



**HAL**  
open science

## **DiatOmicBase, a gene-centered platform to mine functional omics data across diatom genomes**

Emilie Villar, Nathanaël Zweig, Pierre Vincens, Helena Cruz de Carvalho, Carole Duchene, Shun Liu, Raphael Monteil, Richard G Dorrell, Michele Fabris, Klaas Vandepoele, et al.

### ► To cite this version:

Emilie Villar, Nathanaël Zweig, Pierre Vincens, Helena Cruz de Carvalho, Carole Duchene, et al.. DiatOmicBase, a gene-centered platform to mine functional omics data across diatom genomes. 2024. hal-04760696

**HAL Id: hal-04760696**

**<https://hal.science/hal-04760696v1>**

Preprint submitted on 30 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 **DiatOmicBase, a gene-centered platform to mine functional omics** 2 **data across diatom genomes**

3  
4  
5

6 Emilie Villar<sup>1,2</sup>, Nathanaël Zweig<sup>1</sup>, Pierre Vincens<sup>1</sup>, Helena Cruz de Carvalho<sup>1,3</sup>, Carole Duchene<sup>4,§</sup>,  
7 Shun Liu<sup>1,#</sup>, Raphael Monteil<sup>4</sup>, Richard G. Dorrell<sup>5</sup>, Michele Fabris<sup>6</sup>, Klaas Vandepoele<sup>7,8,9</sup>, Chris  
8 Bowler<sup>1,\*</sup> & Angela Falciatore<sup>4,5,\*</sup>

9  
10  
11

12 1 - Institut de Biologie de l'École Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS,  
13 INSERM, Université PSL, 75005 Paris, France

14 2 - EV Consulting, Marseille, France

15 3 - Faculté des Sciences et Technologie, Université Paris Est-Créteil (UPEC), 94000 Créteil, France

16 4 - Institut de Biologie Physico-Chimique, Laboratory of Chloroplast Biology and Light Sensing in  
17 Microalgae, UMR7141 Centre National de la Recherche Scientifique (CNRS), Sorbonne Université,  
18 75005 Paris, France

19 5- CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative - UMR 7238, Sorbonne  
20 Université, 4 place Jussieu, 75005 Paris, France

21 6 - SDU Biotechnology, Department of Green Technology, University of Southern Denmark,  
22 Campusvej 55, 5230 Odense M, Denmark

23 7 - Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 71,  
24 9052 Ghent, Belgium

25 8 - VIB-UGent Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

26 9 - VIB Center for AI & Computational Biology, VIB, Ghent, Belgium

27  
28

29 <sup>§</sup>current address: Department of Algal Development and Evolution, Max Planck Institute for Biology,  
30 72076 Tuebingen, Germany

31 <sup>#</sup>current address: Guangzhou Marine Geological Survey, Guangzhou, China

32  
33

\*Corresponding authors: [angela.falciatore@ibpc.fr](mailto:angela.falciatore@ibpc.fr), [cbowler@biologie.ens.fr](mailto:cbowler@biologie.ens.fr)

34  
35  
36  
37

38 **Abstract :**

39

40 Diatoms are prominent microalgae found in all aquatic environments. Over the last 20 years,  
41 thanks to the availability of genomic and genetic resources, diatom species such as  
42 *Phaeodactylum tricornutum* have emerged as valuable experimental model systems for  
43 exploring topics ranging from evolution to cell biology, (eco)physiology and biotechnology.  
44 Since the first genome sequencing in 2008, numerous genome-enabled datasets have been  
45 generated, based on RNA-Seq and proteomics, epigenomes, and ecotype variant analysis.  
46 Unfortunately, these resources, generated by various laboratories, are often in disparate  
47 formats and challenging to access and analyze. Here we present DiatOmicBase, a genome  
48 portal gathering comprehensive omics resources from *P. tricornutum* and two other diatoms  
49 to facilitate the exploration of dispersed public datasets and the design of new experiments  
50 based on the prior-art.

51 DiatOmicBase provides gene annotations, transcriptomic profiles and a genome browser with  
52 ecotype variants, histone and methylation marks, transposable elements, non-coding RNAs,  
53 and read densities from RNA-Seq experiments. We developed a semi-automatically updated  
54 transcriptomic module to explore both publicly available RNA-Seq experiments and users'  
55 private datasets. Using gene-level expression data, users can perform exploratory data  
56 analysis, differential expression, pathway analysis, biclustering, and co-expression network  
57 analysis. Users can create heatmaps to visualize precomputed comparisons for selected gene  
58 subsets. Automatic access to other bioinformatic resources and tools for diatom comparative  
59 and functional genomics is also provided. Focusing on the resources currently centralized for  
60 *P. tricornutum*, we showcase several examples of how DiatOmicBase strengthens molecular  
61 research on diatoms, making these organisms accessible to a broad research community.

62

63 **Significance statement :**

64

65 In recent years, diatoms have become the subject of increasing interest because of their  
66 ecological importance and their biotechnological potential for natural products such as  
67 pigments and polyunsaturated fatty acids. Here, we present an interactive web-based server  
68 that integrates public diatom 'omics data (genomics, transcriptomics, epigenomics,  
69 proteomics, sequence variants) to connect individual diatom genes to broader-scale functional  
70 processes.

71

72 **Keywords :**

73

74 Diatoms, *Phaeodactylum tricornutum*, genome portal, genome browser, gene models,  
75 ecotype variants, histone marks, non-coding RNAs, protein domains, RNA-Seq datasets

76

77 **Introduction:**

78 Diatoms are unicellular algae that play a major ecological role by contributing up to 20% of  
79 carbon fixation in aquatic ecosystems (Tréguer et al., 2021). The group encompasses up to  
80 100,000 species (Alverson et al., 2007; Malviya et al., 2016) with a large diversity of  
81 morphologies, sizes and life histories. They are classified into two morphogroups according  
82 to their shape: centric diatoms are radially symmetric while pennates are bilaterally  
83 symmetrical. Their main common characteristic is the silica cell wall, or “frustule”, that make  
84 diatoms key players in the oceanic silica biogeochemical cycle (Tréguer et al., 2021) in  
85 addition to the carbon cycle. Widely distributed in any aquatic or humid environments,  
86 diatoms are particularly abundant in nutrient-rich coastal ecosystems as well as at high  
87 latitudes (Malviya et al., 2016). The diverse habitats in which they occur reflect their extreme  
88 adaptive capacities: they can be planktonic and benthic, and can be found even as epiphytes  
89 in terrestrial forests, in soil or associated with sea ice (Vanormelingen et al., 2009, Singer et  
90 al., 2021).

91 Besides their ecological significance, diatoms have recently emerged as interesting novel  
92 experimental systems to explore still largely uncharacterized features of phytoplankton  
93 biology (Falciatore and Mock, 2022). Two species are widely used to decipher diatom  
94 biology and molecular functioning: *Thalassiosira pseudonana* (a centric diatom) and  
95 *Phaeodactylum tricornutum* (a pennate diatom). They both present significant advantages:  
96 they are easily cultivated in the laboratory with rapid growth rates, their genetic  
97 transformation is mastered and they have small genomes (32.1 and 27.4 Mb, respectively)  
98 encoding around 12,000 genes (Armbrust et al., 2004; Bowler et al., 2008). In recent years, a  
99 growing number of different genomic, genetic, physiological and metabolic information and  
100 resources have been generated for these two species, making them suitable models to study  
101 diatom gene functions and metabolic pathways (Brodrick et al., 2019; Falciatore et al.,  
102 2020; Poulsen and Kröger, 2023). In parallel, additional diatom species have also been  
103 proposed as models to understand specific eco-physiological adaptations (e.g., *Fragilariopsis*  
104 *cylindrus* for polar habitats (Mock et al., 2017), *Thalassiosira oceanica* for open-oceans  
105 (Lommer et al., 2012), and *Seminavis robusta* (Osuna-Cruz et al., 2020) for benthic

106 environments) or to address specific features of diatom life cycles (e.g., *Pseudo-nitzschia*  
107 *multistriata* for studying diatom sex (Ferrante et al., 2023)).

108 As of March 2024, 117 complete genomes of pennate and centric diatoms have been  
109 deposited in Genbank, and an ongoing project has announced the sequencing of 100 new  
110 diatom genomes (JGI initiative). The Marine Microbial Eukaryotic Transcriptome  
111 Sequencing Project (MMETSP) has additionally provided 92 different transcriptomes from  
112 diverse diatom species (Keeling et al., 2014) which has been recently completed with 6 other  
113 transcriptomes (Dorrell et al., 2021), providing a good representation of the most abundant  
114 diatom species in the ocean (Malviya et al., 2016). Finally, 54 single-cell and metagenome-  
115 assembled genomes (sMAGs) from diatoms have been assembled from meta-transcriptome/-  
116 genome data derived from *Tara Oceans* (Delmont et al., 2022), allowing complementary  
117 insights into the biology and ecology of uncultured and uncultivable species (Nef et al.,  
118 2022). Phylogenetic analyses of diatom genomes have revealed an extensive gene repertoire,  
119 which can be considered in a phylogenetic sense to constitute a patchwork coming from an  
120 ancient host, several endosymbionts acquired at different times, and bacterial horizontal gene  
121 transfers (Dorrell et al., 2021, Vancaester et al., 2020). Reflecting their complex evolutionary  
122 histories and phylogenetic distance (perhaps a billion years) from better-studied model  
123 eukaryotes within the animals, fungi and plants, the functional organization and evolutionary  
124 trajectories of diatom genomes are highly distinctive. These include families of novel  
125 transposable elements (Hermann et al., 2014), novel epigenomic marking of chromatin  
126 (Veluchamy et al., 2015; Zhao et al., 2021) and an apparent lack of structured centromeres  
127 (Bowler et al., 2008).

128

129 Combined functional genomic approaches in model species have been used to begin to  
130 decipher molecular actors regulating diatom physiology and distinct cellular and metabolic  
131 features. These include extensive energetic exchanges between plastids and mitochondria that  
132 augment CO<sub>2</sub> assimilation (Bailleul et al., 2015), a peculiar organisation of the photosystems  
133 in the plastid membranes (Flori et al., 2017), and novel Light Harvesting Complex (LHC)  
134 protein families (Bailleul et al., 2010; Buck et al., 2019). The central role of diatoms in  
135 marine biogeochemical cycles has further been explored through identification of molecular  
136 mechanisms of nutrient uptake and metabolism, e.g., for silica (Nemoto et al., 2020), carbon  
137 (Shen et al., 2017), iron (Gao et al., 2021), nitrogen (Rogato et al., 2015) and phosphorus  
138 (Dell'Aquila and Maier, 2020). Comparative genomics and molecular physiology studies  
139 have greatly contributed to predict the role of nearly half of the diatom gene repertoire

140 (Blaby-Haas and Merchant, 2019): as of July 2024, 6,781 genes from the nuclear genome  
141 (out of 12,357) have at least one annotation from Gene Ontology (GO), InterPro or Kyoto  
142 Encyclopedia of Genes and Genomes (KEGG) databases. Notwithstanding, the functional  
143 significance of many other diatom genes remains largely unexplored.

144 Among all studied diatom species, the genomic resources for *P. tricornutum* are the most  
145 advanced (Russo et al., 2023). The first assembly of the *P. tricornutum* nuclear genome  
146 constituted 27.4 Mb and 10,402 total gene models, whose annotation was assisted by the  
147 availability of an extensive collection of Expressed Sequence Tags (ESTs)  
148 (<https://mycocosm.jgi.doe.gov/Phatr2/Phatr2.home.html>) (Bowler et al., 2008). At the same  
149 time, a complete mitochondrial genome (77 kb, containing 60 genes (Oudot-Le Secq and  
150 Green, 2011)) and plastid genome (117 kb, containing 162 genes (Oudot-Le Secq et al.,  
151 2007)) were assembled. Following the initial assembly, the nuclear genome annotation was  
152 refined using 90 RNA-Seq datasets and more advanced annotation algorithms, resulting in  
153 the Phatr3 annotation. This annotation is available on the Ensembl archive  
154 ([http://protists.ensembl.org/Phaeodactylum\\_tricornutum/Info/Index](http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index); Rastogi et al., 2018).  
155 More recent annotations (e.g., using proteomic data (Yang et al., 2018) and a new telomere-  
156 to-telomere *P. tricornutum* genome assembly (using long read sequencing technology  
157 (Filloramo et al., 2021 and Giguere et al., 2022)) have been performed, but to date no new  
158 annotation has been proposed to the community. Concerning transcriptomics, 123  
159 microarrays have been used to generate a co-expression network (Ashworth et al., 2016),  
160 which is deposited on the DiatomPortal website (<http://networks.systemsbiology.net/diatom-portal>),  
161 while gene clustering of RNA-Seq experiments were recently used to create the  
162 PhaeoNet database (Ait-Mohamed et al., 2020). The epigenomic PhaeoEpiView browser  
163 including published epigenomic data has also been generated (Wu et al., 2023). Comparative  
164 genomics including *P. tricornutum* with a complete set of functional annotations are also  
165 provided in the PLAZA diatom comparative genomics platform (Vandepoele et al., 2013,  
166 Osuna-Cruz et al., 2020).

167

168 Despite the wealth of information generated from *P. tricornutum* summarized above, the  
169 resources are often disconnected, not always updated, or are incomplete, which diminishes  
170 their potential to yield valuable insights to the research community. To improve the  
171 connectivity between all the published omics data, we present here DiatOmicBase, a gene-  
172 centered web portal aiming to centralize all omics data related to diatoms. By focusing on the  
173 *P. tricornutum* model system, we provide different examples of the utility of this resource,

174 e.g., from the analysis of gene and protein characteristics, as well as gene expression analyses  
175 under different conditions using previously published datasets. We demonstrate how  
176 querying the DiatOmicBase resources can enable a deeper understanding of diatom gene  
177 products and their roles in specific physiological and cellular processes, and whole cell  
178 metabolic fluxes that may be exploited for biotechnological and synthetic biology  
179 applications (Kumar et al., 2022). In addition to exploring existing datasets, the portal also  
180 allows submission of a user's own data to provide a platform for common analyses to be  
181 performed.

