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

Nina Guérin[1,2], Marta Ciccarella[1], Elisa Flamant[1,2], Paul Frémont[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], Olivier Jaillon[1,2], Corinne Cruaud[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

The smallest phytoplankton species are key actors in oceans biogeochemical cycling and their abundance and distribution are affected with global environmental changes. Picoalgae (cells <2µm) of the Pelagophyceae class encompass coastal species causative of harmful algal blooms while others are cosmopolitan and abundant in open ocean ecosystems. Despite the ecological importance of Pelagophytes, 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 representing centromeres. These particular genomic structures could be explained by the absence of genes from a recombination pathway involving double Holiday Junctions. We identified a large repertoire of genes involved in inorganic nitrogen sensing and uptake and several genes replacing iron-requiring proteins potentially explaining *P. calceolata* ecological success in oligotrophic waters. Finally, based on this high-quality assembly, we evaluated *P. calceolata* relative abundance in all oceans using environmental *Tara* Oceans datasets. Our results suggest that *P. calceolata* is one of the most abundant eukaryotic species in the oceans with a relative abundance favoured by high temperature and iron-poor conditions. Climate change projections based on its relative abundance suggest an extension of the *P. calceolata* habitat toward the poles at the end of this century. Collectively, these findings reveal the ecological importance of *P. calceolata* and lay the

foundation for a global scale analysis of the adaptation and acclimation strategies of picoalgae in a changing environment.

Introduction

Marine phytoplankton account for more than 45% of photosynthetic 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, present in all oceans, are 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-to-volume ratio of PPEs compared to larger phytoplankton cells is advantageous for resource acquisition and 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 have 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 global 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 chloroplast 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[12]. A global scale analysis of pelagophyte 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.

102 Here we sequenced, assembled and annotated the *Pelagomonas calceolata* nuclear genome, with a
103 combination of long- and short-reads. We examined genomic structure and gene content relative to
104 other unicellular phytoplankton. Specific analyses were performed to gain new insights into its life
105 cycle and its genetic capacity for nutrient uptake. Finally, we used this genome to detect *P. calceolata*
106 in environmental datasets of *Tara* expeditions across all oceans, to characterize its ecological niche
107 and to identify the environmental conditions controlling its relative abundance.

108 Material and Methods

109 **Pelagomonas culture**

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

115 **DNA extraction, library preparation and sequencing**

116 We pelleted cells from 500 ml of culture by two successive centrifugations at 10,000 g for 15 minutes
117 at 4°C. Genomic DNA was extracted using the NucleoSpin Plant II Mini kit according to the
118 manufacturer's instructions (Macherey-Nagel, Hoerdt, France) with the following exception for the
119 lysis step : 400µL of lysis buffer PL1 and 25µL of proteinase K 25mg/mL were added to strain pellets,
120 and lysates were incubated at 55°C for 1 hour at 900 rpm. DNA quantity and integrity were respectively
121 evaluated on a Qubit 2.0 spectrofluorometer (Invitrogen, Carlsbad, CA, USA) and a Nanodrop1000
122 spectrophotometer (Thermo Fisher Scientific, MA, USA).
123 For Illumina sequencing, DNA (1.5µg) was sonicated using a Covaris E220 sonicator (Covaris, Woburn,
124 MA, USA). Fragments were end-repaired, 3'-adenylated and Illumina adapters (Bioo Scientific, Austin,
125 TX, USA) were then added using the Kapa Hyper Prep Kit (KapaBiosystems, Wilmington, MA, USA).
126 Ligation products were purified with AMPure XP beads (Beckmann Coulter Genomics, Danvers, MA,
127 USA). The library was then quantified by qPCR using the KAPA Library Quantification Kit for Illumina
128 Libraries (KapaBiosystems), and library profile was assessed using a High Sensitivity DNA kit on an
129 Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was sequenced on an
130 Illumina NovaSeq instrument (Illumina, San Diego, CA, USA) using 150 base-length read chemistry in
131 paired-end mode.
132 For Oxford Nanopore Technologies (ONT) sequencing, the library was prepared using the 1D Native
133 barcoding genomic DNA (with EXP-NBD104 and SQK-LSK109). Genomic DNA fragments (1 µg) were

134 repaired and 3'-adenylated with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End
135 Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Adapters with barcode provided
136 by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) were then ligated
137 using the NEB Blunt/TA Ligase Master Mix (NEB). After purification with AMPure XP beads (Beckmann
138 Coulter, Brea, CA, USA), the sequencing adapters (ONT) were added using the NEBNext Quick T4 DNA
139 ligase (NEB). The library was purified with AMPure XP beads (Beckmann Coulter), then mixed with the
140 Sequencing Buffer (ONT) and the Loading Bead (ONT) and loaded on a MinION R9.4.1 flow cell. Reads
141 were basecalled using Guppy 3.1.5.

142 **RNA extraction, library preparation and sequencing**

143 When the cell concentration reached 10 million cells/mL in mid-exponential growth phase, 160 mL of
144 culture were collected by three successive filtrations on 1.2 µm polycarbonate filters of 47mm to avoid
145 prokaryotic contamination. To preserve cell and RNA integrity, we kept filtration time and pressure
146 below 10 min and 20 mmHg, respectively. Then filters were stored in 15 mL Falcon tubes with 3 mL of
147 Trizol (Invitrogen, Carlsbad, CA, USA), mixed and flash-frozen in liquid nitrogen for further processing.
148 RNA was extracted by incubation at 65°C for 15 min, followed by a chloroform extraction. Aqueous
149 phase was purified using Purelink RNA Isolation kit (Ambion Invitrogen, Carlsbad, CA, USA) according
150 to the manufacturer's instructions. DNA contamination was removed by digestion using the TURBO
151 DNA-free™ Kit (Ambion Invitrogen) according to the manufacturer's DNase treatment protocol. After
152 two rounds of 30 min incubation at 37 °C, the efficiency of DNase treatment was assessed by PCR.
153 Quantity and quality of extracted RNA were analyzed with RNA-specific fluorimetric quantitation on a
154 Qubit 2.0 Fluorometer using Qubit RNA HS Assay (Invitrogen). The qualities of total RNA were checked
155 by capillary electrophoresis on an Agilent Bioanalyzer, using the RNA 6,000 Nano LabChip kit (Agilent
156 Technologies, Santa Clara, CA).
157 RNA-Seq library preparation was carried out from 1 µg total RNA using the TruSeq Stranded mRNA kit
158 (Illumina, San Diego, CA, USA), which allows mRNA strand orientation. Briefly, poly(A)+ RNAs were
159 selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using
160 random hexamer priming. After second strand synthesis, double-stranded cDNA was 3'-adenylated
161 and ligated to Illumina adapters. Ligation products were PCR-amplified following the manufacturer's
162 recommendations. Finally, ready-to-sequence Illumina library was quantified by qPCR using the KAPA
163 Library Quantification Kit for Illumina libraries (KapaBiosystems, Wilmington, MA, USA), and evaluated
164 with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was
165 sequenced using 101 bp paired end read chemistry on a HiSeq2000 Illumina sequencer. Low-quality
166 nucleotides (Q < 20) from both ends of the reads were discarded. Illumina sequencing adapters and
167 primer sequences were removed and reads shorter than 30 nucleotides after trimming were discarded.

