase. In addition, the number of gene copies for enzymes involved in the urea cycle (carbamoyl-phosphate synthetase, ornithine carbamoyltransferase, argininosuccinate synthase and argininosuccinate lyase) is equal or slightly lower than in other algae (Figure 2 and Table S7). These results suggest that *P. calceolata* is not particularly adapted to recycle nitrogen from organic molecules but could be capable of incorporating inorganic nitrogen compounds even in N-poor environments.

Genes related to iron uptake, storage and usage in *P. calceolata*

Iron is a critical metal for all photosynthetic organisms. Iron-containing molecules are essential for the photosynthesis, the nitrogen cycle and the protection against reactive oxygen species. We identified gene coding for iron uptake, storage in *P. calceolata* genome and compared them to other small photosynthetic eukaryotes (Table S7). *P. calceolata* has 5 genes encoding the phytotransferrin ISIP2 involved in Fe^{3+} uptake via endosomal vesicles and 2 putative iron storage protein ISIP3. These genes

are transcribed following starvation in diatoms suggesting a capability for thriving in iron-poor environments for *P. calceolata*. In addition, 3 genes encode the iron transporter ferroportin. These proteins are iron exporter in multicellular organisms but its function in micro-algae remains to be studied[76]. Zinc/iron permeases, transmembrane ferric reductases and multicopper oxidases genes, involved in iron uptake from the environment, are present in *P. calceolata* genome (8, 5 and 2 genes respectively) but in equal or lower gene copy number than in other studied species. The iron permease FRT1 and the ISIP1 gene, involved in endocytosis of iron-chelator molecules (siderophores), are absent in the *P. calceolata* genome. Furthermore, *P. calceolata* has no gene encoding for iron storing protein ferritin. The absence of these genes suggests that iron uptake and storage is not a major asset of *P. calceolata* compared to the other photosynthetic protists.

Several important ferrous proteins can be substituted by non-ferrous equivalents in iron-poor environments[13]. In *P. calceolata* genome we identified 11 flavodoxin and 3 phytocyanin genes, encoding non-ferrous proteins involved in electron transfer during photosynthesis potentially replacing Ferredoxin and Cytochrome C_6 respectively. The Fructose-bisphosphate Aldolase (FBA) necessary for the gluconeogenesis and the Calvin cycle is encoded by 6 genes in *P. calceolata*. Two genes are dependent of a bivalent cation (FBA type II), the four other are Zinc/Iron-independent (FBA type I). Finally, all types of Superoxide dismutases (SOD) are in the *P. calceolata* genome. Non-ferrous SOD (Cu/Zn and Ni) are encoded by 3 genes, while the Mn/Fe-SOD are encoded by 2 genes. The genetic content of *P. calceolata* indicates that several essential processes are potentially iron-independent thanks to the presence of many iron-free alternative proteins.

Relative abundance of *P. calceolata* across oceanic basins.

Firstly, we used the OTU table computed from metabarcoding samples of the *Tara* Oceans expedition to estimate the relative abundance of *P. calceolata* across all oceans[59]. The relative rRNA abundance of *P. calceolata* OTU in the 0.8-5 µm size-fraction is 1.30% in average of the 111 surface samples and 0.81% in average of the 62 DCM samples (Table S8). According to this method of abundance estimation, *P. calceolata* is the third most abundant eukaryote of the 0.8-5 µm size-fraction among *Tara* samples covering all oceans except the Arctic Ocean. However, the number of rRNA copies in each organism biases this metabarcoding-based abundance estimation. Therefore, we used the mapping of metagenomic reads on the *P. calceolata* genome to estimate more precisely its abundance. The relative abundance obtained with the entire genome is strongly correlated to the metabarcoding-based relative abundance (Pearson correlations of 0.91 and 0.70 for 0.8-2000 and 0.8-5 µm size-fractions respectively). However, the metabarcoding-based abundance is underestimated by a factor of 2.3 in the 0.8-5 µm size-fraction and a factor of 3.1 in the 0.8-2000 size-fraction compared to the metagenomic-based abundance (Figure S6). This underestimation is probably due to the low copy

number of rRNA in *P. calceolata* (2 complete copies) compared to most other species with larger genomes.

The relative abundance of *P. calceolata* calculated with the number of read aligned on the genome is estimated above 1% in 66 oceanic stations with a maximum of 6.7% for the 0.8-5 µm size-fraction at the DCM in the North Indian Ocean (Figure 3A and Figure S7). In the Indian Ocean, Red Sea and Mediterranean Sea, *P. calceolata* is significantly more abundant in the DCM than in the surface (Figure 3B). In cold waters (below 10°C), *P. calceolata* is not detected above our threshold of 25% of genomic horizontal coverage. Important variations between and within each oceanic basin are observed, suggesting that many biotic or abiotic factors influence *P. calceolata* abundance.

Relative high abundance of *P. calceolata* in temperate and iron-poor regions

In order to identify factors controlling *P. calceolata* abundance in the oceans, we used physical-chemical parameters available for each oceanic station (see methods). Principal component analysis reveals a positive relation between *P. calceolata* abundance, the temperature and the coast distance and a negative relation with iron concentration (Figure 4). The other parameters do not seem related to *P. calceolata* abundance. This result is consistent in the 3 size-fractions containing *P. calceolata* cells (0.8-5 µm, 0.8-2000 µm and 0.2-3 µm size-fractions) (Figure S8).

Despite the numerous factors potentially influencing *P. calceolata* abundance, we observed a weak but significant Pearson's positive correlation with the temperature and a negative correlation with iron concentration (Table 3). In addition, we used a general additive model to estimate the contribution of the combination of temperature and iron concentration to *P. calceolata* relative abundance (Table 3). The two factors contribute significantly and explain 17.9% of the variations of *P. calceolata* abundance in the 0.8-5 µm size-fraction and 47.9% in the 0.8-2000 µm size-fraction. The high relative abundance of *P. calceolata* in iron-poor waters suggests that this species is particularly able to acclimate to this environmental condition.