182

## 183 **Results and Discussion**

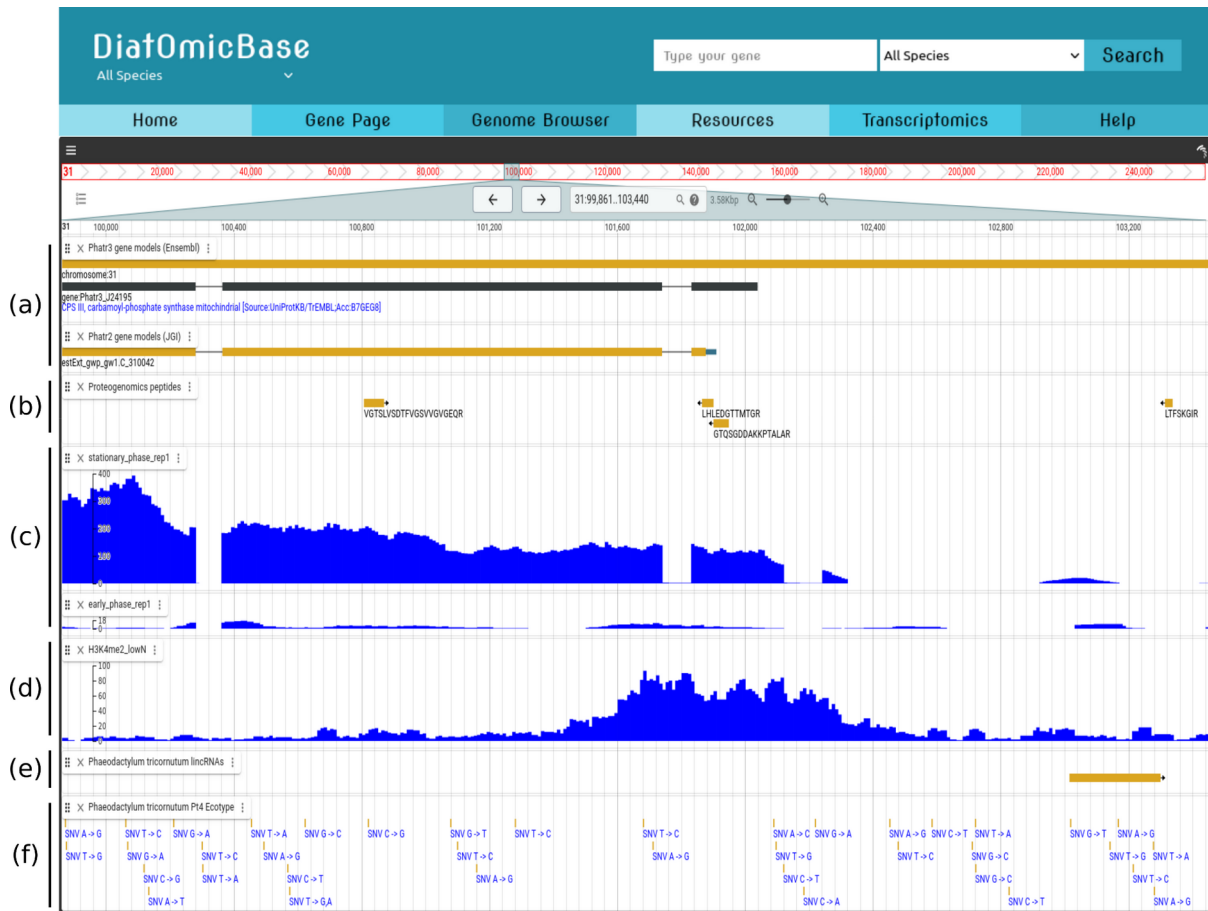
184

### 185 ***Gathering information to develop a complete database centered on *P. tricornutum* genes***

186

187 The DiatOmicBase website offers a gene-centered approach to facilitate exploration of the  
188 functional roles and regulation of diatom genes using omics-based resources. Considering *P.*  
189 *tricornutum* as a primary example, 12,392 gene pages corresponding to the latest annotation  
190 of protein coding genes (12,357 from the nuclear genome, 35 from the chloroplast, 60 genes  
191 from the mitochondria) contain sequence information gathered on a genome browser (Figure  
192 1), alongside general descriptions related to their functional annotation and evolutionary  
193 history (Figure 2).





194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

**Figure 1.** Snapshot of the DiatOmicBase genome browser illustrating the different features used for an integrative analysis of the CPS III gene (*Phatr3\_J24195*).

(a) *Phatr3* and *Phatr2* gene models can be compared at different zoom levels.

(b) Zooming in on the 5' region reveals two peptide sequences mapped using a proteogenomic pipeline on an exon predicted by the *Phatr3* gene model but not by *Phatr2*.

(c) Visualization of read mapping densities (here from Kwon et al., 2021) also validates the *Phatr3* prediction.

(d) Density plots of histone marks (H3K4 methylation) under nitrate-depleted conditions.

(e) Long non-coding RNAs (Cruz de Carvalho et al., 2016)

(f) Single nucleotide variants from 10 different ecotypes (Rastogi et al., 2020). This figure specifically shows only the Pt4 variant.

DiatOmicBase uses the 27.4 Mb **assembly** obtained in 2008 (Bowler et al., 2008) from the sequencing of *P. tricornutum* accession Pt1 8.6 (deposited as CCMP2561). As the latest published **gene prediction**, **Phatr3** (Rastogi et al., 2018) is defined as the standard in the website. This reannotation was obtained using existing gene models, expression data and



212 protein sequences from related species to train prediction programs and predicts 12,089 gene  
213 models. To preserve the continuity of research by conserving gene identifier correspondence,  
214 the previous gene annotation, **Phatr2** (Bowler et al., 2008), comprising 10,402 gene models,  
215 is also available considering that several genes have been manually annotated on this version  
216 by the diatom research community. Only 4,667 Phatr3 gene models display a perfect  
217 correspondence with Phatr2. The other Phatr3 gene models can be new (1,489), have a  
218 modified 5' and/or 3' (4,709), can be merged (194), split (262), antisense (346), or required  
219 manual curation (566).

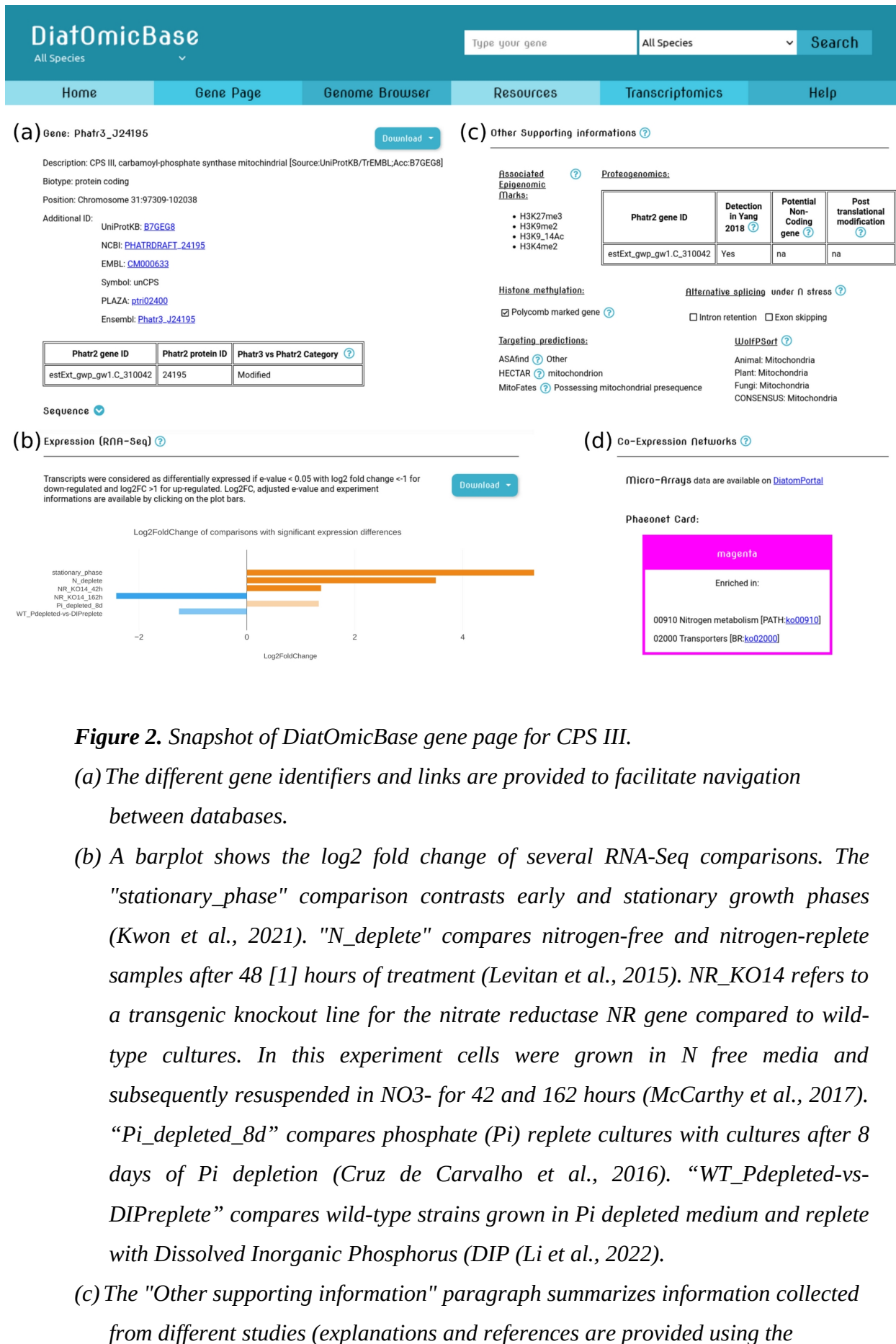
220

221 The genome browser also displays the annotation based on **proteomics** data. Yang et al.,  
222 (2018) analyzed the proteome of 45 samples from *P. tricornutum* grown under eight different  
223 conditions. The peptide sequences resulting from the protein digestion and mass spectrometry  
224 analysis are shown on the genome browser. Using a proteomic pipeline integrating a  
225 dedicated protein search database, Yang et al., (2018) confirmed 8,300 Phatr2 genes, and  
226 identified 606 novel proteins, 506 revised genes, and 94 splice variants that can all be found  
227 on the genome browser. Discrepancies in gene model prediction between Phatr2 and Phatr3  
228 can be resolved either using these proteomic data, or by analyzing mRNA expression and  
229 possible splicing variants using the multiple transcriptomics datasets centralized in the  
230 DiatOmicBase genome browser (see case study below, Figure 1 and Figure S1). To allow  
231 visualization of within-species gene diversity beyond the Pt1 8.6 strain, the database also  
232 includes genomic data from ten *P. tricornutum* ecotypes, covering broad geospatial and  
233 temporal scales (Martino et al., 2007), that have been re-sequenced (Rastogi et al., 2020).  
234 Variant calling results including SNPs, small insertions and deletions, can be visualized on  
235 the genome browser.

236

237 In addition to the sequence-related features displayed on the genome browser (Figure 1),  
238 several other descriptions are available on each gene page (Figure 2 and Figure S2). Gene  
239 pages are accessible by querying the database using Phatr2 or Phatr3 identifiers, terms or  
240 identifiers for GO, KEGG and Interpro databases. Keyword searches are available on the  
241 general search bar and on the gene page search. Finally, gene pages can be retrieved using  
242 nucleotide, protein and translated nucleotide (blastx) BLAST, with the possibility to tune  
243 various parameters.

244



245  
 246  
 247  
 248  
 249  
 250  
 251  
 252  
 253  
 254  
 255  
 256  
 257  
 258  
 259  
 260  
 261

**Figure 2.** Snapshot of DiatOmicBase gene page for CPS III.

(a) The different gene identifiers and links are provided to facilitate navigation between databases.

(b) A barplot shows the log2 fold change of several RNA-Seq comparisons. The "stationary\_phase" comparison contrasts early and stationary growth phases (Kwon et al., 2021). "N\_deplete" compares nitrogen-free and nitrogen-replete samples after 48 [1] hours of treatment (Levitani et al., 2015). NR\_KO14 refers to a transgenic knockout line for the nitrate reductase NR gene compared to wild-type cultures. In this experiment cells were grown in N free media and subsequently resuspended in NO<sub>3</sub><sup>-</sup> for 42 and 162 hours (McCarthy et al., 2017). "Pi\_depleted\_8d" compares phosphate (Pi) replete cultures with cultures after 8 days of Pi depletion (Cruz de Carvalho et al., 2016). "WT\_Pdepleted-vs-DIPreplete" compares wild-type strains grown in Pi depleted medium and replete with Dissolved Inorganic Phosphorus (DIP) (Li et al., 2022).