168 These trimming and cleaning steps were achieved using in-house-designed software based on the
169 FastX package (https://www.genoscope.cns.fr/externe/fastxtend/). The last step identifies and
170 discards read pairs that are mapped to the phage phiX genome, using the SOAP aligner[26] and the
171 Enterobacteria phage PhiX174 reference sequence (GenBank: NC_001422.1). This processing,
172 described in Alberti et al, resulted in high-quality data[27]. Moreover, ribosomal RNA-like reads were
173 excluded using SortMeRNA[28].

174 **Long-read based genome assembly**

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

186 **Repeat masking and GC analyses**

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

197 **Transcriptome assembly**

198 RNA sequencing reads from *P. calceolata* RCC100 were assembled using Velvet 1.2.07 and Oases 0.2.08
199 with a k-mer size of 63 bp[42,43]. Reads were mapped back to the contigs with BWA-mem[44] and only
200 consistent paired-end reads were kept. Uncovered regions were detected and used to identify

201 chimeric contigs. In addition, open reading frames (ORF) and domains were searched using
202 respectively TransDecoder (http://transdecoder.sourceforge.net) and CDDsearch[45]. Contig extremities
203 without predicted ORFs or functional domains were removed. Lastly, we used the read strand
204 information to correctly orient RNA contigs. We completed the RNA contigs dataset with the two
205 transcriptome assemblies of the RCC100 strain of *P. calceolata* from the Marine Eukaryotes
206 Transcriptomes database (METdb) (http://metdb.sb-roscoff.fr/metdb/)[46].

207 **Gene prediction**

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

221 We selected alignments from the newly generated transcriptome assembly and the two assemblies
222 available in METdb belonging to the *P. calceolata* CCMP1214 strain to build a training set for
223 AUGUSTUS *ab initio* gene predictor. Only gene models with complete coding DNA sequences were
224 retained for training and 1,000 genes were set aside for testing AUGUSTUS accuracy. Initial training
225 produced exon and intron parameters for *P. calceolata* species. Parameters were optimized using
226 successive steps of training and testing. We calculated the accuracy of gene prediction by running
227 AUGUSTUS on the test set. At the exon level, AUGUSTUS performed well in terms of sensitivity (0.619)
228 and specificity (0.669). We thus run AUGUSTUS on the masked genome based on trained parameters.
229 The *ab initio* prediction and all the transcriptomic and protein alignments were combined using Gmove
230 which is an easy-to-use predictor with no need for a pre-calibration step[47]. Briefly, putative exons and
231 introns, extracted from predictions and alignments, were used to build a graph, where nodes and
232 edges represent exons and introns respectively. From this graph, Gmove extracts all paths and
233 searches open reading frames (ORFs) which are consistent with the protein evidence. We trimmed
234 untranslated transcribed regions that overlap the 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. Chloroplast 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 chloroplast 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.

NIT-sensing domain containing genes were identified based on Interproscan annotations. Eukaryotic homologs of the 3 *P. calceolata* NIT-sensing genes were retrieved with a BLASTP (e-value < 10^{-5}, coverage > 100 aa) against 27.7 million proteins from NR, the METdb[46] transcriptome database, *Tara* Oceans single-cell amplified genomes and metagenome assembled genomes (SMAGs)[55] and eukaryotic algal proteomes from the JGI database. The 60 retrieved proteins and the 3 *P. calceolata* NIT-domain containing proteins were then aligned with Mafft v7.0 (https://mafft.cbrc.jp/alignment/server/) and a Maximum Likelihood phylogenetic tree (Jones-Taylor-Thornton substitution model and 100

268 bootstraps) was made with MEGAX software. Transmembrane regions in NIT-domain containing
269 proteins were identified with TMHMM v 2.055.

270 **Mapping and filtering of environmental metagenomic reads.**

271 We used metagenomics datasets of *Tara* Oceans and *Tara Polar Circle* expeditions to detect the *P.*
272 *calceolata* genome in the oceans. Datasets from water samples collected on the photic zone: surface
273 (SUR) and deep-chlorophyll maximum (DCM) were analysed. Size-fractionated water samples
274 containing pico-nano algae (organisms <5µm in cell diameter) were selected: 0.2-3 µm (100 samples),
275 0.8-2000 µm (119 samples) and 0.8-5 µm (148 samples)[27]. Metagenomic reads were aligned on the
276 *Pelagomonas calceolata* genome with BWA-mem version 0.7.15 with default parameters[56].
277 Alignments with at least 90% of identity over 80% of read length were retained for further analysis. In
278 the case of several possible best matches, a random one was picked. In order to remove putative PCR
279 duplicates, multiple read pairs aligned at the exact same position on the *P. calceolata* genome were
280 removed with samtools rmdup version 1.10.2[44].

281 **Relative abundance calculation.**

282 The relative abundance of *P. calceolata* was calculated from both metabarcoding data and
283 metagenomic data. For the metabarcoding abundance, we used the 18SV9 rRNA OTU table published
284 in 2021 and available here https://zenodo.org/record/3768510#.YEX2S9zjJaQ[57]. Bacterial and archaea
285 OTUs were removed for this analysis. For the metagenomic abundance, we divided the total number
286 of reads aligned on the *P. calceolata* genome by the total number of sequenced reads for each sample.

287 ***P. calceolata* relative abundance modelling**

288 Two *P. calceolata* relative abundance models were performed based on the *in situ* environmental
289 conditions measured at the time of sampling or using the World Ocean Atlas 2018 (WOA18) datasets.
290 The environmental parameters measured during the expedition are available in the Pangaea database
291 (https://www.pangaea.de/) and are described in[58]. Iron concentrations are annual means derived from
292 PISCES2 model[59] and described in[12]. Ammonium concentrations at the date and location of sampling
293 are derived from the MITgcm Darwin model and available in the Pangaea database[60]. Environmental
294 parameters for each sample are available in Table S3. Pearson's correlation between the relative
295 abundance of *P. calceolata* and all environmental parameters were calculated with the *cor* function in
296 R and the GGally package version 2.1.0. The 8 parameters with a correlation above 0.2 or below -0.2
297 were selected for downstream analysis. Principal component analysis (PCA) was performed with the R
298 package FactoMineR version 2.4. We used a Generalized Additive Model (GAM) for its ability to fit non-
299 linear and non-monotonic functions and for its low sensitivity to extreme values to model the relative
300 abundance of *P. calceolata* as a function of iron concentration and temperature[61]. This function is

301 implemented in the mgcv R package version 1.8-33. All figures were generated with R version 4.0.3
302 and ggplot2 package version 3.3.2.
303 The mean of several climatologies of the Earth System Models under RCP8.5 scenario (ESM8.5) were
304 used to define the environmental conditions at the end of the century following the method of
305 Frémont et al[62]. *P. calceolata* relative abundance modelling based on the World Ocean Atlas 2018
306 (WOA18) at locations, depths and months of *Tara* Oceans samples or ESM8.5 were obtained using four
307 machine learning techniques[62]. Briefly, Generalized Additive Models (gam), Neural Networks (nn),
308 Random Forest (rf) and Gradient Boosted Trees (bt) were trained in regression mode. Over 100-fold
309 random cross validation, all models performed well: mean Pearson correlation coefficient were
310 respectively 0.64 (gam), 0.67 (nn), 0.71 (rf) and 0.70 (bt). We used the ensemble model approaches[63]
311 for final global-scale modelling of the relative abundance of *P. calceolata* (i.e. the mean projections of
312 the validated machine learning techniques).