A		**GAM model**		**GAM verification**		**Pearson correlations**	
0.8 - 5 µm	edf	F value	p-value	k-index	k p-value	r	p-value
s(Temperature)	2.754	4.229	0.00506	1.00	0.46	0.23	0.001
s(Iron concentration)	1.897	5.549	0.00326	0.92	0.16	-0.25	0.001
Adjusted R^2	0.151						
Deviance explained	17.9%						
B		**GAM model**		**GAM verification**		**Pearson correlations**	
0.8 - 2000 µm	edf	F value	p-value	k-index	k p-value	r	p-value
s(Temperature)	4.223	8.769	7.89e-07	0.98	0.405	0.57	0.0001
s(Iron concentration)	1.450	0.793	0.348	0.86	0.045	-0.47	0.0001
Adjusted R^2	0.449						

Deviance explained	47.9%

Table 3: Environmental parameters explaining *P. calceolata* relative abundance for the (A) 0.8-5 µm and the (B) 0.8-2000 µm size-fractions.

Discussion

The essential roles of phytoplankton in oceanic ecosystems have been illustrated many times, however numerous lineages are still poorly explored and model organisms are restricted to a few taxa (mainly diatoms, prasinophytes, haptophytes…) limiting the global understanding of phytoplankton activity. *P. calceolata* genome assembled and annotated in this study brings new insights into specific genomic features of this algae class related to its adaptation to specific environments and reveal a previously underestimated high abundance of *P. calceolata* in the oceans.

Large low-GC centromeres

The chromosome level assembly of *P. calceolata* genome thanks to nanopore read sequencing has reveal unique genomic features that are not usually studied in short-read assembled genomes. The most striking observation is the presence of a unique GC trough (50%) in each scaffold contrasting with the high GC content (63%) of other genomic regions of *P. calceolata*. In eukaryotes, centromeres have a large variety of structures and characteristics. They are composed of many repeated sequences or contain genes, they are determined genetically or epigenetically and their size can vary from 125 bases (in *S. cerevisiae*) to several Mb (in mammals)[77]. Short regional centromeres (1-5 Kb) are generally low-GC compared to the genome and the sequence is unique at each centromere. Among stramenopiles, centromeres were characterized in the diatom *P. tricornatum* where a short low-GC sequence (>500 bp) is enough to define a centromere[78]. Large regional centromeres (>10 Kb) are generally gene-free, contain repeated elements and are not transcribed[77]. *P. calceolata* putative centromeres derive from this general pattern with large low-GC regions containing genes carrying essential functions of DNA replication, transcription and microtubule assembly. Interestingly, the red alga *Cyanidioschyzon* as well as several yeast species seem to have similar centromeres structures with large low-GC centromeres containing genes[79,74].

GC content, meiosis and recombination

The main hypothesis to explain these low-GC patterns in centromeres is the importance of GC-biased gene conversions (gBGC) during recombination and the inhibition of these recombination near centromeres[80]. Indeed, gBGCs increase the GC content of recombining DNA over evolutionary time inducing GC content variations within and between genomes[81]. The kinetochore formation at centromeric regions inhibits recombination and double strand breaks formation during meiosis resulting in rare gBGC in these regions[82]. Centromeric and peri-centromeric regions may therefore

have a lower GC content. Furthermore, the genetic content of *P. calceolata* indicate that this species is capable of meiosis despite the absence of some genes including MSH4 and MSH5 necessary to perform meiotic recombination through double Holliday junctions. In yeasts (*Yarrowia lipolytica*, *Candida lusitaniae*, and *Pichia stipites*), a correlation has been observed between the importance of the GC trough near the centromeres (>10%) and the absence of MSH4/MSH5 genes[74]. It is therefore possible that the absence of this recombination pathway in *P. calceolata* induces more frequently double-strand break repairs by synthesis dependent strand annealing and a more rapid GC-biased conversion on the entire genome except at the centromeres where double-strand breaks are inhibited. This recombination inhibition may have important consequences on the evolution of *P. calceolata* genome. Genes within low-GC regions are significantly longer and contains more introns than genes in other genomic regions. Because intron gain and loss are closely related to double-strand breaks repair and homologous recombination, we then suggest that centromere genes retain more introns because double-strand breaks are reduced[83]. Variant analysis in *P. calceolata* populations could be targeted specifically in future studies to infer an estimation of recombination rate and more generally characterize the evolution processes controlling these large centromere regions.

P. calceolata is one of the most abundant eukaryotes in the oceans

We have shown in this study that *P. calceolata* is cosmopolite in oceanic samples above 10°C with a relative abundance generally >1% of all sequenced reads. In contrast to the coastal Pelagophyceae *A. anophagefferens* that can present high peaks of abundance[84] no *P. calceolata* blooms are reported but *P. calceolata* is well-adapted to a large range of environmental conditions as suggested by previous studies[20,22].

Although the abundance of an organism calculated from metabarcoding or metagenomic data provide only an indirect and relative quantification of organism abundances, both methods suggest that *P. calceolata* is one of the most abundant pico-nano eukaryote in offshore data. The high relative abundance of *P.calceolata* measured with a metabarcoding approach has recently been confirmed with a qPCR method[20]. In addition, we have shown that the metabarcoding approach probably underestimates the relative abundance of *P.calceolata* compared to the metagenomic analysis due to the few copy number of rRNAs in this organism. Further studies may combine microscopic and flow sorting approaches with genomic data to assess the number of cells and biomass of this organism in the oceans.

Iron uptake and storage and modulation of iron needs

Iron is an essential element for the growth, photosynthesis, primary production, nitrogen fixation and reduction for PPEs[85]. Our results show that *P. calceolata* thrives in iron-poor waters and thus occupies

a large ecological niche for a PPE. Two main strategies exist against iron deprivation: 1) optimisation of iron uptake and 2) modulation of iron needs. Optimisation of iron uptake does not seem to be the main strategy of *P. calceolata* in iron-poor environments. Genes coding for iron chelators and ferritin are absent in its genome, and gene coding for passive iron transporters are under- or equally represented compared to other PPEs. In contrast, phytotransferrins (ISIP2), putative iron storage proteins (ISIP3) and ferroportins are overrepresented in *P. calceolata* genome. In Pelagophyceae, expression levels of ISIP genes are correlated with iron concentration at the global scale suggesting a transcriptomic regulation in response to iron availability[12]. Because phytotransferrins are dependent on carbonate ions, ocean acidification may reduce the efficiency of iron uptake in many species[86,87]. In consequence, organisms able to grow with very low iron concentration like *P. calceolata* may be favoured.