(c) The "Other supporting information" paragraph summarizes information collected from different studies (explanations and references are provided using the

262 *interactive "question mark"). Notably, HECTAR and Mitofates predict the*  
263 *expression of CPS III in the mitochondrion.*

264 *(d) To further integrate functional data, a link is provided to access the corresponding*  
265 *microarray-based co-expression network and the Phaeonet clusters (based on*  
266 *RNA-Seq data).*

267 *A snapshot of the complete page is available as Figure S2.*

268

269 The **gene annotations** include identifier correspondences with UniprotKB and NCBI,  
270 retrieved from their respective databases, and the former gene annotation Phatr2 provided in  
271 Rastogi et al., (2018). Annotations from the GO project (Ashburner et al., 2000) were  
272 retrieved from UniprotKB (The UniProt Consortium, release 2021\_01). GO terms for  
273 functional analyses were generated via automatic annotation, and cover three domains: the  
274 Subcellular Component where the gene products are localized, the Molecular Function  
275 informing the main activities of the gene product, and the Biological Process, the set of  
276 molecular events involving the gene product. Several domain and protein family predictions  
277 can be retrieved from UniprotKB, integrating 20 different databases (e.g., InterPro, Pfam,  
278 Gene3D, TIGRFAMs, CDD; the complete list is available [here](#)). These databases integrate  
279 different automated and/or manually curated protein signatures to ease the identification of  
280 protein functions. Functional orthologs, as previously reported in Aït-Mohamed et al., (2020)  
281 using the KEGG orthology database (Kanehisa, 2002), provide insights about molecular  
282 functions using hierarchically-structured biochemical pathways.

283

284 **Transposable Elements** (TEs) represent around 75% of the detected repetitive elements in  
285 the Phatr3 genome annotation (Rastogi et al., 2018; Giguere et al., 2021). With their ability to  
286 insert into genes or regulatory sequences, TEs act as key players in the organisation and  
287 expression of the genome, contributing to phenotypic diversity and, ultimately to the species  
288 evolution (Abbriano et al., 2023). A specific track has been created on the DiatOmicBase  
289 genome browser to visualize their positions by transforming the flat tables produced in  
290 Rastogi et al., (2018) into gff files. The majority (2,790, ~75%) of the TEs are associated  
291 with epigenetic marks (see below), which can also be important regulators of gene expression  
292 (Figure S2).

293

294 **Epigenetic modifications** play a pivotal role in regulating gene expression, influencing  
295 development, adaptation to environmental changes, and maintaining genome stability. In *P.*

296 *tricornutum*, whole genome methylation (Veluchamy et al., 2013) and the distribution of five  
297 histone marks (Veluchamy et al., 2015) have been described for the most used *P. tricornutum*  
298 strain, Pt1 8.6 (CCMP2561) grown in standard culture conditions. The methylome was  
299 obtained by digestion of three replicate DNA samples with the methyl-sensitive endonuclease  
300 McrBC followed by hybridization to a 2.1-million-probe McrBc-chip tiling array of the *P.*  
301 *tricornutum* genome (Veluchamy et al., 2013). The 3,950 methylated regions shown on the  
302 genome browser result from normalization on these three biological replicates. Five histone  
303 marks were chosen as they are known to be involved in transcriptional activation or  
304 repression: H3K4me2, H3K9me2, H3K9me3, H3K27me3 and H3AcK9/14. Two biological  
305 replicates of each histone modification were analyzed using ChIP-Seq, resulting in the  
306 discovery of 119,000 regions annotated with a set of chromatin states covering almost 40% of  
307 the genome. Following this first description of a diatom epigenomic landscape, changes have  
308 been examined in response to nitrate depletion, both by analyzing histone modifications  
309 (H3K4me2, H3K9/14Ac and H3K9me3 using Chip-seq) and DNA methylation (bisulfite  
310 deep sequencing) (Veluchamy et al., 2015). These marks are also displayed on the genome  
311 browser. DiatOmicBase also provides a direct link to the PhaeoEpiView epigenome browser  
312 which include epigenomic data (mostly DNA methylation and post-translational  
313 modifications of histones) on the newly assembled genome (Filloramo et al., 2021).

314

315 Different kinds of **small noncoding (sn)RNAs** (25 to 30 nt-long) have been described in *P.*  
316 *tricornutum* (Rogato et al., 2014) after sequencing of short RNA fragments isolated from  
317 cells grown under different conditions of light and nutrients. Their sequences have been  
318 mapped onto the genome browser. The majority of snRNAs map to repetitive and silenced  
319 TEs marked by DNA methylation and recent evidence indicates their role in the regulation of  
320 epigenetic processes (Grypioti et al., 2024). Other snRNAs target DNA-methylated protein-  
321 coding genes, or are derived from longer noncoding RNAs (tRNAs and U2 snRNA) or are of  
322 unknown origin. Long noncoding (lnc)RNA sequences have been shown to play a significant  
323 role in transcriptional mechanisms and post-translational modifications (Statello et al., 2021;  
324 Mattick, 2023). Although most of the well characterized lncRNAs stem from mammalian  
325 systems, recent work has shown that marine protists, including diatoms, all express lncRNAs  
326 (Debit et al., 2023). Among the different categories of lncRNAs, over 1,500 lincRNAs  
327 (intergenic lncRNAs) and ~3,200 lncNATs (antisense lncRNAs), that have previously been  
328 predicted from transcriptome mapping to the *P. tricornutum* genome (Cruz de Carvalho et al.,

329 2016; Cruz de Carvalho & Bowler, 2020), are likewise available on the genome browser, but  
330 are not integrated on the gene pages.

331

332 Links to a range of already existing web databases for **comparative genomics** are also  
333 provided. Information can be mined using PLAZA Diatoms (Osuna-Cruz et al., 2020) that  
334 includes structural and functional annotation of genome sequences derived from 26 different  
335 species, including 10 diatoms. Complementary annotations, homologous and orthologous  
336 gene families, synteny information, as well as a toolbox enabling a graphical exploration of  
337 orthologs and phylogenetic relationships are available on the corresponding gene pages of  
338 PLAZA. Alongside this, **evolutionary history annotations** based on a ranked BLAST top  
339 hit approach obtained from 75 combined libraries from different taxonomic groups across the  
340 prokaryotic and eukaryotic tree of life have been generated (Rastogi et al., 2018).

341

342 **Co-expression networks** gather genes with similar expression patterns across samples,  
343 suggesting that they could be related functionally, regulated in the same way, or belong to the  
344 same protein complex or pathway. In *P. tricornutum*, two different studies of co-expression  
345 networks have been published. The first one, published by Ashworth et al., (2016) explored  
346 the hierarchical clustering of 123 **microarray** datasets generated from studies of silica  
347 limitation, acclimation to high light, exposure to cadmium, acclimation to light and dark  
348 cycles, exposure to a panel of pollutants, darkness and re-illumination, and exposure to red,  
349 blue and green light. A link to the corresponding gene page on DiatomPortal, the web  
350 platform containing this data, is available. More recently, Ait-Mohamed et al., (2020)  
351 performed Weighted Gene Correlation Network Analysis (WGCNA) of 187 publicly  
352 available and normalised **RNA-Seq** datasets generated under varying nitrogen, iron and  
353 phosphate growth conditions (Cruz de Carvalho et al., 2016; McCarthy et al., 2017) to  
354 identify 28 merged modules of co-expressed genes. The gene cluster identifier (denoted  
355 **PhaeoNet card**) and its KEGG pathway enrichments are indicated for each gene within these  
356 modules in DiatOmicBase. PhaeoNet modules for each gene are detailed in the  
357 supplementary information file available on the resource page.

358

359 Finally, supplementary information regarding the functions of each protein are provided:  
360 **post-translational modifications** extracted from the work of Yang et al., (2018); alternative  
361 splicing from the work of Rastogi et al., (2018) and also by analyzing the RNA-Seq data  
362 centralized in DiatOmicBase; *in silico* **targeting predictions** regrouping multiple different

363 predictive tools: HECTAR under default conditions (Gschloessl et al., 2008); ASAFind using  
364 Signal P v 3.0 (Gruber et al., 2015; Dyrlov Bendtsen et al., 2004); MitoFates with cutoff  
365 threshold 0.35 (Fukasawa et al., 2015); and WolfPSort with animal, plant and fungal  
366 reference models (Horton et al., 2007), as previously assembled in Ait Mohamed et al.,  
367 (2020).

368

369 Users have the possibility to leave comments on each gene page, indicating a reference and  
370 one or more predefined labels to facilitate indexing (Figure S2). This tool is expected to help  
371 the community to centralize information that is not easily accessible or cannot be  
372 automatically retrieved, such as the availability of transgenic lines, or evidence for  
373 transcriptional regulation or alternative splicing. A peer-reviewed reference is mandatory and  
374 comments can be moderated by the DiatOmicBase committee. The comments can also be  
375 used to specify a correct gene model when different gene annotations (Phatr2 and Phatr3) are  
376 not consistent.

377

### 378 ***Transcriptomic Data and Analysis***

379 We collected raw data available at NCBI from the RNA-Seq short reads BioProjects  
380 published from *P. tricornutum* (different ecotypes and selected transgenic lines), exposed to  
381 different conditions. These kinds of studies cover nutrient limitation (nitrate, phosphate, iron,  
382 vitamin B12), responses to chemical exposure (decadenial, naphthenic acids, glufosinate-  
383 ammonium, L-methionine sulfoximine, rapamycin, and nocodazole), different CO<sub>2</sub> and light  
384 levels, response to grazing stress or competition. Samples involving transgenic lines (nitrate  
385 reductase and aureochrome photoreceptor knock-outs, alternative oxidase and cryptochrome  
386 knock-downs, chitin synthase transgenic cell lines, *etc.*) were compared to the corresponding  
387 wild-type samples. Morphotype-related transcriptomes were all compared together as well as  
388 growth stage transcriptomes. When studies consisted of time series, sample sets were  
389 compared in control *versus* treatment pairs for each timepoint but not between timepoints. As  
390 of July 2024, DiatOmicBase includes 33 studies, encompassing 1,398 samples and 135  
391 comparisons.

392

393 For each BioProject, the most relevant pairwise comparisons were chosen to assess gene  
394 expression regulation in the different sample sets using the R package DESeq2 (Love et al.,  
395 2014). Bar plots showing up- and down-regulation illustrate the results of the pre-computed



396 comparisons on each gene page. Only significant comparisons are graphically shown but the  
397 list of non-significant comparisons and samples without read matches are provided.

398

399 On the transcriptomics page, users can also reanalyze public Bioprojects, defining the  
400 samples to compare (see case study 2). For public data, the gene-level read counts are already  
401 computed for each sample, and users only have to select the samples to compare. Moreover,  
402 users can also analyze their own data (see case study 3). For private data, inputs are a gene-  
403 level read count table or any equivalent expression matrix and a table informing how the  
404 samples should be grouped to be compared.

405

406 Gene expression analyses can subsequently be performed using the web application  
407 “integrated Differential Expression and Pathway analysis” (iDEP; Ge et al., 2018).  
408 Connecting several widely used R/Bioconductor packages and gene annotation databases,  
409 iDEP provides a user-friendly platform for comprehensive transcriptomics analysis, including  
410 quality control plots, normalization, PCA, differential expression analysis, heatmaps,  
411 pathway and GO analysis, KEGG pathway diagrams, functional annotation, co-expression  
412 networks, interactive visualizations, and downloadable gene lists. iDEP enables pairwise  
413 comparisons or more complex statistical models including up to 6 factors. Co-expression  
414 networks can be examined using WGCNA. All the analysis steps are customisable; methods  
415 and parameters can be easily tuned using dialog boxes.

416 Finally, expression patterns of several genes can be analyzed by drawing customized  
417 heatmaps (Figure 3). In this case the user can provide a gene list and select the sample  
418 comparisons to be shown on the plot.

419

## 420 **Case studies**

421 Improving gene model prediction and visualization of gene expression and regulation

422

423 DiatOmicBase can aid functional analyses of diatom genes by improving prediction of  
424 protein-coding genes and their regulation. In Figures 1 and 2, we show the information that  
425 can be readily obtained in DiatOmicBase, using as an example the *P. tricornutum* *CPS III*  
426 gene, encoding a carbamoyl phosphate synthase involved in the urea cycle. In diatoms, the  
427 discovery of an ornithine-urea cycle (OUC) was unexpected as it was previously thought to  
428 be specific to animals, to allow removal of excess  $\text{NH}_4^+$  derived from a protein-rich diet.



429 Rather than eliminate NH<sub>4</sub>, diatoms have been proposed to conserve this precious resource by  
430 using the pathway to cope with fluctuations in nitrogen availability in the ocean (Allen et al.,  
431 2011, Smith et al., 2019), maintaining the cellular balance of carbon and nitrogen. The key  
432 enzyme of the OUC, carbamoyl phosphate synthase (CPS) is encoded in the *P. tricornutum*  
433 genome by two copies. One copy, named *unCPS* or *CPS III* (Phatr3\_J24195) and using  
434 ammonium as a substrate, is predicted both by HECTAR and MitoFates to be targeted to the  
435 mitochondria (Figure 2c), in agreement with the mitochondrial localization observed by  
436 microscopy (Allen et al., 2011). Furthermore, a *T. pseudonana* homologue of *unCPS*  
437 (Thaps3a\_40323) has been recovered from proteomic data derived from purified  
438 mitochondrial fractions. A second *P. tricornutum* paralog, pgCPS2 (encoded by  
439 Phatr3\_EG01947) has been inferred to be cytosolic.

440 In DiatOmicBase, the peptide alignment from proteogenomic studies confirmed the Phatr3  
441 gene model of *CPS III*, compared to the Phatr2 gene model (estExt\_gwp\_gw1.C\_310042;  
442 Figures 1, 2 and S1), predicting a protein extended by 133 amino acids at the N-terminus, due  
443 to the translation from an earlier ATG initiation site in the genome (chromosome 31 position  
444 102308 reverse strand, c.f. chromosome 31 position 101909 reverse strand). This gene model  
445 and also intron position were confirmed by analyzing RNA-Seq data from conditions where  
446 the *CPS III* gene is strongly expressed (Figure S1).