313 # Results

314 **A compact genome revealed by a chromosome-scale assembly.**

315 To investigate its gene repertoire and its distribution across the oceans we sequenced and assembled
316 the genome of *P. calceolata* using ONT long-reads and Illumina short-reads. Using k-mer distribution
317 of short reads, the genome was estimated to be homozygous with a size of 31 Mb (Figure S1). The ONT
318 long-reads were assembled using Flye into 6 nuclear contigs for a total of 32.4 Mb, 1 plastid circular
319 contig (90 Kb) and 1 mitochondrial circular contig (39 Kb) (Figure 1 and Figure S2A,B). Two large and
320 highly similar duplicated regions (>99% of identity) were detected at the extremity of contig 1 and 5
321 (393 Kb) and at the extremities of contig 3 and 6 (192 Kb). The vertical read coverage of these two
322 regions is similar to that of other genomic regions suggesting that they are duplicated in all cells of this
323 *P. calceolata* culture (Figure 1 and Figure S2C). In addition, 150 Kb at one extremity of contig 4 present
324 a higher vertical coverage suggesting that this region is also duplicated in the *P. calceolata* genome but
325 collapsed in the assembly (Figure S2). Interestingly, duplicate regions at the end of sequences have
326 already been observed in the chlorophyte *Ostreococcus tauri* which has been maintained for several
327 years in culture[64]. These observations suggest that culture could affect not only the sequence of a given
328 organism's genome but also its structure. (TTAGGG)n telomeric repeats were detected at both ends of
329 contigs 2, 3, 4 and 6 indicating that these 4 contigs represent complete chromosomes (Figure S2). For
330 contig 1 and contig 5, telomeric sequences are present at only one extremity, the other extremity
331 ending in the duplicated region. This result suggests that the six contigs correspond to six
332 chromosomes of *P. calceolata*.

333 A total of 16,667 genes was predicted on the *P. calceolata* genome with an average of 1.45 intron per
334 gene (Table 1). The distribution of intron lengths reveals a peak at around 210 bp (Figure S3A), which
335 is the characteristic length of Introner Elements (IE) described in *A. anophagefferens*[65]. Putative IE were
336 then identified (n=956, see method). Both canonical (GT) and non-canonical (GC) donor splicing sites
337 were present at the ends of putative IE, while the acceptor site was canonical (AG) in all cases (Table
338 S4). The logo representation of the putative IE sequences reveals the presence of GT and GC donor
339 sites at the 5' end, AG acceptor site at the 3' end, and conserved TIR in the flanking regions (Figure
340 S3B). 9, 583 *P. calceolata* predicted proteins (58%) are homologous with at least one gene in a
341 Stramenopile genome, including 2,631 (16%) only shared with the Pelagophyceae *A. anophagefferens*
342 (Figure S4). A conserved functional domain (Pfam, KO or InterProScan) was found in 11,240 proteins
343 (67%).

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 eukaryotes.

344 **GC content distribution along *P. calceolata* chromosomes**

345 A remarkable feature in *P. calceolata* genome is the distribution of GC content along *P. calceolata*
346 chromosomes (Figure 1). While the average GC content of the nuclear genome is 63%, one large region
347 in each contig (259 Kb in average) is 52% GC. These unique large troughs in GC content in each
348 chromosome suggest that these regions may encompass centromeres. We did not observe
349 accumulation of repeated elements or transposons in these low-GC regions; however, the gene
350 structures differ from that of other genomic regions (Figure 1 and Table S5). The slight decrease of
351 gene density observed in Figure 1 is explained by a longer gene size (average of 3 Kb compared to 1.9
352 Kb in other regions). In addition, the 453 genes in low-GC regions contain more introns (2.56 introns
353 per gene compared to 0.49) and the introns are shorter (120 compared to 214 bases). We also noticed
354 a higher proportion of bases belonging to intergenic regions, 18% in low-GC regions and 11% in other
355 chromosomic regions (Table S5). Interestingly, three repeats of more than 500 bases are present in
356 several low-GC regions (R_13 in contigs 2, 3 and 6; R_25 in all contigs except contig 1 and R_80 in

357 contigs 1 to 4) (Figure 1). No homology was found between these sequences and known repeat
358 elements.
359 Gene function analysis of low-GC regions reveals an enrichment of genes involved in specific cellular
360 mechanisms (Table S6). Thirteen genes are involved in DNA replication including the Anaphase-
361 promoting complex subunit 4, the sister chromatid cohesion protein Dcc1 and 3 Mini-chromosome
362 maintenance genes. Twenty-five genes are involved in microtubule synthesis and microtubule-binding
363 motor proteins (9 genes carrying dynein domains, 6 genes carrying kinesin motor domains and 4
364 tubulin genes). These genes indicate that the low-GC regions contain many genes required for *P.*
365 *calceolata* cellular division. Finally, 18 genes are involved in transcription including 3 genes encoding
366 RNA Pol II rpb2 subunits and 7 genes encoding transcription factors suggesting an important role of
367 these chromosomic regions for the regulation of gene expression.

368 **Sex related genes in *P. calceolata***

369 Because low-GC regions could be related to meiotic recombination[74], we looked for genes involved in
370 sexual reproduction in the *P. calceolata* genome. Among 20 genes specifically involved in meiosis, 16
371 homologs are present in the *P. calceolata* genome (Table 2). These genes include the double-strand
372 DNA break (DSB) initiator SPO11; RAD50, RAD52 and MRE11 to bind DSBs; HOP2, MND1, DMC1 and
373 RAD51 to ensure pairing and invade the homologous strand; MSH2, MSH3 PMS1 and MSH6 genes
374 involved in the synthesis-dependent strand annealing pathway and MUS81 necessary for non-
375 interfering (class II) crossing over. Interestingly, MSH4 and MSH5 genes are absent from the *P.*
376 *calceolata* genome. Indeed, these genes necessary to perform the interfering (class I) recombination
377 pathway through Double Holliday Junctions are present in most eukaryotic lineages. The large low-GC
378 regions could be a consequence of the absence of the MSH4/5 genes as suggested for yeasts[75] (see
379 discussion). There are also no homologs of ZIP, HOP1 and RED1 genes in the *P. calceolata* genome.
380 These genes are known to be involved in homologous pairing of chromatids and construction of the
381 synaptonemal complex in animals, plants and fungi but are not required to perform meiosis. Indeed
382 these 3 genes are absent in several phyla capable of meiosis like diatoms[76] and ciliates[52]. Taken
383 together, the genetic content of *P .calceolata* strongly suggests that this species perform 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 Pca_1p10430	K08735	3.5E-190* 4.00E-152	MSH2	Dna mismatch repair protein ; Forms a heterodimer with Msh3 or Msh6
Pca_2p23180	K08736	1.50E-150	MSH3	Dna mismatch repair proteinn ; Forms a heterodimer with Msh2
Absent Absent			MSH4 MSH5	The meiosis-specific MutS homologs, MSH4 and MSH5, function as a heterodimer and have specialized roles in meiotic recombination and Holliday junction resolution
Pca_1p10290 Pca_4p14280	K08737	0* 4.70E-50	MSH6	Forms a heterodimer with Msh2; binds base mismatches
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.