The presence of three ferroportin genes in *P. calceolata* is interesting since these transmembrane iron-exporters proteins play a major role in iron homeostasis in multicellular organisms[76]. Ferroportin function in micro-algae is unknown but could act to export iron from endocytosis vesicles to the cytoplasm[88]. In the green alga *Chlamydomonas reinhardtii*, a ferroportin gene is overexpressed under low Fe conditions[89]. The function of the 3 ferroportin genes of *P. calceolata* could be investigated to understand their role in iron-poor environments. In addition, modulation of iron needs seems to be a major strategy for *P. calceolata*. All known molecular switches between ferrous and non-ferrous proteins are genetically possible in *P. calceolata* and the non-ferrous encoded proteins genes FBA I, flaovodoxin and phytocyanin present a higher number of gene copies than the ferrous ones. It could be a sign that those switches are easier for *P. calceolata* than for other PPEs, but transcriptomic experiments are required to determine if this genetic overrepresentation has metabolic consequences.

Nitrogen cycle

Expressing more than 90% of all nitrate transporters transcripts, Pelagophytes may dominate nitrate uptake and assimilation in the North Pacific Ocean[24]. Indeed, *P. calceolata* contains a large collection of genes of nitrate, nitrite and urea transporter. Interestingly, a few genes are ammonium transporters suggesting that the main source of nitrogen for *P. calceolata* is the nitrite and/or nitrate. One remarkable feature of the *P. calceolata* genome is the presence of 3 genes carrying NIT domains (PF08376). This NIT domain was first described in bacteria as nitrite and nitrate sensor proteins [90]. This sensor is an alpha-helical protein playing a signal transduction role regulating gene expression, cell motility and enzyme activity in *Klebsiella oxytoca*[91]. Most of NIT-domain genes in microalgae are associated with a serine-threonine/tyrosine-kinase domain suggesting a signal transduction according to the presence of intracellular nitrate or nitrite. In Pelagophytes and in the coral symbiont dinoflagellate *Symbiodinium*, genes carrying a NIT domain are surrounded by 2 transmembrane

domains suggesting extracellular sensing of nitrate/nitrite. *P. calceolata* is the only known species among eukaryotic algae carrying both types of NIT proteins. The NIT-domain genes in *P. calceolata* are highly expressed in subtropical Pacific N-depleted waters[23], suggesting that these proteins have a transcriptional regulation role according to nitrate availability[24].

Conclusion

Due to its widespread distribution and its high abundance in the open oceans, *Pelagomonas calceolata* can serve as an ecologically-relevant model to study marine photosynthetic protists. The chromosome-scale genome sequence, mostly telomere-to-telomere generated in this study is an essential starting point for its detection in environmental datasets. This will allow to study its ecological importance in the oceans at a molecular level. We have shown that *P. calceolata* genome has specific genomic features potentially explaining its ecological success. The large repertoire of genes involved in nutrient acquisition from the environment is coherent with its widespread pattern of relative abundance distribution across different environments. The ecological niche of *P. calceolata* suggests that this alga will benefit from the global climate change with the extension of oligotrophic regions and global ocean warming. Future studies could use this *P. calceolata* genome to explore adaptation and acclimation processes controlling the distribution and abundance of this alga.

Data availability

Pelagomonas calceolata genomic and transcriptomic reads, the genome assembly and gene prediction are available at the ENA (EMBL-EBI) website under the accession number PRJEB47931. Tara Oceans and Tara Polar Circle metagenomic sequences are archived at the ENA under the following accession numbers: PRJEB9740, PRJEB9691, PRJEB4352 and PRJEB1787.

Acknowledgments

We thank the commitment of the following people who made this work possible: the CNRS (in particular the Federation de Recherche R2022/Tara Oceans GO-SEE), the Genoscope/CEA, Marie-José Garet-Delmas from the Roscoff Culture Collection for growing the RCC100 strain, Claude Scarpelli for support in high-performance computing. Computations were performed using the cobalt HPC machine provided through FRANCE GENOMIQUE (ANR-10-INBS-09–08). We also thank the *Tara* Expedition Foundation and their partners for the organization of marine scientific expeditions (http://oceans.taraexpeditions.org).

References

1. Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science *281*, 237–240.

2. Boyce, D.G., Lewis, M.R., and Worm, B. (2010). Global phytoplankton decline over the past century. Nature *466*, 591–596.

3. Henson, S.A., Cael, B.B., Allen, S.R., and Dutkiewicz, S. (2021). Future phytoplankton diversity in a changing climate. Nat Commun *12*, 5372.

4. Vaulot, D., Eikrem, W., Viprey, M., and Moreau, H. (2008). The diversity of small eukaryotic phytoplankton (≤3 µm) in marine ecosystems. FEMS Microbiology Reviews *32*, 795–820.

5. Morán, X.A.G., López-Urrutia, Á., Calvo-Díaz, A., and Li, W.K.W. (2010). Increasing importance of small phytoplankton in a warmer ocean. Global Change Biology *16*, 1137–1144.

6. Li, W.K.W., McLaughlin, F.A., Lovejoy, C., and Carmack, E.C. (2009). Smallest algae thrive as the arctic ocean freshens. Science *326*.

7. Benner, I., Irwin, A.J., and Finkel, Z. V. (2020). Capacity of the common Arctic picoeukaryote Micromonas to adapt to a warming ocean . Limnology and Oceanography Letters *5*.