447

448 Considering quantitative gene expression trends, reanalysis of transcriptomic data in  
449 DiatOmicBase has shown that *CPS III* is highly expressed in cells experiencing three days of  
450 nitrogen deprivation (Levitan et al., 2015). Interestingly, its expression is also induced under  
451 phosphorus (P) depletion conditions, whereas Pi repletion of starved cells reverse this trend.  
452 These responses to Pi availability could be the result of the rapid cross-talk between Pi and N  
453 metabolism observed in diatoms (Helliwell et al., 2021). *CPS III* was also overexpressed in a  
454 nitrate reductase (NR) knockout line compared with the wild-type (wt) under conditions of  
455 nitrogen repletion (e.g., comparing the response of NR knockout versus wt in cells repleted  
456 with nitrogen for 42 h), but was down-regulated when cells became limited for this nutrient  
457 (e.g., NR KO versus wt after 162 h of nitrogen resupply) (McCarthy et al., 2017; Figure 2b).  
458 *CPS III* is also down-regulated in cell lines incubated in the complete absence of nitrogen  
459 compared with nitrogen-replete media for 4 and 20 hours (Matthijs et al., 2016), but was up-  
460 regulated when cells experienced prolonged nitrogen starvation (Levitan et al., 2015). It is  
461 therefore possible that nitrate uptake and internal cellular nitrate concentrations play a role in  
462 the hierarchical regulation of *CPS III* expression. We also note a dramatic increase in *CPS III*

463 expression in stationary-phase cells compared with exponential-phase cells, which may relate  
464 to functions in amino acid scavenging and recycling, but also in nutrient starvation responses.  
465 Information on *CPS III* gene expression in additional growth conditions and physiological  
466 states can be obtained by analyzing other centralized omics data (e.g., Figure S2).  
467 The histone marks H3K4me2 and H3K9/14Ac, consistent with transcriptional activity,  
468 mapped to the 5' region of the gene both in nitrate replete and deplete conditions, suggesting  
469 that CPS III is also important in nitrate replete cells (Figure 2c). We furthermore note that  
470 *CPS III* groups in the PhaeoNet magenta card, a module of co-expressed genes that appear to  
471 be enriched in functions implicated in organelle amino acid and nitrate metabolism (Figure  
472 2d). Other members of this module include the plastidial glutamate synthetase  
473 (Phatr3\_J50912) and N-acetyl-gamma-glutamyl-phosphate reductase implicated in the  
474 plastidial ornithine cycle (Phatr3\_J36913), as well as several other genes encoding proteins  
475 involved in nitrate assimilation. This might be consistent with a role of the urea cycle under  
476 nitrate replete conditions in recycling excess assimilated plastidial amines. Finally, while no  
477 TEs were found in the coding region, a lincRNA was detected in the 5' region of the gene. A  
478 dozen small RNAs were mapped onto the gene, and ecotypes displayed between 7 and 31  
479 single nucleotide variants (Figure S2), suggesting further possible hierarchical factors  
480 influencing the function and evolution of this gene.

481

482 Identification of common or distinct regulators of diatom responses to light and nutrient  
483 variations

484

485 Here we show how transcriptomic resources centralized in DiatOmicBase can be exploited to  
486 perform novel comparative functional analyses across multiple datasets (i.e., meta-analyses)  
487 and derive new information about common or specific regulators possibly implicated in  
488 diatom acclimation to environmental cues. Fluctuations in nutrient availability are recurrent  
489 in the marine environment, with nitrogen (N) and phosphorus (P) being amongst the main  
490 limiting nutrients for primary productivity. Both nitrogen and phosphorus deficiencies lead to  
491 a reduction in photosynthetic activity and a halt in cell division in diatoms (Jaubert et al.,  
492 2022; Levitan et al., 2015; Cruz de Carvalho et al., 2016; Matthjis et al., 2016). This may  
493 result in a cellular quiescent state, which is reversed when nutrients are resupplied (Cruz de  
494 Carvalho et al., 2016; Matthjis et al., 2016; Dell'Aquila and Maier, 2020). On the other hand,  
495 light is the major source of energy for photosynthesis, and a critical source of information

496 from the external environment. In recent years, the exploration of diatom genomes and of  
497 transcriptomic data obtained from cells exposed to different light conditions (wavelength,  
498 intensity and photoperiod cycles) and targeted functional analyses of selected genes in *P.*  
499 *tricornutum* has identified peculiar regulators of diatom photosynthesis (Lepetit et al., 2022),  
500 novel photoreceptors responsible for light perception, and an endogenous regulator  
501 controlling responses to periodic light-dark cycles (Jaubert et al., 2022). However, the  
502 regulatory cross-talk between light and nutrient signaling pathways remains poorly  
503 understood in diatoms. To this aim, here we reanalyzed already available RNA-Seq data from  
504 cells experiencing different light conditions (high light stress, fluctuating light, different  
505 colours) as well as different nitrate and phosphate depletion and repletion conditions.

506

507 In Figures 3, S3 and S4, we focused on gene expression changes for two-component  
508 regulators and transcription factors (TFs) encoded in the *P. tricornutum* genome. These  
509 regulators orchestrate the physiological response(s) to a given stress by participating in  
510 signaling cascades through differential expression regulation. The genome of *P. tricornutum*  
511 contains a high number of bacterial-like two-component sensory histidine kinases (Bowler et  
512 al., 2008), defined on the basis of the presence of a histidine kinase domain (PF00512) and/or  
513 a response regulator domain (PF00072). In addition, the histidine phosphotransmitter protein  
514 (Hpt), which was reported missing in the first version of the genome (Bowler et al., 2008),  
515 was later identified in Phatr3 (Phatr3\_J33969, PF01627) and was included in the analysis  
516 (Figure 3). Regarding two-component regulators, we observed significant gene expression  
517 changes for some of these regulators following a red to blue light shift (e.g., *DHK1*, *DHK2*,  
518 *DDR3*, *EGO2384*, *EGO2387*). The loss of these responses in two independent Aureochrome  
519 blue light receptor KO lines (Mann et al., 2020) compared to wild-type cells indicates that  
520 this blue light photoreceptor participates in the regulation of their expression. By contrast,  
521 *Phatr3\_J46628* is not affected by any changes in light conditions, but is over-expressed under  
522 prolonged phosphate depletion (4d) (Figure 3). Interestingly, *DDR3* is found to be up-  
523 regulated under blue-light treatment, but also under both N (from 4h to 48h) and Pi depletion  
524 (4 days) as well as under Pi resupply (Figure 3 and S3), but not under HL. It therefore  
525 appears likely that a blue light-activated signaling cascade participates via *DDR3* in the  
526 regulation of cellular responses to nutrient availability, but not in photoprotection.  
527 *Phatr3\_EGO2384* is expressed in the opposite fashion in response to HL and nitrogen  
528 depletion, suggesting a possible antagonistic role in light and nutrient signaling pathways. On

529 the other hand, *DRR1* appears not to be affected by light signals, but is specifically expressed  
530 under prolonged N deficiency and Pi depletion.

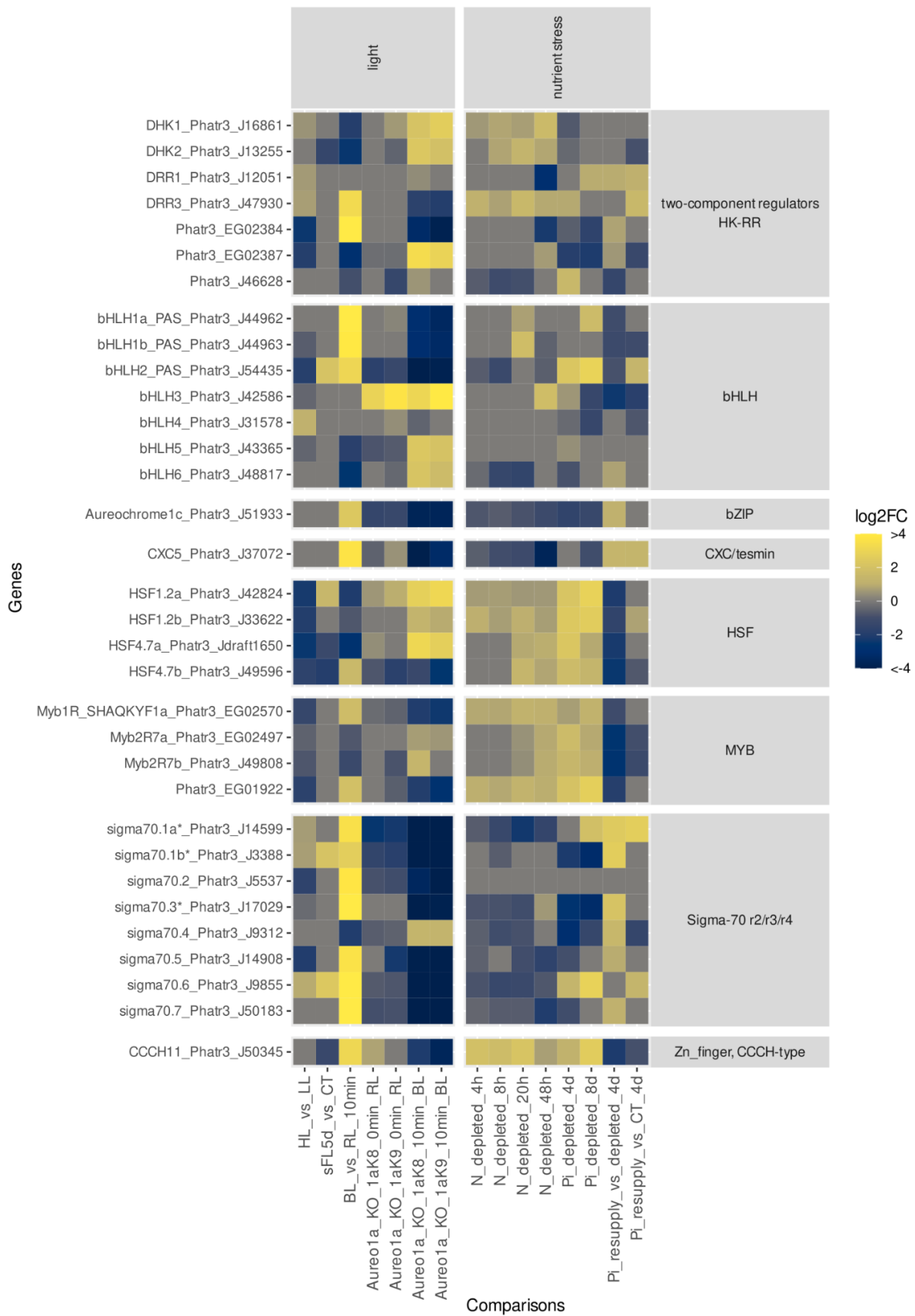
531

532 The analysis of TF expression patterns across treatments also provides interesting  
533 information (Figures 3 and S4). Between the genes encoding bHLH domain-containing  
534 proteins, the bHLH1a-PAS also known as RITMO1 shows a rapid induction by blue light  
535 treatment, which is repressed in *Aureo1* photoreceptor mutants, as previously shown (Mann  
536 et al., 2020). Neither HL treatments, nor short N and Pi deficiency affect its expression, as  
537 expected considering the role of this protein as an endogenous and robust timekeeper  
538 (Annunziata et al., 2019). Its expression and possible activity is, however, affected by  
539 prolonged nutrient stress treatments. bHLH2-PAS is differentially expressed under the tested  
540 light conditions as well as Pi deficiency, while bHLH3 appears to be expressed specifically  
541 under N and Pi deficiency. Interestingly, its expression is not induced by blue light, but is  
542 significantly affected in the *Aureo1a* mutants under red light. It is thus possible that *Aureo1a*,  
543 which is also a TF with a bZIP domain, can participate in regulation to nutrient changes,  
544 independently of its activity as a blue light sensor. A light-independent activity of *Aureo1a* is  
545 also supported by a recent study reporting altered rhythmic gene expression in *Aureo1*  
546 mutants compared to wild-type cells in constant darkness (Madhuri et al., 2024). Of the genes  
547 encoding sigma 70 factors, it is interesting to note that their expression is strongly modulated  
548 by HL or blue light, and for the three factors predicted to be plastid localized (sigma 70.1a,  
549 1.b and 70.3\*) also by Pi availability. This is consistent with a possible role for these factors  
550 in regulating plastid gene expression, which could also be sensitive to the physiological state  
551 of this organelle. Other TFs modulated by nutrient availability and blue light include  
552 Aureochrome1C of the bZIP family and Pt-CXC5 of the CXC/tesmin family, both of which  
553 are downregulated under nutrient stress and upregulated upon nutrient resupply, namely P, as  
554 well as under blue light (Figure 3). On the other hand, the opposite trend can be found for  
555 TFs of the HSF (HSF1.2a/b; HSF4.7a/b) and Myb families, which are also modulated by  
556 nutrient status and light, being up-regulated under nutrient and high light stress (Myb2R7a/b),  
557 as well as blue light (Myb1R\_SHAQKYF1a; Phatr3\_EG01922).

558

559 This analysis reveals the complex cross-talk between the regulatory pathways operating  
560 under light and nutrient fluctuations, which could enable opening new exploratory and  
561 functional research avenues aiming to shed light into the molecular processes underpinning  
562 diatom resilience to environmental stresses.