384 **Genes involved in nitrogen uptake, storage and recycling.**

385 *P. calceolata* could be an important player in the nitrogen (N) cycle in oceanic ecosystems[24], therefore
386 we explored the genomic capacities of *P. calceolata* to assimilate and use nitrogen-containing
387 compounds. We systematically identified and counted genes involved in the nitrogen cycle in *P.*
388 *calceolata* compared to seven other pico-nano photosynthetic eukaryotes (Figure 2 and Table S7).
389 The uptake of nitrogen-containing inorganic compounds is supported by 13 genes in *P. calceolata*.
390 Among them, 8 genes encode nitrate/nitrite or formate/nitrite transporters, which is on average
391 higher than in other algae. In contrast, only 5 genes encode ammonium transporters which is low
392 compared to other species suggesting that nitrite and/or nitrate is the main external source of
393 inorganic nitrogen for *P. calceolata*. The number of enzymes incorporating ammonium into organic
394 compounds (GS/GOGAT pathway) are higher in *P. calceolata* than in other species: 5 glutamine
395 synthetase and 4 glutamate synthase genes are present in the *P. calceolata* genome. On the contrary,
396 fewer genes assure the reduction of nitrate into nitrite (1 nitrate reductase) then nitrite into
397 ammonium (2 nitrite reductases) (Table S7).
398 We identified 3 genes carrying the nitrate and nitrite sensing (NIT) domain (IPR013587) in the *P.*
399 *calceolata* genome. One *P. calceolata* protein carries an NIT-sensing domain surrounded by 2

400 transmembrane domains suggesting a capacity for external Nitrate/Nitrite sensing while the 2 other
401 NIT genes carry a protein-kinase domain (IPR000719) suggesting a phosphorylation-based signal
402 transduction dependent on intracellular nitrate or nitrite concentration (Figure S5). Using NCBI non-
403 redundant proteins and marine genomic databases (see methods), 60 homologous proteins of the NIT-
404 sensing domains of *P. calceolata* were identified. These homologs are restricted to the Pelagophyceae
405 class (16 transcriptomes), the Dictyochophyceae class (6 transcriptomes) and one putative
406 Cryptophyte transcriptome. The phylogenetic tree of this protein family shows three subfamilies
407 diverging before the Dictyochophyte/Pelagophyte separation (Figure S5).
408 We then identified genes involved in nitrogen recycling from organic compounds which are important
409 in several species in case of inorganic nitrogen deprivation. One arginase gene and one cyanase gene
410 were detected in the *P. calceolata* genome but no gene encoding formamidase. In addition, the
411 number of gene copies for enzymes involved in the urea cycle (carbamoyl-phosphate synthetase,
412 ornithine carbamoyltransferase, argininosuccinate synthase and argininosuccinate lyase) is equal or
413 slightly lower than in other algae (Figure 2 and Table S7). These results suggest that *P. calceolata* is not
414 particularly adapted to recycle nitrogen from organic molecules but could be capable of incorporating
415 inorganic nitrogen compounds even in N-poor environments.

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

417 Iron is a critical metal for all photosynthetic organisms. Iron-containing molecules are essential for the
418 photosynthesis, the nitrogen cycle and the protection against reactive oxygen species. We identified
419 gene coding for iron uptake and storage in *P. calceolata* genome and compared them to other small
420 photosynthetic eukaryotes (Table S7). *P. calceolata* has 5 genes encoding the phytotransferrin ISIP2
421 involved in Fe^{3+} uptake via endosomal vesicles and 2 putative iron storage protein ISIP3. These genes
422 are transcribed following starvation in diatoms suggesting a capability for thriving in iron-poor
423 environments for *P. calceolata*. In addition, 3 genes encode the iron transporter ferroportin. These
424 proteins are iron exporters in multicellular organisms but their function in micro-algae remains to be
425 studied[77]. Zinc/iron permeases, transmembrane ferric reductases and multicopper oxidase genes,
426 involved in iron uptake from the environment, are present in the *P. calceolata* genome (8, 5 and 2
427 genes respectively) but in equal or lower gene copy number than in other studied species. The iron
428 permease FRT1 and the ISIP1 gene, involved in endocytosis of iron-chelator molecules (siderophores),
429 are absent in the *P. calceolata* genome. Furthermore, *P. calceolata* has no gene encoding the iron
430 storage protein ferritin. The absence of these genes suggests that iron uptake and storage is not a
431 major asset of *P. calceolata* compared to the other photosynthetic protists.
432 Several important ferrous proteins can be substituted by non-ferrous equivalents in iron-poor
433 environments[13]. In the *P. calceolata* genome we identified 11 flavodoxin and 3 phytocyanin genes,

434 encoding non-ferrous proteins involved in electron transfer during photosynthesis potentially
435 replacing Ferredoxin and Cytochrome C_6 respectively. The Fructose-bisphosphate Aldolase (FBA)
436 necessary for gluconeogenesis and the Calvin cycle is encoded by 6 genes in *P. calceolata*. Two genes
437 are dependent on a bivalent cation (FBA type II), the four others are Zinc/Iron-independent (FBA type
438 I). Finally, all types of Superoxide dismutases (SOD) are found in the *P. calceolata* genome. Non-ferrous
439 SODs (Cu/Zn and Ni) are encoded by 3 genes, while the Mn/Fe-SOD is encoded by 2 genes. The genetic
440 content of *P. calceolata* indicates that several essential processes are potentially iron-independent
441 thanks to the presence of many iron-free alternative proteins.

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

443 First, we used the OTU table computed from metabarcoding samples of the *Tara* Oceans expedition
444 to estimate the relative abundance of *P. calceolata* across all oceans[57]. The relative rRNA abundance
445 of *P. calceolata* OTU in the 0.8-5 µm size-fraction is 1.30% on average for the 111 surface samples and
446 0.81% on average for the 62 DCM samples (Table S8). According to this method of abundance
447 estimation, *P. calceolata* is the third most abundant eukaryote of the 0.8-5 µm size-fraction among
448 *Tara* samples covering all oceans except the Arctic Ocean. However, the number of rRNA copies in
449 each organism biases this metabarcoding-based abundance estimation. Therefore, we used the
450 mapping of metagenomic reads on the *P. calceolata* genome to estimate more precisely its abundance.
451 The relative abundance obtained with the entire genome is strongly correlated to the metabarcoding-
452 based relative abundance (Pearson correlations of 0.91 and 0.70 for 0.8-2000 and 0.8-5 µm size-
453 fractions respectively). However, the metabarcoding-based abundance is underestimated by a factor
454 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
455 metagenomic-based abundance (Figure S6). This underestimation is probably due to the low copy
456 number of rRNA in *P. calceolata* (2 complete copies) compared to most other species with larger
457 genomes.
458 The relative abundance of *P. calceolata* calculated with the number of reads aligned on the genome is
459 estimated above 1% in 66 oceanic stations with a maximum of 6.7% for the 0.8-5 µm size-fraction at
460 the DCM in the North Indian Ocean (Figure 3A and Figure S7). In the Indian Ocean, Red Sea and
461 Mediterranean Sea, *P. calceolata* is significantly more abundant in the DCM than at the surface (Figure
462 3B). In cold waters (below 10°C), *P. calceolata* is not detected above our threshold of 25% of genomic
463 horizontal coverage. Important variations between and within each oceanic basin are observed,
464 suggesting that many biotic or abiotic factors influence *P. calceolata* abundance.
465
466 **High relative 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 4A and B). The other parameters do not seem to be related to *P. calceolata* abundance. This result is consistent over 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 capable of acclimation 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.