8. Sunda, W.G., and Huntsman, S.A. (1995). Iron uptake and growth limitation in oceanic and coastal phytoplankton. Marine Chemistry *50*, 189–206.

9. Raven, J.A. (1998). The twelfth Tansley Lecture. Small is beautiful: the picophytoplankton. Functional Ecology *12*, 503–513.

10. Morel, F.M.M., and Price, N.M. (2003). The Biogeochemical Cycles of Trace Metals in the Oceans. Science *300*, 944–947.

11. Gao, X., Bowler, C., and Kazamia, E. (2021). Iron metabolism strategies in diatoms. Journal of Experimental Botany *72*, 2165–2180.

12. Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Karlusich, J.J.P., Vieira, F.R.J., Villar, E., Chaffron, S., Malviya, S., et al. (2019). Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. Global Biogeochemical Cycles *33*, 391–419.

13. Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of eukaryotic genes. Nat Commun *9*, 373.

14. Morrissey, J., Sutak, R., Paz-Yepes, J., Tanaka, A., Moustafa, A., Veluchamy, A., Thomas, Y., Botebol, H., Bouget, F.-Y., McQuaid, J.B., et al. (2015). A Novel Protein, Ubiquitous in Marine Phytoplankton, Concentrates Iron at the Cell Surface and Facilitates Uptake. Current Biology *25*, 364–371.

15. Moore, C.M., Mills, M.M., Arrigo, K.R., Berman-Frank, I., Bopp, L., Boyd, P.W., Galbraith, E.D., Geider, R.J., Guieu, C., Jaccard, S.L., et al. (2013). Processes and patterns of oceanic nutrient limitation. Nature Geosci *6*, 701–710.

16. Kumar, A., and Bera, S. (2020). Revisiting nitrogen utilization in algae: A review on the process of regulation and assimilation. Bioresource Technology Reports *12*, 100584.

17. Smith, S.R., Dupont, C.L., McCarthy, J.K., Broddrick, J.T., Oborník, M., Horák, A., Füssy, Z., Cihlář, J., Kleessen, S., Zheng, H., et al. (2019). Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. Nat Commun *10*, 4552.

18. Berg, G.M., Glibert, P.M., Lomas, M.W., and Burford, M.A. (1997). Organic nitrogen uptake and growth by the chrysophyte Aureococcus anophagefferens during a brown tide event. Marine Biology *129*, 377–387.

19. Andersen, R.A., Saunders, G.W., Paskind, M.P., and Sexton, J.P. (1993). Ultrastructure and 18s rrna gene sequence for Pelagomonas calceolata gen. et sp. nov. and the description of a new algal class, the pelagophyceae classis nov. Journal of Phycology *29*, 701–715.

20. Choi, C.J., Jimenez, V., Needham, D.M., Poirier, C., Bachy, C., Alexander, H., Wilken, S., Chavez, F.P., Sudek, S., Giovannoni, S.J., et al. (2020). Seasonal and Geographical Transitions in Eukaryotic Phytoplankton Community Structure in the Atlantic and Pacific Oceans. Front Microbiol *11*, 542372.

21. Duerschlag, J., Mohr, W., Ferdelman, T.G., LaRoche, J., Desai, D., Croot, P.L., Voß, D., Zielinski, O., Lavik, G., Littmann, S., et al. (2021). Niche partitioning by photosynthetic plankton as a driver of CO2-fixation across the oligotrophic South Pacific Subtropical Ocean. ISME J, 1–12.

22. Worden, A.Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R.M., Malfatti, S., Tringe, S.G., and Keeling, P.J. (2012). Global distribution of a wild alga revealed by targeted metagenomics. Current Biology *22*, R675–R677.

23. Dimier, Cé., Brunet, C., Geider, R., and Raven, J. (2009). Growth and photoregulation dynamics of the picoeukaryote Pelagomonas calceolata in fluctuating light. Limnology and Oceanography *54*, 823–836.

24. Dupont, C.L., McCrow, J.P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., Roth, R., Hogle, S.L., Bai, J., Johnson, Z.I., et al. (2015). Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. ISME J *9*, 1076–1092.

25. Kang, Y., Harke, M.J., Berry, D.L., Collier, J.L., Wilhelm, S.W., Dyhrman, S.T., and Gobler, C.J. (2021). Transcriptomic Responses of Four Pelagophytes to Nutrient (N, P) and Light Stress. Front. Mar. Sci. *8*.

26. Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713–714.

27. Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.-M., Belser, C., Bertrand, A., et al. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data *4*, 170093.

28. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217.

29. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202–2204.

30. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol *37*, 540–546.

31. Liu, H., Wu, S., Li, A., Ruan, J., Wu, S., Li, A., and Ruan, J. (2021). SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte *2021*, 1–9.

32. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat Methods *17*, 155–158.

33. Aury, J.-M., and Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genomics and Bioinformatics *3*.

34. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res *27*, 737–746.

35. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol *35*, 543–548.

36. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*, 573–580.

37. Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol *13*, 1028–1040.

38. Smit, AFA, Hubley, R & Green, P. RepeatMasker. http://repeatmasker.org/.

39. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

40. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

41. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

42. Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086–1092.

43. Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821–829.

44. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

45. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Research *43*, D222–D226.

46. Niang, G., Hoebeke, M., Meng, A., Liu, X., Scheremetjew, M., Finn, R., Pelletier, E., and Corre, E. (2020). METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES . F1000Research *9*.

47. Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C.D., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M. (2016). Gmove a tool for eukaryotic gene predictions using various evidences . F1000Research *5*.

48. Sibbald, S.J., Lawton, M., and Archibald, J.M. (2021). Mitochondrial Genome Evolution in Pelagophyte Algae. Genome Biology and Evolution *13*.

49. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res *33*, W116-120.

50. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods *18*, 366–368.

51. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251–2252.

52. Chi, J., Mahé, F., Loidl, J., Logsdon, J., and Dunthorn, M. (2014). Meiosis gene inventory of four ciliates reveals the prevalence of a synaptonemal complex-independent crossover pathway. Mol Biol Evol *31*, 660–672.