563



564

565

566 **Figure 3.** RNA-Seq heatmaps of two components regulators and transcription factors  
567 Heatmap showing gene expression differences in light and nutrient stress experiments for  
568 different genes of interest. HL\_vs\_LL compares low light and high light acclimated wild-type  
569 (WT) strains (Agarwal et al., 2022). sFL5d\_vs\_CT compares strains acclimated to constant  
570 light conditions to severe light fluctuation conditions (Zhou et al., 2022). BL\_vs\_RL\_10min  
571 compares WT strains under red light (RL) and 10 minutes after a shift to blue light (BL); In  
572 samples KO\_1aK8\_10min\_BL and KO\_1aK9\_10min\_BL, *Aureo1a* knockout K8 and K9  
573 strains are compared to WT under RL, and after a 10 minutes shift to BL (Mann, et al.,  
574 2020).  
575 *N\_deplete\_* compares nitrate-deplete and control samples 4h, 8h and 20h after transfer to  
576 nitrogen-replete medium (Matthijs et al., 2016). *N\_deplete\_48h* compares nitrogen-free and  
577 nitrogen-replete samples after 48h of treatment (Levitan et al., 2015). *Pi\_depleted\_4d* and  
578 *Pi\_depleted\_8d* compare control samples and cultures without phosphate supplement during  
579 4 and 8 days. *Pi\_resupply-vs-depleted\_4d* and *Pi\_resupply-vs-CT\_4d* compare samples with  
580 phosphate re-supplementation after 4 days of starvation (Cruz de Carvalho et al., 2016).  
581 \* based on the prediction supported by the database, these genes are localized in the plastid.  
582

583 DiatomicBase for the *de novo* analysis of previously unpublished RNA-Seq data

584

585 In DiatomicBase, we have also developed the possibility for users to analyze their own RNA-  
586 Seq data. Here, we show the analyses of new data obtained from *P. tricornutum* lines over a  
587 progressive two-week iron (Fe) starvation time-course. More specifically, we compared gene  
588 expression changes between cell lines grown in Fe replete media and those experiencing  
589 short-, medium- and long-term Fe withdrawal conditions (Figure S5).

590

591 Principal Component Analysis of the RNA-Seq data, performed with the integrated iDEP  
592 platform in DiatOmicBase, revealed three groups, corresponding to Fe-replete, short-term (3  
593 days), and medium/ long-term (7, 14 days) Fe limitation (Figure 4a). Calculation of the k-  
594 means revealed four distinct clusters enriched in different GO terms that were differentially  
595 regulated in response to different periods of Fe limitation (Figure 4b). The first cluster  
596 (generated as Cluster A by the iDEP calculations) was strongly induced after 3 days Fe  
597 withdrawal relative to Fe-replete conditions and was significantly ( $P < 10^{-15}$ ) enriched in genes  
598 encoding proteins with functions relating to photosynthesis (photosynthesis light reactions,



599 protein-chromophore linkage, generation of precursor metabolites and energy), alongside  
600 translation and peptide biosynthesis-associated processes (Figure 4c). This might relate to a  
601 transcriptional and translational upregulation of light-harvesting protein complexes in  
602 particular to compensate for diminished Fe availability limiting the synthesis of photosystem  
603 I (Gao et al., 2021). This contrasted with a second cluster (Cluster C) that was uniquely  
604 downregulated in short-term Fe limitation conditions, and showed a weaker ( $P < 0.001$ )  
605 enrichment in photosynthesis and transcriptional regulation processes but no other enriched  
606 GO terms (Figure 4b). The final two clusters, Clusters B and D, were strongly up- and  
607 downregulated, respectively, by medium and long-term Fe limitation compared to short-term  
608 and Fe-replete conditions. These showed weak enrichments ( $P < 0.0001$ ) in transport  
609 processes, i.e., ion and organic metabolite transport (Figure 4b). These may relate to the  
610 induction of Fe uptake systems in response to prolonged Fe withdrawal, alongside changes in  
611 the internal metabolite profiles and organic acid transport activities of Fe-limited cell lines  
612 (Gao et al., 2021).

613

614 While the data from DiatOmicBase provide insights into transcriptomic changes, they can be  
615 used to inform user construction of more complex evaluations of the links between gene  
616 expression and function. As an example of this, in Figure 4c we present Volcano Plots for  
617 DEGs generated from the DiatomicBase site, alongside tabulated fold-change and P-values  
618 calculated for genes known to be involved in Fe stress metabolism (Figure 4d; Gao et al.,  
619 2021). Concerning known Fe-stress associated proteins, many were already strongly  
620 upregulated after three days Fe limitation compared to Fe-replete lines, confirming the  
621 immediate impacts of Fe withdrawal on cell physiology (Figure 4c) (Gao et al., 2021). The  
622 most significantly upregulated of these genes at all three time points ( $P\text{-value} < 10^{-100}$  for 3, 7  
623 and 14-day Fe limitation) encoded a plastidial cytochrome  $c_6$  (Phatr3\_J44056), which  
624 mediates photosynthetic electron transfer between cytochrome  $b_6f$  and photosystem I,  
625 underlining the importance of photosystem remodelling in both short-term and sustained Fe  
626 limitation (Allen, de Paula et al., 2011). We note that *P. tricornutum* does not possess an  
627 endogenous plastocyanin gene (Groussman, Parker et al., 2015) that could be induced to  
628 compensate for cytochrome  $c_6$  under Fe-limitation.

629

630 Consistent with previous studies, we see the upregulation of known iron-stress related genes  
631 at all Fe-limitation time points studied. Constitutively upregulated genes include *ISIP2a/*  
632 *phytotransferrin* (Phatr3\_54465/ Phatr3\_54987) and *ferric reductase 2* (Phatr3\_J54982),



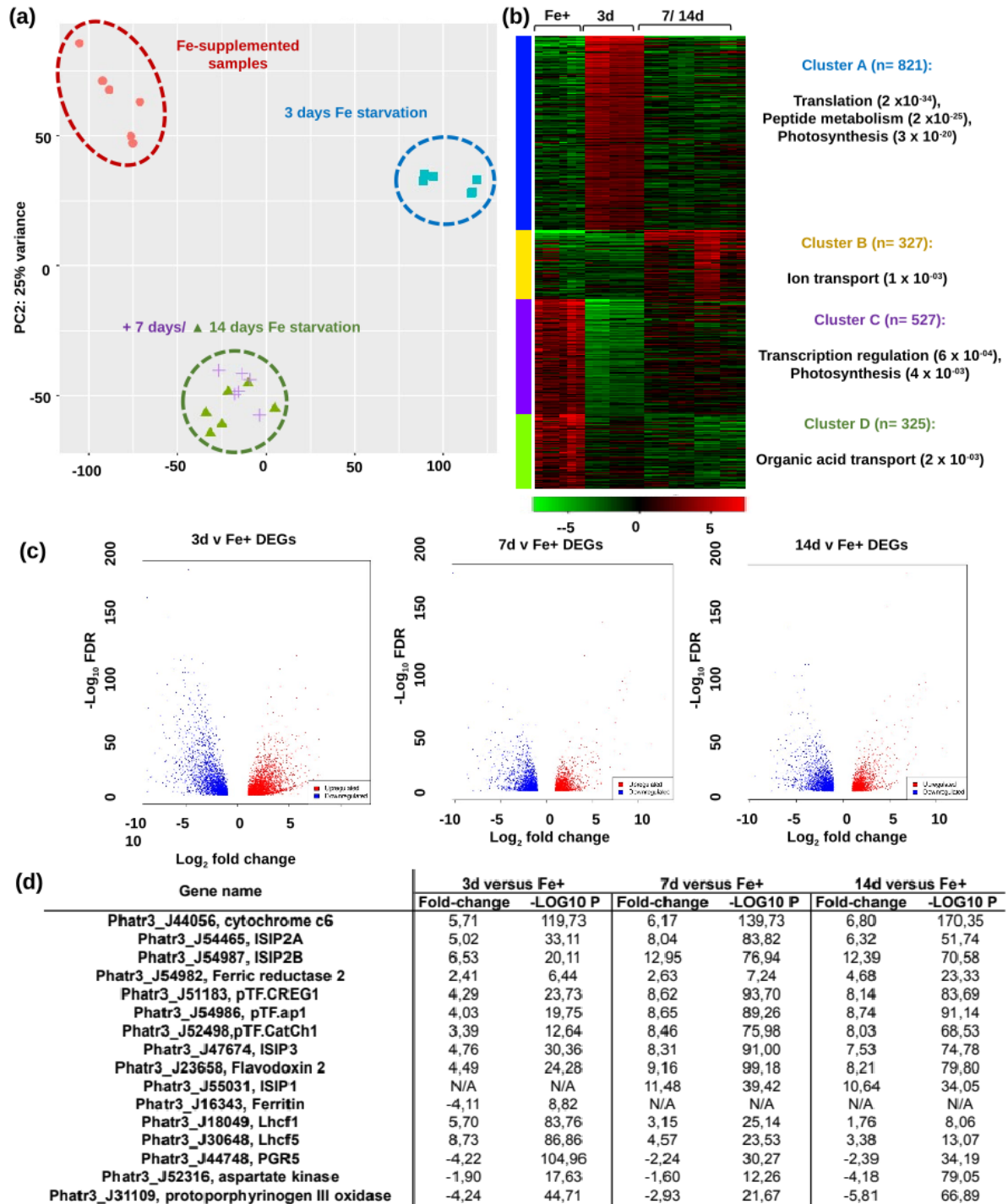
633 implicated in the reductive uptake of Fe from the cell surface (Morrissey, Sutak et al., 2015,  
634 McQuaid, Kustka et al., 2018); alongside genes encoding the newly identified plastid-  
635 targeted partners of ISIP2a *pTF.CREG1* (*Phatr3\_J51183*), *pTF.ap1* (*Phatr3\_J54986*), and  
636 the gene encoding the plasma membrane-located ISIP2a-binding protein pTF.CatCh1  
637 (*Phatr3\_J52498*) (Allen, Laroche et al., 2008, Turnšek, Brunson et al., 2021). These results  
638 suggest that reductive Fe uptake in diatom cells is relevant to both short- and long-term Fe  
639 limitation. We also observed upregulation at all time points for *ISIP3* (*Phatr3\_J47674*),  
640 encoding a protein of unknown function but proposed to participate in the intracellular  
641 trafficking or storage of iron (Behnke and LaRoche 2020, Kazamia, Mach et al., 2022); and  
642 flavodoxin 2 (*Phatr3\_J23658*), which can diminish cellular iron requirements of flavodoxin  
643 in photosystem I (Setif 2001, Lodeyro, Ceccoli et al., 2012).

644

645 Our data also provide insight into Fe-stress associated diatom genes that show more  
646 distinctive responses to Fe limitation. For example, ISIP1 (*Phatr3\_J55031*), implicated in the  
647 non-reductive transport of Fe-siderophore complexes from the cell surface to the chloroplast,  
648 shows no evidence of induction in the three day treatment, but strong induction in medium-  
649 and long-term limitation datasets (Figure 4c) (Kazamia et al., 2018), suggesting that it is  
650 induced more slowly than reductive iron uptake strategies. This might reflect the greater  
651 energetic cost or gene coordination required by diatoms to produce siderophores in response  
652 to Fe limitation, whose synthesis pathway remains unknown. Most dramatically, the gene  
653 encoding the proposed plastid iron storage protein ferritin (*Phatr3\_J16343*) showed no  
654 response to either medium or long-term Fe-limitation but was significantly downregulated ( $P$   
655  $< 10^{-08}$ ) in response to short-term Fe withdrawal. The role of ferritin in diatoms has  
656 historically been unclear, with some studies suggesting that it facilitates long-term Fe storage  
657 and tolerance of chronic starvation (Marchetti et al., 2009); and others that it may be  
658 transiently upregulated in response to Fe enrichment, allowing competitive removal of iron  
659 from the environment (Lampe et al., 2018, Cohen et al., 2018). Our data are broadly more  
660 consistent with the latter role for the *P. tricornutum* ferritin, although we note it is  
661 phylogenetically distinct to other diatom ferritins and may confer different physiological  
662 roles (Gao et al 2021). A greater similarity of this *P. tricornutum* ferritin gene with sequences  
663 from *Nannochloropsis* and ciliates than from other diatoms is also observed by analyzing  
664 PLAZA Diatoms.

665 Finally, our data provide some insights into the broader physiological responses that mediate  
666 short- and long-term diatom responses to Fe deprivation. The importance of light-harvesting

667 complexes for short-term responses to Fe-limitation (Figure 4c) is underlined by the strong  
668 induction of genes encoding two Lhc proteins (Lhcf1, *Phatr3\_18049*; Lhcf5, *Phatr3\_J30648*)  
669 that directly interact with one another in the PSI LHC (Joshi-Deo et al., 2010), although are  
670 also both found in PSII (Gundermann et al., 2013; Nagao et al., 2021), and are implicated in  
671 low-light adaptation (Gundermann et al., 2013). In contrast, we observed strong  
672 downregulation of PGR5 (*Phatr3\_J44748*), debatably implicated in cyclic electron flow  
673 around PSI in diatoms (Grouneva et al., 2011, Johnson et al., 2014). It has recently been  
674 proposed that *Chlamydomonas reinhardtii* and plant PGR5 indirectly participate in the  
675 delivery of Fe to PSI, which might explain its immediate sensitivity to Fe depletion in our  
676 data (Leister et al., 2022). In contrast, under long-term Fe-limitation but not under short- or  
677 medium-term conditions, we identified dramatic ( $P < 10^{-50}$ ) downregulation of the gene  
678 encoding Aspartokinase (*Phatr3\_J52316*), involved in lysine biosynthesis, as well as the  
679 gene encoding Protoporphyrinogen oxidase (*Phatr3\_J31109*), involved in the tetrapyrrole  
680 branch of chlorophyll/haem biosynthesis, which may point to an overall quiescence of core  
681 organelle metabolic pathways in response to sustained Fe deprivation (Allen et al., 2008; Ait-  
682 Mohamed, Novák Vanclová et al., 2020).



683

684 **Figure 4. de novo RNA-Seq analysis of iron starvation response.**

685 Comparison of gene expression from cells adapted to 12h light:12h night cycles under 19°C,

686 and either Fe-replete (“Fe+”), 3 day (“FeS”), 7 day (“FeM”) or 14 day (“FeL”) Fe-

687 limitation. A schematic experimental design is provided in Figure S5.

688

689 (a) PCA of RNA-Seq Transcripts Per Million (TPM) values calculated with the  
690 integrated iDEP module in DiatOmicBase. Three distinct treatment groups are visible,  
691 suggesting relative equivalence of FeM and FeL treatments.