Finally, we projected the relative abundance of *P. calceolata* at the end of the century. We modelled the ecological niche of *P. calceolata* using the WOA18 datasets at the time and location of sampling or using the projected climatology in 2099 using the RPC8.5 scenario (see methods). We projected an increase (from +0.4% to +0.92%) of *P. calceolata* relative abundance from latitude 40° to latitude 50° in the North and South hemispheres and a slight decrease in temperate and tropical waters (-0.5% maximum) (Figure 4C and Figure S9). The decrease in the tropical waters is uncertain (stars on the world map) because this environment in 2099 is out of the range of the training dataset.

Discussion

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

495 **Large low-GC centromeres**

496 The chromosome level assembly of the *P. calceolata* genome thanks to nanopore read sequencing
497 revealed unique genomic features that are not usually studied in short-read assembled genomes. The
498 most striking observation is the presence of a unique GC trough (50%) in each scaffold contrasting with
499 the high GC content (63%) of other genomic regions of *P. calceolata*. In eukaryotes, centromeres have
500 a large variety of structures and characteristics. They are composed of many repeated sequences or
501 contain genes, they are determined genetically or epigenetically and their size can vary from 125 bases
502 (in *S. cerevisiae*) to several Mb (in mammals)[78]. Short regional centromeres (1-5 Kb) are generally low-
503 GC compared to the genome and the sequence is unique at each centromere. Among stramenopiles,
504 centromeres were characterized in the diatom *P. tricornatum* where a short low-GC sequence (>500
505 bp) is enough to define a centromere[79]. Large regional centromeres (>10 Kb) are generally gene-free,
506 contain repeated elements and are not transcribed[78]. *P. calceolata* putative centromeres derive from
507 this general pattern with large low-GC regions containing genes carrying essential functions of DNA
508 replication, transcription and microtubule assembly. Interestingly, the red alga *Cyanidioschyzon* as well
509 as several yeast species seem to have similar centromere structures with large, low-GC centromeres
510 containing genes[80,75].

511 **GC content, meiosis and recombination**

512 The main hypothesis to explain these low-GC patterns in centromeres is the importance of GC-biased
513 gene conversions (gBGC) during recombination and the inhibition of this recombination near
514 centromeres[81]. Indeed, gBGCs increase the GC content of recombining DNA over evolutionary time
515 inducing GC content variations within and between genomes[74]. The kinetochore formation at
516 centromeric regions inhibits recombination and double strand break formation during meiosis
517 resulting in rare gBGC in these regions[82]. Centromeric and peri-centromeric regions may therefore
518 have a lower GC content. Furthermore, the genetic content of *P. calceolata* indicates that this species
519 is capable of meiosis despite the absence of some genes including MSH4 and MSH5 necessary to
520 perform meiotic recombination through double Holliday junctions. In yeasts (*Yarrowia lipolytica*,
521 *Candida lusitaniae*, and *Pichia stipites*), a correlation has been observed between the importance of

522 the GC trough near the centromeres (>10%) and the absence of MSH4/MSH5 genes[75]. It is therefore
523 possible that the absence of this recombination pathway in *P. calceolata* induces more frequently
524 double-strand break repair by synthesis-dependent strand annealing and a more rapid GC-biased
525 conversion across the genome except at the centromeres where double-strand breaks are inhibited.
526 This recombination inhibition may have important consequences on the evolution of the *P. calceolata*
527 genome. Genes within low-GC regions are significantly longer and contain more introns than genes in
528 other genomic regions. Because intron gain and loss are closely related to double-strand break repair
529 and homologous recombination, we suggest that centromere genes retain more introns because
530 double-strand breaks are reduced[83]. Variant analysis in *P. calceolata* populations could be targeted
531 specifically in future studies to infer an estimation of recombination rate and more generally
532 characterize the evolutionary processes controlling these large centromere regions.

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

534 We have shown in this study that *P. calceolata* is cosmopolitan in oceanic samples above 10°C with a
535 relative abundance generally >1% of all sequenced reads. In contrast to the coastal Pelagophyceae *A.*
536 *anophagefferens* that can present high peaks of abundance[84] no *P. calceolata* blooms are reported but
537 *P. calceolata* is well-adapted to a large range of environmental conditions as suggested by previous
538 studies[20,22].
539 Although the abundance of an organism calculated from metabarcoding or metagenomic data provide
540 only an indirect and relative quantification of organism abundances, both methods suggest that *P.*
541 *calceolata* is one of the most abundant pico-nano eukaryote in offshore data. The high relative
542 abundance of *P.calceolata* measured with a metabarcoding approach has recently been confirmed
543 with a qPCR method[20]. In addition, we have shown that the metabarcoding approach probably
544 underestimates the relative abundance of *P.calceolata* compared to the metagenomic analysis owing
545 to the low copy number of rRNAs in this organism. Further studies may combine microscopic and flow
546 sorting approaches with genomic data to assess the number of cells and the biomass of this organism
547 in the oceans. Our modelling analysis has revealed a probable increase of *P. calceolata* relative
548 abundance at the end of the century in high latitudes where the temperature is today too cold for this
549 species. This result is coherent with previous studies suggesting a global increase of phytoplankton in
550 subpolar regions[85,86].

551 **Iron uptake and storage and modulation of iron needs**

552 Iron is an essential element for the growth, photosynthesis, primary production, nitrogen fixation and
553 reduction for PPEs[87]. Our results show that *P. calceolata* thrives in iron-poor waters and thus occupies
554 a large ecological niche for a PPE. Two main strategies exist against iron deprivation: 1) optimisation

555 of iron uptake and 2) modulation of iron needs. Optimisation of iron uptake does not seem to be the
556 main strategy of *P. calceolata* in iron-poor environments. Genes coding for iron chelators and ferritin
557 are absent from its genome, and gene coding for passive iron transporters are under- or equally
558 represented compared to other PPEs. In contrast, phytotransferrins (ISIP2), putative iron storage
559 proteins (ISIP3) and ferroportins are overrepresented in the *P. calceolata* genome. In Pelagophyceae,
560 expression levels of ISIP genes are correlated with iron concentration at the global scale suggesting a
561 transcriptomic regulation in response to iron availability[12]. Because phytotransferrins are dependent
562 on carbonate ions, ocean acidification may reduce the efficiency of iron uptake in many species[88,89]. In
563 consequence, organisms able to grow with very low iron concentration like *P. calceolata* may be
564 favoured.
565 The presence of three ferroportin genes in *P. calceolata* is interesting since these transmembrane iron-
566 export proteins play a major role in iron homeostasis in multicellular organisms[77]. Ferroportin function
567 in micro-algae is unknown but could act to export iron from endosomes to the cytoplasm[90]. In the
568 green alga *Chlamydomonas reinhardtii*, a ferroportin gene is overexpressed under low Fe conditions[91].
569 The function of the 3 ferroportin genes of *P. calceolata* could be investigated to understand their role
570 in iron-poor environments. In addition, modulation of iron needs seems to be a major strategy for *P.*
571 *calceolata*. All known molecular switches between ferrous and non-ferrous proteins are genetically
572 possible in *P. calceolata* and the non-ferrous encoded protein genes FBA I, flaovodoxin and
573 phytocyanin present a higher number of gene copies than the ferrous ones. It could be a sign that
574 those switches are easier for *P. calceolata* than for other PPEs, but transcriptomic experiments are
575 required to determine if this genetic overrepresentation has metabolic consequences.