53. Ramesh, M.A., Malik, S.-B., and Logsdon, J.M. (2005). A phylogenomic inventory of meiotic genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. Curr Biol *15*, 185–191.

54. Schurko, A.M., and Logsdon, J.M. (2008). Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. Bioessays *30*, 579–589.

55. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567–580.

56. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780.

57. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol *35*, 1547–1549.

58. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

59. Ibarbalz, F.M., Henry, N., Brandão, M.C., Martini, S., Busseni, G., Byrne, H., Coelho, L.P., Endo, H., Gasol, J.M., Gregory, A.C., et al. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. Cell *179*, 1084-1097.e21.

60. Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans data. Sci Data *2*, 150023.

61. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M. (2015). PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geoscientific Model Development *8*, 2465–2513.

62. Clayton, S., Dutkiewicz, S., Jahn, O., Hill, C., Heimbach, P., and Follows, M.J. (2017). Biogeochemical versus ecological consequences of modeled ocean physics. Biogeosciences *14*, 2877–2889.

63. Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A.N. (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. Environment International *132*, 104987.

64. Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F.-Y., Carré, I., Château, A., Eyre-Walker, A., Grimsley, N., Moreau, H., et al. (2014). An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. BMC Genomics *15*, 1103.

65. Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Nature *538*, 533–536.

66. Gobler, C.J., Berry, D.L., Dyhrman, S.T., Wilhelm, S.W., Salamov, A., Lobanov, A.V., Zhang, Y., Collier, J.L., Wurch, L.L., Kustka, A.B., et al. (2011). Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. PNAS *108*, 4352–4357.

67. Guo, L., Liang, S., Zhang, Z., Liu, H., Wang, S., Pan, K., Xu, J., Ren, X., Pei, S., and Yang, G. (2019). Genome assembly of Nannochloropsis oceanica provides evidence of host nucleus overthrow by the symbiont nucleus during speciation. Commun Biol *2*, 1–12.

68. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature *456*, 239–244.

69. Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al. (2004). The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. Science *306*, 79–86.

70. Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., et al. (2009). Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science *324*, 268–272.

71. Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., et al. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. PNAS *104*, 7705–7710.

72. Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M.F., et al. (2012). Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. Genome Biology *13*, R74.

73. Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature *499*, 209–213.

74. Lynch, D.B., Logue, M.E., Butler, G., and Wolfe, K.H. (2010). Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol Evol *2*, 572–583.

75. Patil, S., Moeys, S., von Dassow, P., Huysman, M.J.J., Mapleson, D., De Veylder, L., Sanges, R., Vyverman, W., Montresor, M., and Ferrante, M.I. (2015). Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. BMC Genomics *16*, 930.

76. Ward, D.M., and Kaplan, J. (2012). Ferroportin-mediated iron transport: expression and regulation. Biochim Biophys Acta *1823*, 1426–1433.

77. Talbert, P.B., and Henikoff, S. (2020). What makes a centromere? Experimental Cell Research *389*, 111895.

78. Diner, R.E., Noddings, C.M., Lian, N.C., Kang, A.K., McQuaid, J.B., Jablanovic, J., Espinoza, J.L., Nguyen, N.A., Anzelmatti, M.A., Jansson, J., et al. (2017). Diatom centromeres suggest a mechanism for nuclear DNA acquisition. PNAS *114*, E6015–E6024.

79. Kanesaki, Y., Imamura, S., Matsuzaki, M., and Tanaka, K. (2015). Identification of centromere regions in chromosomes of a unicellular red alga, Cyanidioschyzon merolae. FEBS Letters *589*, 1219–1224.

80. Nambiar, M., and Smith, G.R. (2016). Repression of harmful meiotic recombination in centromeric regions. Semin Cell Dev Biol *54*, 188–197.

81. Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G.A.B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. Genome Biol Evol *4*, 675–682.

82. Vincenten, N., Kuhl, L.-M., Lam, I., Oke, A., Kerr, A.R., Hochwagen, A., Fung, J., Keeney, S., Vader, G., and Marston, A.L. (2015). The kinetochore prevents centromere-proximal crossover recombination during meiosis. eLife *4*, e10850.

83. Farlow, A., Meduri, E., and Schlötterer, C. (2011). DNA double-strand break repair and the evolution of intron density. Trends Genet *27*, 1–6.

84. Gobler, C.J., Lonsdale, D.J., and Boyer, G.L. (2005). A review of the causes, effects, and potential management of harmful brown tide blooms caused byAureococcus anophagefferens (Hargraves et sieburth). Estuaries *28*, 726–749.

85. Martin, J.H., Coale, K.H., Johnson, K.S., Fitzwater, S.E., Gordon, R.M., Tanner, S.J., Hunter, C.N., Elrod, V.A., Nowicki, J.L., Coley, T.L., et al. (1994). Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. Nature *371*, 123–129.

86. Shi, D., Xu, Y., Hopkinson, B.M., and Morel, F.M.M. (2010). Effect of Ocean Acidification on Iron Availability to Marine Phytoplankton. Science *327*, 676–679.

87. McQuaid, J.B., Kustka, A.B., Oborník, M., Horák, A., McCrow, J.P., Karas, B.J., Zheng, H., Kindeberg, T., Andersson, A.J., Barbeau, K.A., et al. (2018). Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. Nature *555*, 534–537.

88. Turnšek, J., Brunson, J.K., Viedma, M. del P.M., Deerinck, T.J., Horák, A., Oborník, M., Bielinski, V.A., and Allen, A.E. (2021). Proximity proteomics in a marine diatom reveals a putative cell surface-to-chloroplast iron trafficking pathway. eLife *10*, e52770.

89. Urzica, E.I., Casero, D., Yamasaki, H., Hsieh, S.I., Adler, L.N., Karpowicz, S.J., Blaby-Haas, C.E., Clarke, S.G., Loo, J.A., Pellegrini, M., et al. (2012). Systems and Trans-System Level Analysis

Identifies Conserved Iron Deficiency Responses in the Plant Lineage[W][OA]. Plant Cell *24*, 3921–3948.