692 (b) k-means clusters, showing selected enriched GO terms ( $P < 0.001$ ), calculated  
693 with the iDEP module in DiatOmicBase. Four clusters with distinctive biological identities  
694 show different relationships to both short-term and medium/ long-term Fe-limitation.

695 (c) Volcano plots of DEGs inferred with iDEP with threshold P-value 0.1 and fold-  
696 change 2.

697 (d) Tabulated fold-changes and  $-\log_{10}$  P-values of genes associated with Fe stress  
698 metabolism, following Gao et al., (2021).

699

700

## 701 **Perspectives**

702

703 DiatOmicBase provides a comprehensive database and analytic modules integrating genomic,  
704 epigenomic, transcriptomic and proteomic data from published diatom genomes. In this work,  
705 we focused on the resources centered around *P. tricornutum*. The first objective of this  
706 centralized database is to provide the community with a powerful tool for assessing correct  
707 gene models. Automated gene annotation remains error-prone, especially considering the  
708 large proportion of diatom genes whose function is unknown and which have no clear  
709 homologues in other systems. The peptide and RNA/DNA sequence mapping provided in  
710 DiatOmicBase may help to identify the correct form, as shown for the *CPS III* gene;  
711 DiatOmicBase makes provisions for user annotation and correction of individual gene pages.  
712 Given that *P. tricornutum* is the diatom species with the largest collection of genomic  
713 resources and data, the correct gene annotation in this species also represents a powerful  
714 support for correct gene prediction in other diatom species and from environmental data.

715

716 Analyzing the expression of genes in different circumstances has been facilitated by grouping  
717 results of several RNA-Seq experiments involving different conditions. In the case studies  
718 described in this work, we have shown how our tools can help to identify common or specific  
719 regulators of responses to various light changes and nutrient stress conditions. This offers  
720 new opportunities to characterize the still largely unknown signalling pathways involved in  
721 the perception and acclimation to complex changing environments. Using the iDEP pipeline,

722 public RNA-Seq datasets can be re-analyzed, enabling to answer questions different from the  
723 ones in the original articles. In order to enrich our resources, we would like to encourage the  
724 community to inform us when new data becomes available.

725

726 DiatOmicBase also facilitates the study of gene functions, essential for deciphering the  
727 physiology and metabolic capacities of these microalgae and harnessing their potential for  
728 biotechnology and synthetic biology Assigning a phenotype to a protein requires overcoming  
729 many challenges (Heydarizadeh et al., 2014) as automated annotations should be validated by  
730 biochemical evidence and comparative analyses between wild-type and mutant lines.  
731 Metabolic engineering and synthetic biology resources in diatoms are rapidly evolving  
732 (Russo et al., 2023), yet they are limited by some aspects of diatom metabolism that are not  
733 yet well understood. For example, we still do not know how many metabolic pathways,  
734 including those that are important for biotechnology such as the biosynthesis of carotenoids  
735 and terpenoids, react to environmental conditions. Nor do we know the mechanisms by which  
736 these pathways are regulated, or the subcellular location of the different enzymes. Protein  
737 targeting predictions, expression data available or newly generated and analyzed in  
738 DiatOmicBase, in conjunction with metabolomics data can be exploited for strain engineering  
739 strategies, by improving gene-enzyme-function associations, identifying unknown enzymes,  
740 and reveal still elusive aspects of pathway regulation. For designing pathway engineering and  
741 optimizing cultivation strategies, DiatOmicBase can be particularly useful for investigating  
742 co-regulation of industrially-relevant metabolic pathways or key pathway nodes with  
743 transcription factors (Figure 3) across various conditions to identify regulators that can  
744 control entire metabolic pathways. These are primary genetic targets for strain engineering  
745 strategies which are promising and feasible (Song et al., 2023), but also remains a largely  
746 untapped strategy in diatom biotechnology.

747

748 The future development of the DiatOmicBase should include improved genome assemblies  
749 and annotations. Unfortunately, *P. tricornutum* genome complexity (especially TEs) has  
750 prevented significant improvement of the reference genome obtained by Sanger sequencing.  
751 A new assembly using the latest technologies of long read sequencing has recently been  
752 published (Filloramo et al., 2021). However, this long-read derived assembly still lacks  
753 continuity resulting in more than 200 scaffolds (while the Sanger assembly comprised 88  
754 scaffolds with 33 chromosome-level scaffolds). As no gene annotation was associated with  
755 this assembly, it was not possible to fully exploit this new resource in this first version of

756 DiatOmicBase, but we aim to include it following it's annotation. As shown for the *CPS III*  
757 gene, DiatOmicBase should help to resolve conflicting gene model predictions derived by  
758 new sequencing and assembly projects. The database makes provisions for user annotation  
759 and correction of individual gene pages.

760

761 The addition of new diatom species in DiatOmicBase is also planned, especially those with  
762 the most developed omic resources. As of September 2024, preliminary DiatOmicBase  
763 portals are available for the model centric species *T. pseudonana*  
764 ([https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?](https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?species=Thalassiosira+pseudonana)  
765 [species=Thalassiosira+pseudonana](https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?species=Thalassiosira+pseudonana)) and the model sexual diatom *P. multistriata*  
766 ([https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?species=Pseudo-](https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?species=Pseudo-nitzschia+multistriata)  
767 [nitzschia+multistriata](https://www.diatomicsbase.bio.ens.psl.eu/genomeBrowser?species=Pseudo-nitzschia+multistriata)). Users can automatically search for gene information in these different  
768 species. Further developments will include the addition of new proteomic data, the  
769 incorporation of comparative genomic and automated phylogenetic analyses of individual  
770 genes, functional genomic information from transgenic lines, and precomputed  
771 biogeographical distributions of environmental homologues of individual diatom genes from  
772 *Tara* Oceans (Vernette et al., 2022).

773

## 774 **Experimental procedures**

775

### 776 **Website architecture**

777 The back-end server of the website consists of an API server coded with the FastAPI  
778 framework. It includes a local PostgreSQL database that is accessed using the SQLAlchemy  
779 library. Data from NCBI/EBI/Ensembl are fetched and inserted or updated in the local  
780 database using a python loader script. The genome browser used is Jbrowse 2 (Buels et al.,  
781 2016). A R-shiny iDEP instance (Ge et al., 2018), hosted on the local server, was modified to  
782 contain *P. tricornutum* genome annotation and to automatically load public data from  
783 DiatOmicBase. A background worker coded in python is used to run more computationally  
784 intensive user calculations (Blast or differential expression analysis) and ensures that these do  
785 not use more resources than allowed, using a queuing system. The possibility to comment  
786 gene pages used Commento widgets. To ensure data privacy, Commento and its PostgreSQL  
787 database are self-hosted.

788



789 The front-end part is developed with the React framework and uses static rendering with  
790 Next.js for performance. The FastAPI and Next.js instance communicates with the user  
791 through an Apache proxy to retrieve user requests and display results. Cite the Github address  
792

### 793 ***Analysis of publicly available transcriptomic data***

794 We collected raw data available at NCBI from the RNA-Seq short reads BioProjects  
795 published from *P. tricornutum* (different ecotypes and selected mutants), exposed to different  
796 conditions. Short read sequences were 35 to 150 bp long, principally generated from Illumina  
797 or DNBSeg. Fastq files were analyzed using the Bioinformatic pipeline nf-core/rnaseq (Patel  
798 et al., 2021). Briefly, raw reads were cleaned and merged before being mapped with Hisat2  
799 on their reference genome and quantified using featureCounts (Liao et al., 2014), an  
800 algorithm specifically designed to quickly and efficiently quantify the expression of  
801 transcripts using RNA-Seq data. Read coverage files are displayed on the genome browser of  
802 each gene page.

803

804 For each BioProject, pairwise comparisons were chosen to assess gene expression regulation  
805 in the different sample sets using the R package DESeq2 (Love et al., 2014). All the  
806 replicates deemed to be available and of good quality (based on mean quality score computed  
807 with FastQC) were used to estimate biological and technical variation. In studies  
808 investigating transcriptomic responses to environmental variations, the tested conditions were  
809 compared with their respective control groups. In cases involving mutant lines, precomputed  
810 comparisons were made between mutants and wild-types.

811

812 Gene expression analyses can be performed with the web application “integrated Differential  
813 Expression and Pathway analysis” (iDEP, Ge et al., 2018). Data can first be explored using  
814 heatmap, k-means clustering, and PCA. Differential expression analysis can be conducted  
815 through two different methods; limma and DESeq2 packages and several visualization plots  
816 are available (e.g., Venn diagrams, Volcano plots, genome maps). Based on Ensembl  
817 annotation, pathway enrichment can be performed from GO and KEGG annotation using  
818 several methods: GSEA, PAGE, GAGE or ReactomePA.

819

### 820 ***RNA-Seq data generated in this study over a two-week iron (Fe) starvation time-course***

821 Wild-type *P. tricornutum* v 1.86 cells were grown in ESAW medium at 19°C (Dorrell, PNAS  
822 2021), under 12hLight:12Dark cycle (50uE light). Fe-replete (Fe<sup>+</sup>) and Fe-depleted (Fe<sup>-</sup>)



823 ESAW media were both produced using iron-free reagents based on a protocol from Xia Gao  
824 (Gao et al., 2019). For the transcriptomic analyses, three different Fe limitation conditions  
825 were designed: three days Fe limitation as a short-term Fe- treatment (FeS), one week Fe  
826 limitation as a medium-term Fe- treatment (FeM) and two weeks Fe limitation as a long-term  
827 Fe- treatment (FeL). Fe-replete conditions (Fe+) were used as a control. RNA-Seq analysis  
828 were performed using two genetically distinct lines of *P. tricornutum* wild-type cells,  
829 transformed with pPhat and HA-Cas9 vectors without guide RNAs (Dorrell et al, 2024), to  
830 allow subsequent comparison of gene expression responses to mutants (data not shown).  
831 Reproducible transcriptome dynamics were observed for each culture, suggesting insertion  
832 of the pPhat and Cas9 vectors did not intrinsically bias Fe metabolism in these strains. Three  
833 biological replicates were performed for each cell line and condition. Cells were collected for  
834 the analyses at the exponential phase, and no other nutrient stress than Fe limitation  
835 influenced the results. Fast Fv/Fm (with 10% FP, 70% SP) tests were performed for all cell  
836 lines using a PAM (PAR-FluorPen FP 110, Photon Systems Instruments) prior to sampling.  
837 Fv/Fm values under Fe-replete values were measured at mean 0.62, suggesting that neither N  
838 nor Pi were growth-limiting in the media. Measured Fv/Fm values after 3 days Fe starvation  
839 were 0.55, suggesting Fe limitation of photosystem activity.

840

841 Total RNA was extracted from around 50 mg cell pellets by using the TRIzol reagent (T9424,  
842 Sigma-Aldrich) and according to (Dorrell et al, 2024). 24 DNase-treated RNA libraries (4  
843 conditions x 3 biological replica x 2 genetically distinct cell lines) were sequenced on a  
844 DNBseq Illumina platform (BGI Genomics Co., Ltd, Hongkong, China) with 100 bp paired-  
845 end sequencing. Raw reads were filtered by removing adaptor sequences, contamination and  
846 low-quality reads (reads containing over 40% bases with Q value < 20%) to obtain clean  
847 reads. Clean reads were mapped to the version 3 annotation of the *P. tricornutum* genome  
848 (Rastogi et al., 2018), and average TPM values for each gene in each library were calculated  
849 using DiatomicBase.

850

#### 851 **Author contributions**

852 AF and CB conceived and supervised the project. EV coordinated the project, provided initial  
853 databases and drafted the manuscript. NZ designed the website, loaded the data with the  
854 technical help of PV. SL designed and performed RNA-Seq analysis under progressive Fe  
855 limitation. HCC, RGD, CD, RM and AF performed and interpreted case studies. KV, MF,  
856 HCC, RGD, AF and CB provided critical suggestions to the manuscript. All authors  
857 proofread and approved the manuscript.

858

## 859 **Acknowledgements**

860 The authors would like to thank the research community working on molecular aspects of  
861 diatoms for their feedback during the creation and curation of the website. NZ and EV thank  
862 Dr Ge and his team from the South Dakota State University for help in deploying iDEP on  
863 DiatOmicBase. NZ would like to thank Catherine Le Bihan, Maël Lefeuvre, Nolwenn  
864 Lavielle, Phi Phong Nguyen from the IBENS computing service for their technical support.  
865 We thank Mariella Ferrante, Svenja Mager, and Anna Santin for helping us to add *Pseudo-*  
866 *nitzschia multistriata* to DiatOmicBase. KV acknowledges Michiel Van Bel and Emmelien  
867 Vancaester for technical assistance and data curation within PLAZA Diatoms.

868 This work was supported principally by a grant from Gordon and Betty Moore Foundation  
869 GBMF8752. CB acknowledges additional support from the European Research Council  
870 (ERC) under the European Union's Horizon 2020 research and innovation programme  
871 (Diatomic; grant agreement No. 835067), the French Government 'Investissements d'Avenir'  
872 programs MEMO LIFE (ANR-10-LABX-54) and PSL Research University (ANR-11-IDEX-  
873 0001-02). AF acknowledges funding from the Fondation Bettencourt-Schueller (Coups d'élan  
874 pour la recherche française-2018), the "Initiative d'Excellence" program (Grant  
875 "DYNAMO," ANR-11-LABX-0011-01) and EMBRC-FR – "Investments d'avenir" program  
876 (ANR-10-INBS-02). AF and CB acknowledges the ANR BrownCut (Project-ANR-19-CE20-  
877 0020) and Horizon Europe BlueRemediomics (grant agreement No. 101082304) projects.  
878 HCC acknowledges funding from ANR DiaLincs (ANR 19-CE43-0011-01). RGD  
879 acknowledges an ERC Starting Grant (grant number 101039760, "ChloroMosaic"). KV  
880 acknowledges Research Foundation-Flanders (FWO) for ELIXIR Belgium [I002819N] and  
881 BOF/GOA No. 01G01715 and 01G01323. MF acknowledges a Villum Young Investigator  
882 Grant (Villum Fonden grant number 37521).