576 **Nitrogen cycle**

577 Expressing more than 90% of all nitrate transporter transcripts, Pelagophytes may dominate nitrate
578 uptake and assimilation in the North Pacific Ocean[24]. Indeed, *P. calceolata* contains a large collection
579 of genes for nitrate, nitrite and urea transporters. Interestingly, a few genes are ammonium
580 transporters suggesting that the main source of nitrogen for *P. calceolata* is nitrite and/or nitrate. One
581 remarkable feature of the *P. calceolata* genome is the presence of 3 genes carrying NIT domains
582 (PF08376). This NIT domain was first described in bacterial nitrite and nitrate sensor proteins[92]. This
583 sensor is an alpha-helical protein playing a signal transduction role regulating gene expression, cell
584 motility and enzyme activity in *Klebsiella oxytoca*[93]. In pico-nano algae, NIT-sensing domains can be
585 associated with a serine-threonine/tyrosine-kinase domain, suggesting signal transduction according
586 to the concentration of intracellular nitrate/nitrite, or surrounded by 2 transmembrane domains
587 suggesting extracellular sensing. Even though NIT-sensing domains can be found across various phyla,
588 homologs of *P. calceolata* NIT proteins are restricted to Pelagophytes and Dictyochophytes. The NIT

589 genes in *P. calceolata* are highly expressed in subtropical Pacific N-depleted waters, suggesting that
590 these proteins have a role in transcription regulation according to nitrate availability[24].

591 Conclusion

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

603 Data availability

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

608 Acknowledgments

609 We thank the commitment of the following people who made this work possible: the Genoscope/CEA,
610 the CNRS (in particular the Federation de Recherche R2022/Tara Oceans GO-SEE), Marie-José Garet-
611 Delmas from the Roscoff Culture Collection for growing the RCC100 strain, Claude Scarpelli for support
612 in high-performance computing. Computations were performed using the cobalt HPC machine. We
613 acknowledge the financial support of FRANCE GENOMIQUE (ANR-10-INBS-09–08) and Oceanomics
614 (ANR-11-BTBR-0008). We also thank the *Tara* Expedition Foundation and their partners for the
615 organization of marine scientific expeditions (http://oceans.taraexpeditions.org). This article is
616 contribution number XXX of *Tara* Oceans.

617

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.

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

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

698 32. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat Methods *17*,
699 155–158.

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

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

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

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

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

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

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

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

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

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

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

722 44. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and
723 Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–
724 2079.

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

728 46. Niang, G., Hoebeke, M., Meng, A., Liu, X., Scheremetjew, M., Finn, R., Pelletier, E., and Corre, E.
729 (2020). METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES .
730 F1000Research *9*.

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

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

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

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

740 51. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H.
741 (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score
742 threshold. Bioinformatics *36*, 2251–2252.

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

746 53. Ramesh, M.A., Malik, S.-B., and Logsdon, J.M. (2005). A phylogenomic inventory of meiotic
747 genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. Curr Biol *15*, 185–
748 191.

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

751 55. Delmont, T.O., Gaia, M., Hinsinger, D.D., Fremont, P., Vanni, C., Guerra, A.F., Eren, A.M.,
752 Kourlaiev, A., d’Agata, L., Clayssen, Q., et al. (2021). Functional repertoire convergence of
753 distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics.

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

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

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

762 59. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M. (2015). PISCES-v2: an ocean
763 biogeochemical model for carbon and ecosystem studies. Geoscientific Model Development *8*,
764 2465–2513.

765 60. Clayton, S., Dutkiewicz, S., Jahn, O., Hill, C., Heimbach, P., and Follows, M.J. (2017).
766 Biogeochemical versus ecological consequences of modeled ocean physics. Biogeosciences *14*,
767 2877–2889.

768 61. Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A.N. (2019). Generalized additive models:
769 Building evidence of air pollution, climate change and human health. Environment International
770 *132*, 104987.

771 62. Frémont, P., Gehlen, M., Vrac, M., Leconte, J., Delmont, T.O., Wincker, P., Iudicone, D., and
772 Jaillon, O. (2021). Restructuring of genomic provinces of surface ocean plankton under climate
773 change.

774 63. Jones, M.C., and Cheung, W.W.L. (2015). Multi-model ensemble projections of climate change
775 effects on global marine biodiversity. ICES Journal of Marine Science *72*, 741–752.

776 64. Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F.-Y., Carré, I., Château, A.,
777 Eyre-Walker, A., Grimsley, N., Moreau, H., et al. (2014). An improved genome of the model
778 marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. BMC Genomics
779 *15*, 1103.

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

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

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

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

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

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

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

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

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

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

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

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

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

818 78. Talbert, P.B., and Henikoff, S. (2020). What makes a centromere? Experimental Cell Research
819 *389*, 111895.

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

823 80. Kanesaki, Y., Imamura, S., Matsuzaki, M., and Tanaka, K. (2015). Identification of centromere
824 regions in chromosomes of a unicellular red alga, Cyanidioschyzon merolae. FEBS Letters *589*,
825 1219–1224.

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

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

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

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

836 85. Agusti, S., Lubián, L.M., Moreno-Ostos, E., Estrada, M., and Duarte, C.M. (2019). Projected
837 Changes in Photosynthetic Picoplankton in a Warmer Subtropical Ocean. Frontiers in Marine
838 Science *5*, 506.

839 86. Anderson, S.I., Barton, A.D., Clayton, S., Dutkiewicz, S., and Rynearson, T.A. (2021). Marine
840 phytoplankton functional types exhibit diverse responses to thermal change. Nat Commun *12*,
841 6413.

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

88. 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.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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.