90. Shu, C.J., Ulrich, L.E., and Zhulin, I.B. (2003). The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors. Trends Biochem Sci *28*, 121–124.

91. Wu, S.Q., Chai, W., Lin, J.T., and Stewart, V. (1999). General Nitrogen Regulation of Nitrate Assimilation Regulatory Gene nasR Expression in Klebsiella oxytoca M5al. Journal of Bacteriology *181*, 7274–7284.

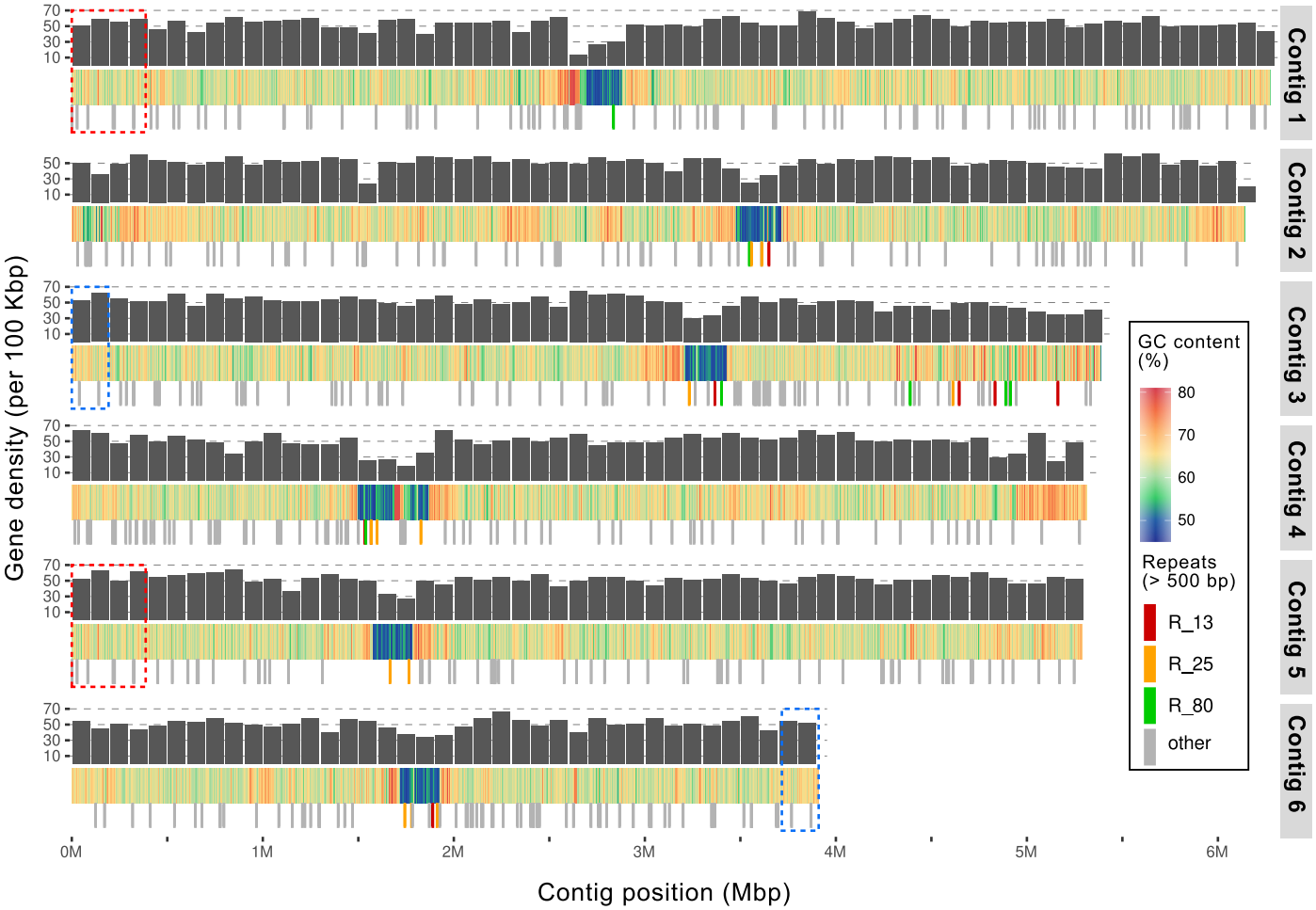

Figure 1: *Pelagomonas calceolata* nuclear chromosomes. Representation of the 6 nuclear contigs of *P. calceolata*. The top layer indicates the number of genes per 100 Kb (black bars), the middle layer represents the GC content in percentage over a window of 200 Kb and the bottom layer is the position of DNA repeats of more than 500 bases repeated at least 5 times over the entire genome. Red, orange and yellow bars indicate three different repeats in low-GC regions present in at least 3 different contigs. Dashed red and blue rectangles are duplicated chromosomic regions.