883

## 884 **Conflict of Interest Statement**

885 The authors have no conflicts of interest to declare.

886

## 887 **Data statement**

888 Raw fastq data corresponding to new RNA-Seq performed under progressive Fe limitation is  
889 provided in NCBI BioProject number PRJNA936812.

## 890 **Supporting information provided in a separate file:**

891

892 **Figure S1** : CPS III (Phatr3\_J24195) genome browser capture displaying Phatr3 and Phatr2  
893 annotations as well as peptides mapped using a proteogenomic pipeline. A zoom in the 5'  
894 region shows that two peptide sequences are mapped on an exon predicted by Phatr3 gene  
895 model but not by Phatr2. This gene model is also predicted by mapping of RNA-Seq reads  
896 (here from a study of McCarthy et al., 2017).

897

898 **Figure S2** : A snapshot of the complete gene page for CPS III (Phatr3\_J24195).

899

900 **Figure S3** : Heatmap showing two-component regulator gene expression in light- and  
901 nutrient stress- related transcriptomes.

902

903 **Figure S4** : Heatmap displaying transcription factor expression in light and nutrient stress  
904 related experiments.

905

906 **Figure S5** : Schematic diagram of the culture regime used for Fe limitation experiments  
907 ( BioProject number PRJNA936812).

908

909

910 **References**

911

912 **Abbriano, R.M., George, J., Kahlke, T., Commault, A.S. and Fabris, M.** (2023)

913 Mobilization of a diatom mutator-like element () transposon inactivates the uridine

914 monophosphate synthase (UMPS) locus in *Phaeodactylum tricornutum*. *The Plant Journal*,

915 **115**, 926–936. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.16271>

916 [Accessed July 17, 2024].

917 **Agarwal, A., Di, R. and Falkowski, P.G.** (2022) Light-harvesting complex gene

918 regulation by a MYB-family transcription factor in the marine diatom, *Phaeodactylum*

919 *tricornutum*. *Photosynth Res*, **153**, 59–70. Available at: [https://doi.org/10.1007/s11120-](https://doi.org/10.1007/s11120-022-00915-w)

920 022-00915-w [Accessed July 22, 2024].

921 **Ait-Mohamed, O., Novák Vanclová, A.M.G., Joli, N., Liang, Y., Zhao, X., Genovesio,**

922 **A., Tirichine, L., Bowler, C. and Dorrell, R.G.** (2020) PhaeoNet: A Holistic RNAseq-

923 Based Portrait of Transcriptional Coordination in the Model Diatom *Phaeodactylum*

924 *tricornutum*. *Frontiers in Plant Science*, **11**. Available at:

925 <https://www.frontiersin.org/article/10.3389/fpls.2020.590949> [Accessed February 28,

926 2022].

927 **Allen, A.E., Dupont, C.L., Oborník, M., et al.,** (2011) Evolution and metabolic

928 significance of the urea cycle in photosynthetic diatoms. *Nature*, **473**, 203–207. Available

929 at: <https://www.nature.com/articles/nature10074> [Accessed February 28, 2022].

930 **Allen, A.E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.J.,**

931 **Finazzi, G., Fernie, A.R. and Bowler, C.** (2008) Whole-cell response of the pennate

932 diatom *Phaeodactylum tricornutum* to iron starvation. *Proceedings of the National*

933 *Academy of Sciences*, **105**, 10438–10443. Available at:

934 <https://www.pnas.org/doi/full/10.1073/pnas.0711370105> [Accessed July 19, 2024].

935 **Allen, J.F., Paula, W.B.M. de, Puthiyaveetil, S. and Nield, J.** (2011) A structural

936 phylogenetic map for chloroplast photosynthesis. *Trends in Plant Science*, **16**, 645–655.

937 Available at: [https://www.cell.com/trends/plant-science/abstract/S1360-1385\(11\)00226-3](https://www.cell.com/trends/plant-science/abstract/S1360-1385(11)00226-3)

938 [Accessed July 22, 2024].

939 **Alverson, A.J., Jansen, R.K. and Theriot, E.C.** (2007) Bridging the Rubicon:

940 Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by

941 thalassiosiroid diatoms. *Molecular Phylogenetics and Evolution*, **45**, 193–210. Available

942 at: <https://www.sciencedirect.com/science/article/pii/S1055790307001005> [Accessed

943 October 5, 2023].

- 944 **Annunziata, R., Ritter, A., Fortunato, A.E., et al.**, (2019) bHLH-PAS protein RITMO1  
945 regulates diel biological rhythms in the marine diatom *Phaeodactylum tricornutum*.  
946 *Proceedings of the National Academy of Sciences*, **116**, 13137–13142. Available at:  
947 <https://www.pnas.org/doi/abs/10.1073/pnas.1819660116> [Accessed July 22, 2024].
- 948 **Armbrust, E.V., Berges, J.A., Bowler, C., et al.**, (2004) The Genome of the Diatom  
949 *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism. *Science*, **306**, 79–86.  
950 Available at: <https://www.science.org/doi/abs/10.1126/science.1101156> [Accessed October  
951 28, 2021].
- 952 **Ashburner, M., Ball, C.A., Blake, J.A., et al.**, (2000) Gene Ontology: tool for the  
953 unification of biology. *Nat Genet*, **25**, 25–29. Available at:  
954 [https://www.nature.com/articles/ng0500\\_25](https://www.nature.com/articles/ng0500_25) [Accessed February 28, 2022].
- 955 **Ashworth, J., Turkarslan, S., Harris, M., Orellana, M.V. and Baliga, N.S.** (2016) Pan-  
956 transcriptomic analysis identifies coordinated and orthologous functional modules in the  
957 diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Marine Genomics*, **26**,  
958 21–28. Available at:  
959 <https://www.sciencedirect.com/science/article/pii/S1874778715300441> [Accessed October  
960 26, 2021].
- 961 **Bailleul, B., Berne, N., Murik, O., et al.**, (2015) Energetic coupling between plastids and  
962 mitochondria drives CO<sub>2</sub> assimilation in diatoms. *Nature*, **524**, 366–369. Available at:  
963 <https://www.nature.com/articles/nature14599> [Accessed October 5, 2023].
- 964 **Bailleul, B., Cardol, P., Breyton, C. and Finazzi, G.** (2010) Electrochromism: a useful  
965 probe to study algal photosynthesis. *Photosynth Res*, **106**, 179–189.
- 966 **Behnke, J. and LaRoche, J.** (2020) Iron uptake proteins in algae and the role of Iron  
967 Starvation-Induced Proteins (ISIPs). *European Journal of Phycology*, **55**, 339–360.  
968 Available at: <https://doi.org/10.1080/09670262.2020.1744039> [Accessed July 22, 2024].
- 969 **Blaby-Haas, C.E. and Merchant, S.S.** (2019) Comparative and Functional Algal  
970 Genomics. *Annu Rev Plant Biol*, **70**, 605–638.
- 971 **Bowler, C., Allen, A.E., Badger, J.H., et al.**, (2008) The *Phaeodactylum* genome reveals  
972 the evolutionary history of diatom genomes. *Nature*, **456**, 239–244. Available at:  
973 <https://www.nature.com/articles/nature07410> [Accessed October 28, 2021].
- 974 **Broddrick, J.T., Du, N., Smith, S.R., et al.**, (2019) Cross-compartment metabolic  
975 coupling enables flexible photoprotective mechanisms in the diatom *Phaeodactylum*

- 976 tricornutum. *New Phytol*, **222**, 1364–1379. Available at:  
977 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594073/> [Accessed July 22, 2024].
- 978 **Buck, J.M., Sherman, J., Bártulos, C.R., et al.**, (2019) Lhcx proteins provide  
979 photoprotection via thermal dissipation of absorbed light in the diatom *Phaeodactylum*  
980 tricornutum. *Nat Commun*, **10**, 4167. Available at:  
981 <https://www.nature.com/articles/s41467-019-12043-6> [Accessed October 28, 2021].
- 982 **Buels, R., Yao, E., Diesh, C.M., et al.**, (2016) JBrowse: a dynamic web platform for  
983 genome visualization and analysis. *Genome Biology*, **17**, 66. Available at:  
984 <https://doi.org/10.1186/s13059-016-0924-1> [Accessed March 3, 2022].
- 985 **Cohen, N.R., Mann, E., Stemple, B., Moreno, C.M., Rauschenberg, S., Jacquot, J.E.,**  
986 **Sunda, W.G., Twining, B.S. and Marchetti, A.** (2018) Iron storage capacities and  
987 associated ferritin gene expression among marine diatoms. *Limnology and Oceanography*,  
988 **63**, 1677–1691. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/lno.10800>  
989 [Accessed July 19, 2024].
- 990 **Cruz de Carvalho, M.H., Sun, H.-X., Bowler, C. and Chua, N.-H.** (2016) Noncoding  
991 and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New*  
992 *Phytologist*, **210**, 497–510. Available at:  
993 <https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.13787> [Accessed February 28, 2022].
- 994 **Debit, A., Charton, F., Pierre-Elies, P., Bowler, C. and Cruz de Carvalho, H.** (2023)  
995 Differential expression patterns of long noncoding RNAs in a pleiomorphic diatom and  
996 relation to hyposalinity. *Sci Rep*, **13**, 2440. Available at:  
997 <https://www.nature.com/articles/s41598-023-29489-w> [Accessed July 22, 2024].
- 998 **Dell’Aquila, G. and Maier, U.G.** (2020) Specific acclimations to phosphorus limitation in  
999 the marine diatom *Phaeodactylum tricornutum*. *Biological Chemistry*, **401**, 1495–1501.  
1000 Available at: <https://www.degruyter.com/document/doi/10.1515/hsz-2020-0197/html>  
1001 [Accessed February 28, 2022].
- 1002 **Delmont, T.O., Gaia, M., Hinsinger, D.D., et al.**, (2022) Functional repertoire  
1003 convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean.  
1004 *Cell Genomics*, **2**, 100123. Available at:  
1005 <https://www.sciencedirect.com/science/article/pii/S2666979X22000477> [Accessed May  
1006 17, 2023].
- 1007 **Dorrell, R.G., Zhang, Y., Liang, Y., et al.**, (2024) Complementary environmental  
1008 analysis and functional characterization of lower glycolysis-gluconeogenesis in the diatom

- 1009 plastid. *The Plant Cell*, koae168. Available at: <https://doi.org/10.1093/plcell/koae168>  
1010 [Accessed July 22, 2024].
- 1011 **Dyrløv Bendtsen, J., Nielsen, H., Heijne, G. von and Brunak, S.** (2004) Improved  
1012 Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783–795.  
1013 Available at: <https://www.sciencedirect.com/science/article/pii/S0022283604005972>  
1014 [Accessed June 2, 2023].
- 1015 **Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. and Mock, T.** (2020) Diatom  
1016 Molecular Research Comes of Age: Model Species for Studying Phytoplankton Biology  
1017 and Diversity[OPEN]. *The Plant Cell*, **32**, 547–572. Available at:  
1018 <https://doi.org/10.1105/tpc.19.00158> [Accessed October 28, 2021].
- 1019 **Falciatore, A. and Mock, T. eds.** (2022) *The Molecular Life of Diatoms*, Cham: Springer  
1020 International Publishing. Available at: [https://link.springer.com/10.1007/978-3-030-92499-](https://link.springer.com/10.1007/978-3-030-92499-7)  
1021 [7](https://link.springer.com/10.1007/978-3-030-92499-7) [Accessed July 19, 2024].
- 1022 **Ferrante, M.I., Broccoli, A. and Montresor, M.** (2023) The pennate diatom *Pseudo-*  
1023 *nitzschia multistriata* as a model for diatom life cycles, from the laboratory to the sea. *J*  
1024 *Phycol*, **59**, 637–643.
- 1025 **Filloramo, G.V., Curtis, B.A., Blanche, E. and Archibald, J.M.** (2021) Re-examination  
1026 of two diatom reference genomes using long-read sequencing. *BMC Genomics*, **22**, 379.  
1027 Available at: <https://doi.org/10.1186/s12864-021-07666-3> [Accessed May 17, 2023].
- 1028 **Flori, S., Jouneau, P.-H., Bailleul, B., et al.,** (2017) Plastid thylakoid architecture  
1029 optimizes photosynthesis in diatoms. *Nat Commun*, **8**, 15885. Available at:  
1030 <https://www.nature.com/articles/ncomms15885> [Accessed October 28, 2021].
- 1031 **Fukasawa, Y., Tsuji, J., Fu, S.-C., Tomii, K., Horton, P. and Imai, K.** (2015)  
1032 MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage  
1033 Sites. *Mol Cell Proteomics*, **14**, 1113–1126. Available at:  
1034 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4390256/> [Accessed June 2, 2023].
- 1035 **Gao, X., Bowler, C. and Kazamia, E.** (2021) Iron metabolism strategies in diatoms.  
1036 *Journal of Experimental Botany*, **72**, 2165–2180. Available at:  
1037 <https://doi.org/10.1093/jxb/eraa575> [Accessed October 28, 2021].
- 1038 **Ge, S.X., Son, E.W. and Yao, R.** (2018) iDEP: an integrated web application for  
1039 differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, **19**,