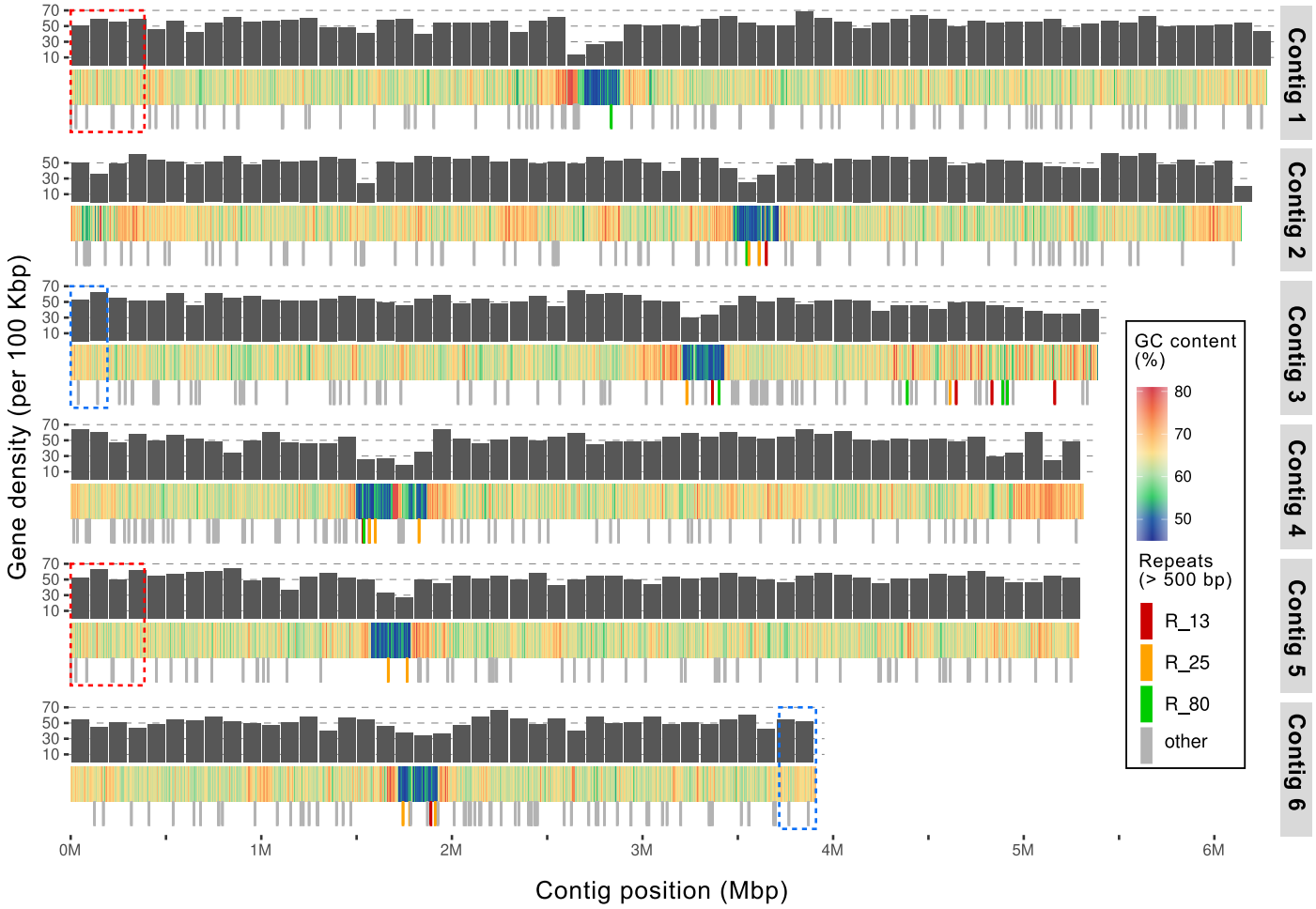

Figure 1: *Pelagomonas calceolata* nuclear genome. Representation of the 6 nuclear contigs of *P. calceolata*. The top layer indicates the number of genes per 100 Kb (black bars), the middle layer represents the GC content in percentage over a window of 200 Kb and the bottom layer is the position of 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.

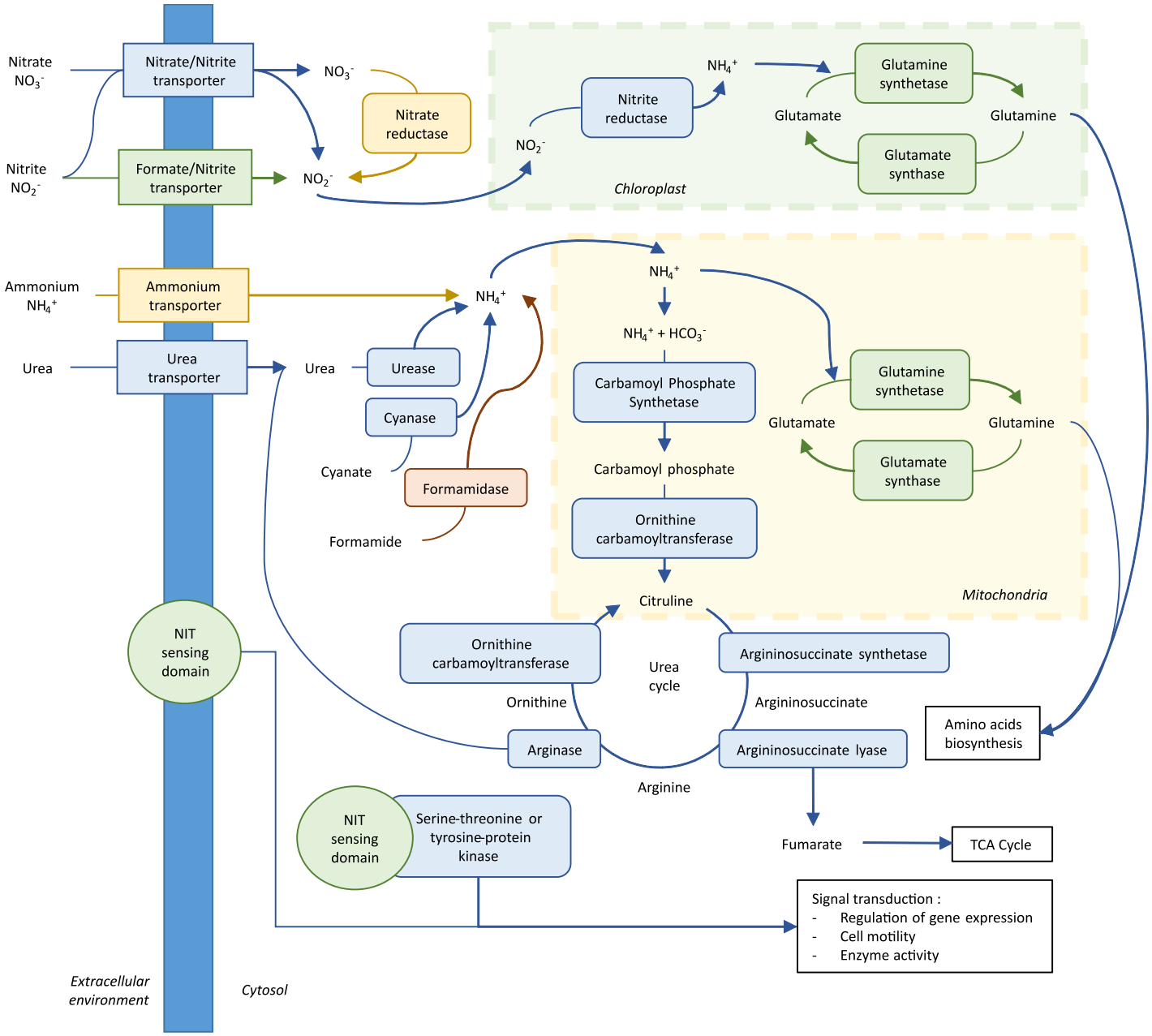

Figure 2: Schematic representation of N transport and assimilation in *P. calceolata* based on the gene content. Rectangles represent transporters of inorganic nitrogenous compounds, ovals represent enzymes and chemicals are unboxed text. The color code indicates if the number of gene copies for a specific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