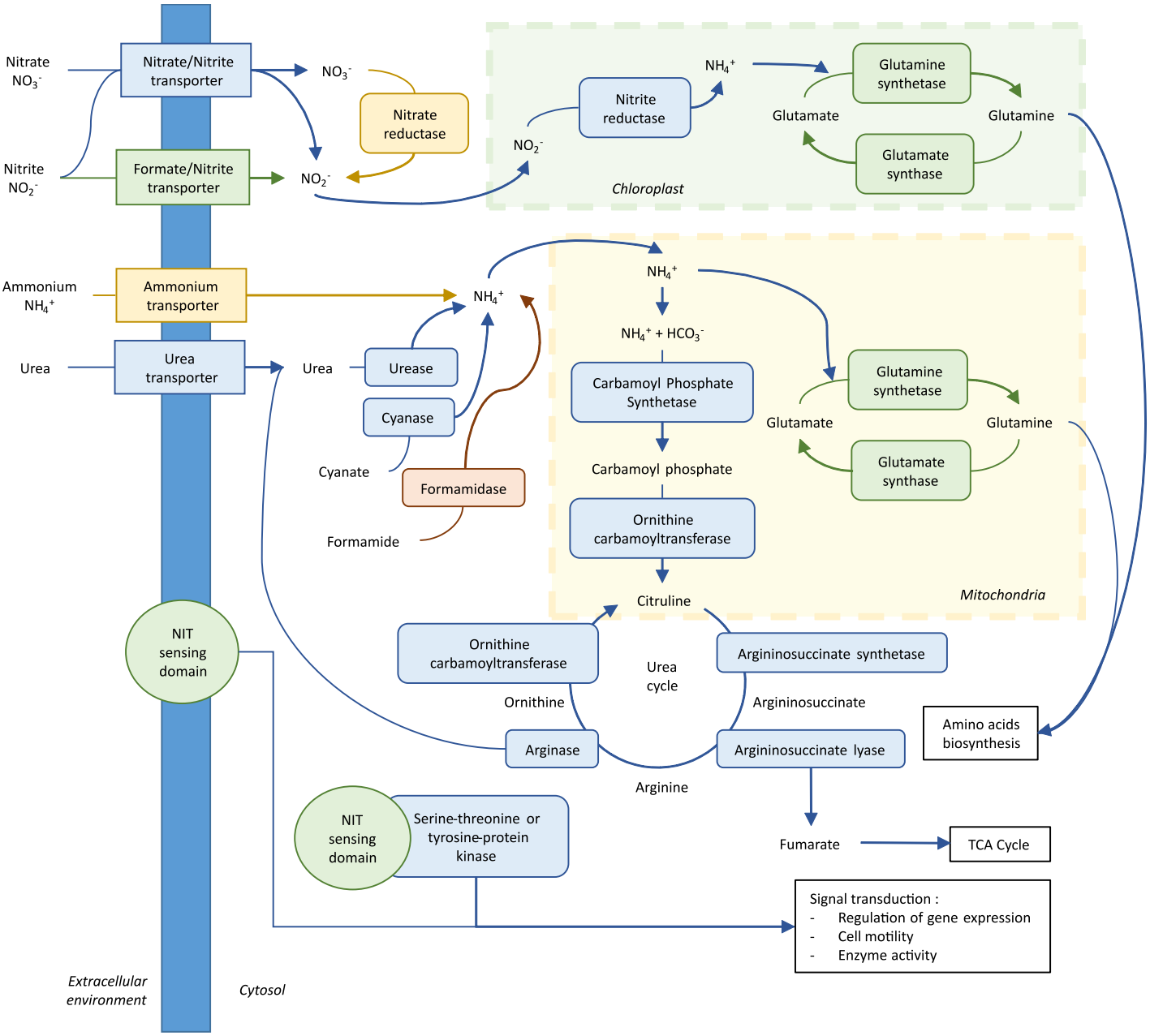

Figure 2: Schematic representation of N transport and assimilation in *P. calceolata* based on the gene content. Rectangles represent transporters of inorganic nitrogenous compounds, ovals represent enzymes and chemicals are unboxed text. The color code indicates if the number of gene copy for a specific function is overrepresented (green), equally represented (blue), underrepresented (orange) or absent (red) in *P. calceolata* compared to the mean of 8 pico-nano photosynthetic species. The number of gene copy for each function is indicated in Table S7.