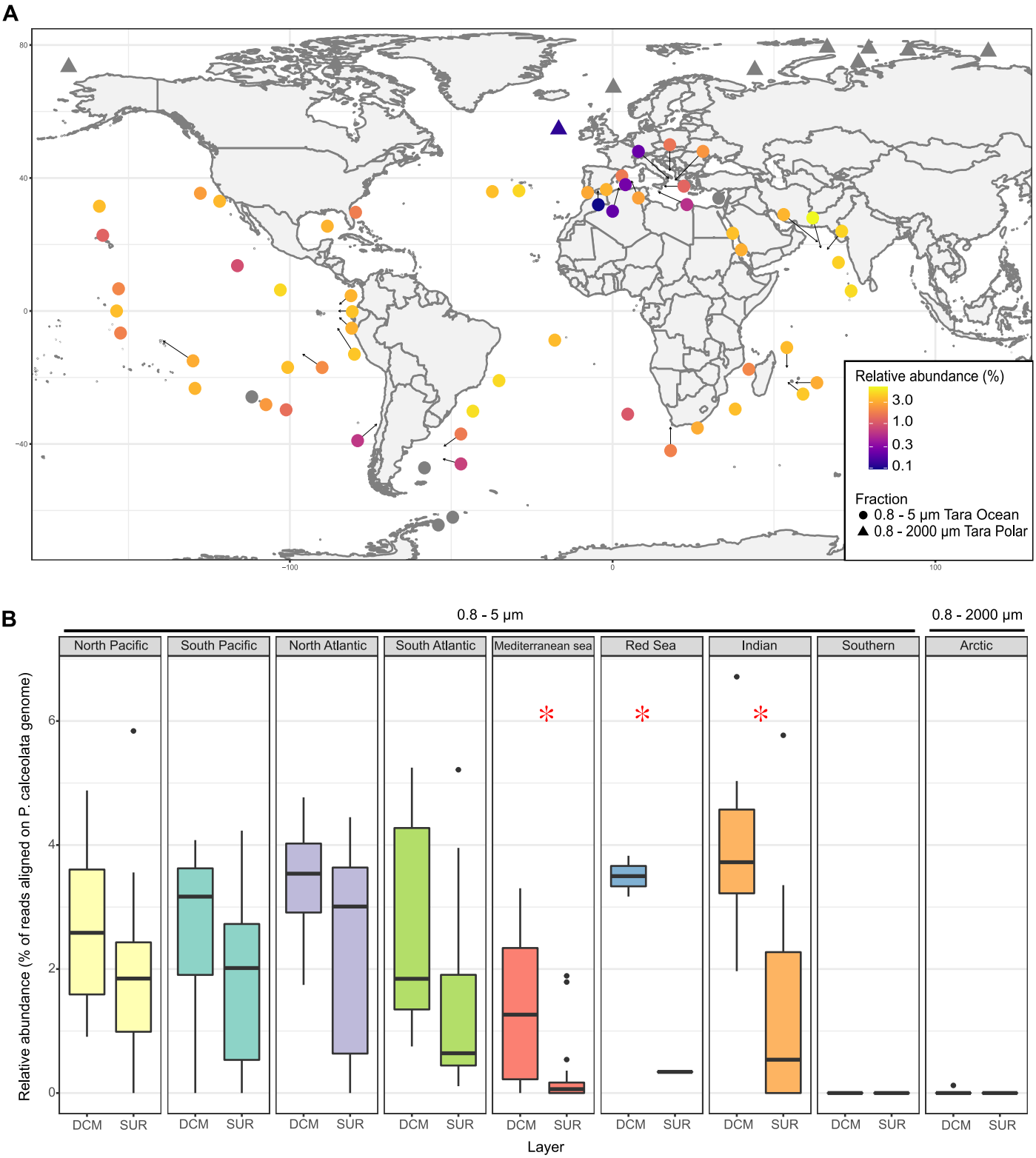

Figure 3: Relative abundance and distribution of *P. calceolata* in the oceans. A) World map of the relative abundance of *P. calceolata.* The colour code indicates the percentage of sequenced reads aligned on the *P. calceolata* genome. The DCM samples of size-fractions 0.8-5µm (circles) or 0.8-2000 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).

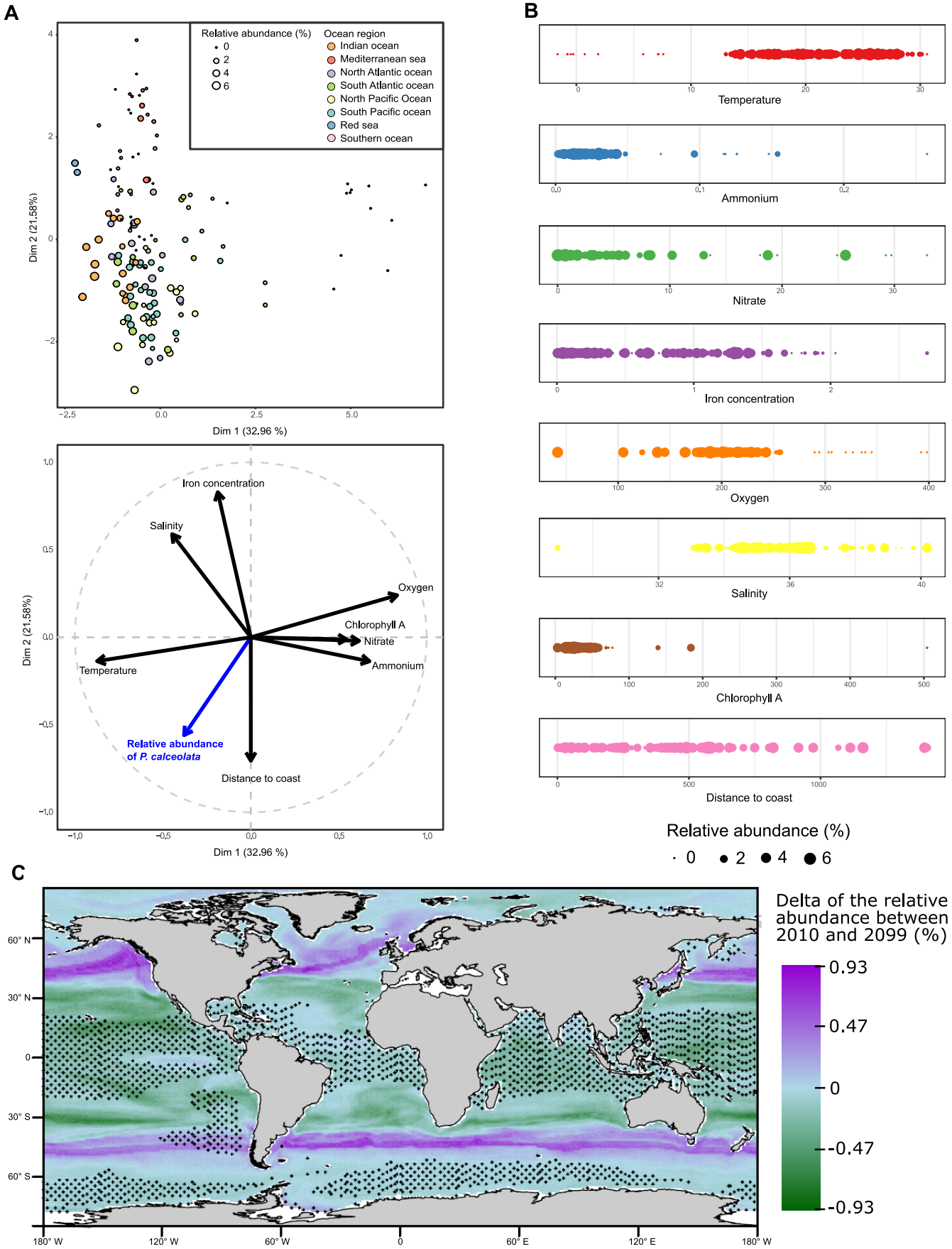

Figure 4: Ecological niche of *P. calceolata*. A) Principal component analysis of the relative abundance of *P. calceolata* and 8 environmental parameters measured during Tara sampling. Each dot represents a sample with a size proportional to the relative abundance of *P. calceolata*. The colours indicate the oceanic basins. B) Bubble plot of the relative abundance of P. calceolata for the 0.8-5 µm size-fraction according to the 8 environmental parameters. C) Difference of the modelled relative abundance of *P. calceolata* between 2010 and 2099. Green areas correspond to a decrease while purple areas correspond to an increase of *P. calceolata* relative abundance. Small stars indicate locations where at least one of the predictor drivers is out of range compared to the training dataset values.