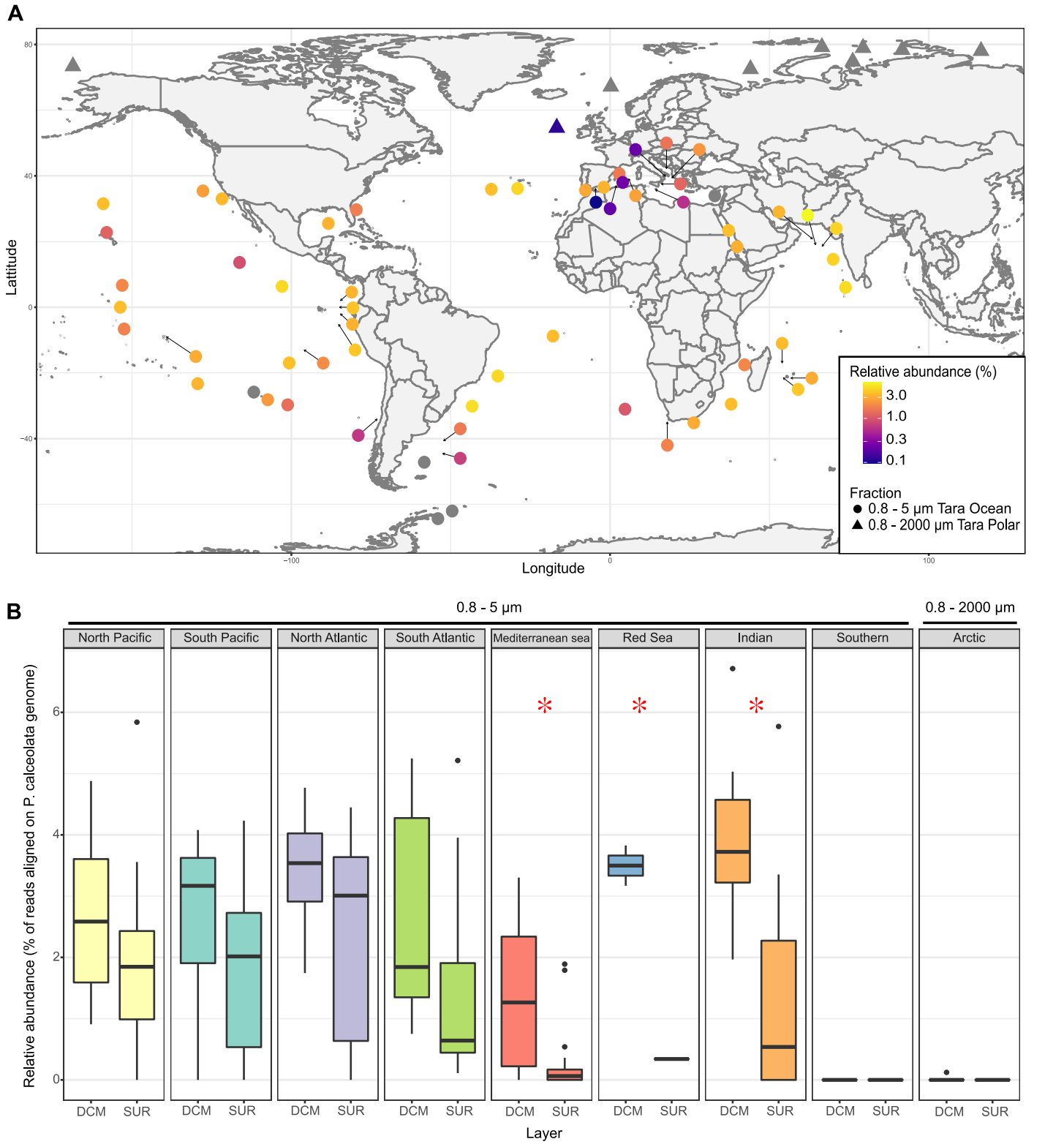

Figure 3: Relative abundance and distribution of *P. calceolata* in the oceans. A) World map of the relative abundance of *P. calceolata*. The colour code indicates the percentage of sequenced reads aligned on *P. calceolata* genome. The DCM samples of size-fractions 0.8-5µm (circles) or 0.8-2000 µm (triangles) are shown. *P. calceolata* is considered to be absent if the horizontal coverage is below 25% of the genome (grey dots). B) Boxplot of the relative abundance of *P. calceolata* in each oceanic region in surface and DCM samples. Red stars indicate a significant difference between SUR and DCM samples (Wilcoxon test, p-value<0.01).

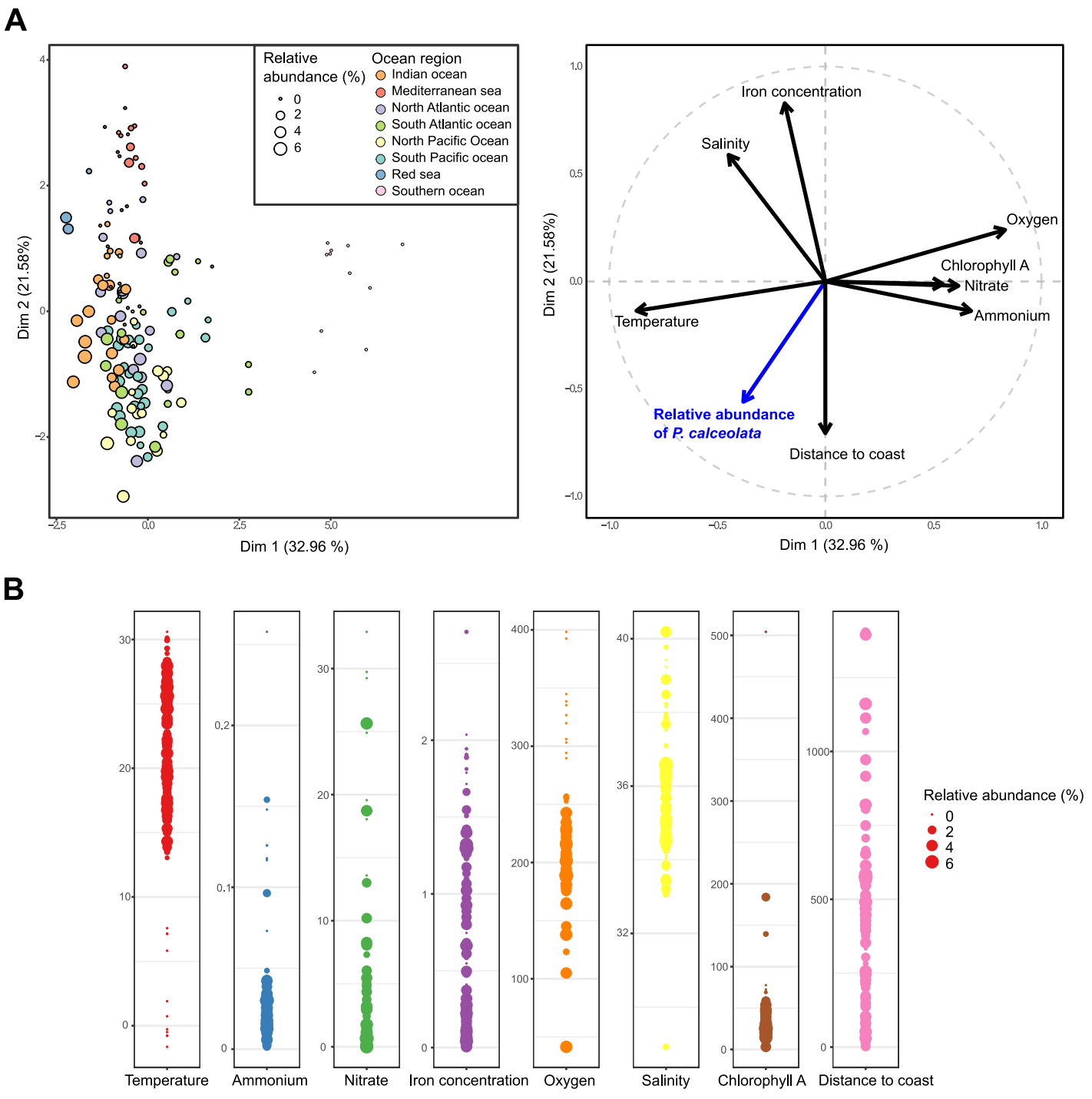

Figure 4: Ecological niche of *P. calceolata*. A) Principal component analysis of the relative abundance of *P. calceolata* and the 8 environmental parameters for the same samples. Each dot represents a sample with a size proportional to the relative abundance of P. calceolata. The colors indicate the oceanic basins. B) Bubble plot of the relative abundance of *P. calceolata* for the 0.8-5 µm size-fraction in surface and DCM samples. Dot sizes indicate the percentage of sequenced read for each environmental parameter.