- 1040 534. Available at: <https://doi.org/10.1186/s12859-018-2486-6> [Accessed February 28,  
1041 2022].
- 1042 **Giguere, D.J., Bahcheli, A.T., Slattery, S.S., Patel, R.R., Browne, T.S., Flatley, M.,**  
1043 **Karas, B.J., Edgell, D.R. and Gloor, G.B.** (2022) Telomere-to-telomere genome  
1044 assembly of *Phaeodactylum tricornutum*. *PeerJ*, **10**, e13607.
- 1045 **Grouneva, I., Rokka, A. and Aro, E.-M.** (2011) The Thylakoid Membrane Proteome of  
1046 Two Marine Diatoms Outlines Both Diatom-Specific and Species-Specific Features of the  
1047 Photosynthetic Machinery. *J. Proteome Res.*, **10**, 5338–5353. Available at:  
1048 <https://doi.org/10.1021/pr200600f> [Accessed July 22, 2024].
- 1049 **Grossman, R.D., Parker, M.S. and Armbrust, E.V.** (2015) Diversity and Evolutionary  
1050 History of Iron Metabolism Genes in Diatoms. *PLOS ONE*, **10**, e0129081. Available at:  
1051 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129081> [Accessed July  
1052 22, 2024].
- 1053 **Gruber, A., Rocap, G., Kroth, P.G., Armbrust, E.V. and Mock, T.** (2015) Plastid  
1054 proteome prediction for diatoms and other algae with secondary plastids of the red lineage.  
1055 *Plant J*, **81**, 519–528. Available at:  
1056 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4329603/> [Accessed June 2, 2023].
- 1057 **Grypioti, E., Richard, H., Kryovrysanaki, N., Jaubert, M., Falciatore, A., Verret, F.**  
1058 **and Kalantidis, K.** (2024) Dicer-dependent heterochromatic small RNAs in the model  
1059 diatom species *Phaeodactylum tricornutum*. *New Phytologist*, **241**, 811–826. Available at:  
1060 <https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.19429> [Accessed July 22, 2024].
- 1061 **Gschloessl, B., Guermeur, Y. and Cock, J.M.** (2008) HECTAR: A method to predict  
1062 subcellular targeting in heterokonts. *BMC Bioinformatics*, **9**, 393. Available at:  
1063 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2567999/> [Accessed June 2, 2023].
- 1064 **Gundermann, K., Schmidt, M., Weisheit, W., Mittag, M. and Büchel, C.** (2013)  
1065 Identification of several sub-populations in the pool of light harvesting proteins in the  
1066 pennate diatom *Phaeodactylum tricornutum*. *Biochimica et Biophysica Acta (BBA) -*  
1067 *Bioenergetics*, **1827**, 303–310. Available at:  
1068 <https://www.sciencedirect.com/science/article/pii/S0005272812010742> [Accessed July 19,  
1069 2024].
- 1070 **Helliwell, K.E., Harrison, E.L., Christie-Oleza, J.A., et al.**, (2021) A Novel Ca<sup>2+</sup>  
1071 Signaling Pathway Coordinates Environmental Phosphorus Sensing and Nitrogen  
1072 Metabolism in Marine Diatoms. *Current Biology*, **31**, 978-989.e4. Available at:



1073 [https://www.cell.com/current-biology/abstract/S0960-9822\(20\)31821-2](https://www.cell.com/current-biology/abstract/S0960-9822(20)31821-2) [Accessed July 19,  
1074 2024].

1075 **Hermann, D., Egue, F., Tastard, E., et al.,** (2014) An introduction to the vast world of  
1076 transposable elements – what about the diatoms? *Diatom Research*, **29**, 91–104. Available  
1077 at: <https://doi.org/10.1080/0269249X.2013.877083> [Accessed October 28, 2021].

1078 **Heydarizadeh, P., Marchand, J., Chenais, B., Sabzalian, M.R., Zahedi, M., Moreau,**  
1079 **B. and Schoefs, B.** (2014) Functional investigations in diatoms need more than a  
1080 transcriptomic approach. *Diatom Research*, **29**, 75–89. Available at:  
1081 <https://doi.org/10.1080/0269249X.2014.883727> [Accessed February 28, 2022].

1082 **Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and**  
1083 **Nakai, K.** (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, **35**,  
1084 W585–W587. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933216/>  
1085 [Accessed June 2, 2023].

1086 **Jaubert, M., Duchêne, C., Kroth, P., Rogato, A., Bouly, J.-P. and Falciatore, A.** (2022)  
1087 Sensing and Signalling in Diatom Responses to Abiotic Cues. In pp. 607–639.

1088 **Johnson, G.N., Cardol, P., Minagawa, J. and Finazzi, G.** (2014) Regulation of Electron  
1089 Transport in Photosynthesis. In S. M. Theg and F.-A. Wollman, eds. *Plastid Biology*. New  
1090 York, NY: Springer, pp. 437–464. Available at: [https://doi.org/10.1007/978-1-4939-1136-](https://doi.org/10.1007/978-1-4939-1136-3_16)  
1091 [3\\_16](https://doi.org/10.1007/978-1-4939-1136-3_16) [Accessed July 22, 2024].

1092 **Joshi-Deo, J., Schmidt, M., Gruber, A., Weisheit, W., Mittag, M., Kroth, P.G. and**  
1093 **Büchel, C.** (2010) Characterization of a trimeric light-harvesting complex in the diatom  
1094 *Phaeodactylum tricornutum* built of FcpA and FcpE proteins. *Journal of Experimental*  
1095 *Botany*, **61**, 3079–3087. Available at: <https://doi.org/10.1093/jxb/erq136> [Accessed July  
1096 22, 2024].

1097 **Kanehisa, M.** (2002) The KEGG database. *Novartis Found Symp*, **247**, 91–101; discussion  
1098 101-103, 119–128, 244–252.

1099 **Kazamia, E., Mach, J., McQuaid, J.B., et al.,** (2022) In vivo localization of iron  
1100 starvation induced proteins under variable iron supplementation regimes in. *Plant Direct*,  
1101 **6**, e472. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pld3.472> [Accessed  
1102 July 22, 2024].

1103 **Kazamia, E., Sutak, R., Paz-Yepes, J., et al.,** (2018) Endocytosis-mediated siderophore  
1104 uptake as a strategy for Fe acquisition in diatoms. *Science Advances*, **4**, eaar4536.

1105 Available at: <https://www.science.org/doi/full/10.1126/sciadv.aar4536> [Accessed July 22,  
1106 2024].

1107 **Keeling, P.J., Burki, F., Wilcox, H.M., et al.,** (2014) The Marine Microbial Eukaryote  
1108 Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of  
1109 Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, **12**,  
1110 e1001889. Available at:  
1111 <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001889> [Accessed  
1112 February 28, 2022].

1113 **Kumar, M., Zuniga, C., Tibocha-Bonilla, J.D., Smith, S.R., Coker, J., Allen, A.E. and**  
1114 **Zengler, K.** (2022) Constraint-Based Modeling of Diatoms Metabolism and Quantitative  
1115 Biology Approaches. In A. Falciatore and T. Mock, eds. *The Molecular Life of Diatoms*.  
1116 Cham: Springer International Publishing, pp. 775–808. Available at:  
1117 [https://doi.org/10.1007/978-3-030-92499-7\\_26](https://doi.org/10.1007/978-3-030-92499-7_26) [Accessed May 22, 2023].

1118 **Kwon, D.Y., Vuong, T.T., Choi, J., Lee, T.S., Um, J.-I., Koo, S.Y., Hwang, K.T. and**  
1119 **Kim, S.M.** (2021) Fucoxanthin biosynthesis has a positive correlation with the specific  
1120 growth rate in the culture of microalga *Phaeodactylum tricornutum*. *J Appl Phycol*, **33**,  
1121 1473–1485. Available at: <https://doi.org/10.1007/s10811-021-02376-5> [Accessed July 22,  
1122 2024].

1123 **Lampe, R.H., Mann, E.L., Cohen, N.R., Till, C.P., Thamatrakoln, K., Brzezinski,**  
1124 **M.A., Bruland, K.W., Twining, B.S. and Marchetti, A.** (2018) Different iron storage  
1125 strategies among bloom-forming diatoms. *Proceedings of the National Academy of*  
1126 *Sciences*, **115**, E12275–E12284. Available at:  
1127 <https://www.pnas.org/doi/full/10.1073/pnas.1805243115> [Accessed July 19, 2024].

1128 **Leister, D., Marino, G., Minagawa, J. and Dann, M.** (2022) An ancient function of  
1129 PGR5 in iron delivery? *Trends in Plant Science*, **27**, 971–980. Available at:  
1130 [https://www.cell.com/trends/plant-science/abstract/S1360-1385\(22\)00127-3](https://www.cell.com/trends/plant-science/abstract/S1360-1385(22)00127-3) [Accessed  
1131 July 22, 2024].

1132 **Lepetit, B., Campbell, D., Lavaud, J., Büchel, C., Goss, R. and Bailleul, B.** (2022)  
1133 Photosynthetic Light Reactions in Diatoms. II. The Dynamic Regulation of the Various  
1134 Light Reactions. In Springer International Publishing, p. 423. Available at:  
1135 <https://hal.science/hal-03672201> [Accessed July 19, 2024].

1136 **Levitan, O., Dinamarca, J., Zelzion, E., et al.,** (2015) Remodeling of intermediate  
1137 metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress. *PNAS*, **112**,

- 1138 412–417. Available at: <https://www.pnas.org/content/112/2/412> [Accessed February 28,  
1139 2022].
- 1140 **Li, J., Zhang, K., Lin, X., Li, L. and Lin, S.** (2022) Phytate as a Phosphorus Nutrient  
1141 with Impacts on Iron Stress-Related Gene Expression for Phytoplankton: Insights from the  
1142 Diatom *Phaeodactylum tricornutum*. *Applied and Environmental Microbiology*, **88**,  
1143 e02097-21. Available at: <https://journals.asm.org/doi/10.1128/aem.02097-21> [Accessed  
1144 July 22, 2024].
- 1145 **Liao, Y., Smyth, G.K. and Shi, W.** (2014) featureCounts: an efficient general purpose  
1146 program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
- 1147 **Lodeyro, A.F., Ceccoli, R.D., Pierella Karlusich, J.J. and Carrillo, N.** (2012) The  
1148 importance of flavodoxin for environmental stress tolerance in photosynthetic  
1149 microorganisms and transgenic plants. Mechanism, evolution and biotechnological  
1150 potential. *FEBS Letters*, **586**, 2917–2924. Available at:  
1151 <https://www.sciencedirect.com/science/article/pii/S0014579312005947> [Accessed July 22,  
1152 2024].
- 1153 **Lommer, M., Specht, M., Roy, A.-S., et al.,** (2012) Genome and low-iron response of an  
1154 oceanic diatom adapted to chronic iron limitation. *Genome Biology*, **13**, R66. Available at:  
1155 <http://dx.doi.org/10.1186/gb-2012-13-7-r66> [Accessed June 26, 2016].
- 1156 **Love, M.I., Huber, W. and Anders, S.** (2014) Moderated estimation of fold change and  
1157 dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550. Available at:  
1158 <https://doi.org/10.1186/s13059-014-0550-8> [Accessed February 28, 2022].
- 1159 **Madhuri, S., Lepetit, B., Fürst, A.H. and Kroth, P.G.** (2024) A Knockout of the  
1160 Photoreceptor *PtAureo1a* Results in Altered Diel Expression of Diatom Clock  
1161 Components. *Plants*, **13**, 1465. Available at: <https://www.mdpi.com/2223-7747/13/11/1465>  
1162 [Accessed July 22, 2024].
- 1163 **Malviya, S., Scalco, E., Audic, S., et al.,** (2016) Insights into global diatom distribution  
1164 and diversity in the world’s ocean. *PNAS*, **113**, E1516–E1525. Available at:  
1165 <https://www.pnas.org/content/113/11/E1516> [Accessed October 26, 2021].
- 1166 **Mann, M., Serif, M., Wrobel, T., et al.,** (2020) The Aureochrome Photoreceptor  
1167 *PtAUREO1a* Is a Highly Effective Blue Light Switch in Diatoms. *iScience*, **23**. Available  
1168 at: [https://www.cell.com/iscience/abstract/S2589-0042\(20\)30927-5](https://www.cell.com/iscience/abstract/S2589-0042(20)30927-5) [Accessed February 28,  
1169 2022].

- 1170 **Marchetti, A., Parker, M.S., Moccia, L.P., Lin, E.O., Arrieta, A.L., Ribalet, F.,**  
1171 **Murphy, M.E.P., Maldonado, M.T. and Armbrust, E.V.** (2009) Ferritin is used for iron  
1172 storage in bloom-forming marine pennate diatoms. *Nature*, **457**, 467–470. Available at:  
1173 <https://www.nature.com/articles/nature07539> [Accessed July 22, 2024].
- 1174 **Martino, A.D., Meichenin, A., Shi, J., Pan, K. and Bowler, C.** (2007) Genetic and  
1175 phenotypic characterization of *Phaeodactylum tricornutum* (Bacillariophyceae)  
1176 accessions1. *Journal of Phycology*, **43**, 992–1009. Available at:  
1177 <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1529-8817.2007.00384.x> [Accessed  
1178 February 28, 2022].
- 1179 **Matthijs, M., Fabris, M., Broos, S., Vyverman, W. and Goossens, A.** (2016) Profiling  
1180 of the Early Nitrogen Stress Response in the Diatom *Phaeodactylum tricornutum* Reveals a  
1181 Novel Family of RING-Domain Transcription Factors. *Plant Physiology*, **170**, 489–498.  
1182 Available at: <https://doi.org/10.1104/pp.15.01300> [Accessed February 28, 2022].
- 1183 **Matthijs, M., Fabris, M., Obata, T., Foubert, I., Franco-Zorrilla, J.M., Solano, R.,**  
1184 **Fernie, A.R., Vyverman, W. and Goossens, A.** (2017) The transcription factor bZIP14  
1185 regulates the TCA cycle in the diatom *Phaeodactylum tricornutum*. *EMBO J*, **36**, 1559–  
1186 1576. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5452028/> [Accessed  
1187 February 28, 2022].
- 1188 **Mattick, J.S.** (2023) RNA out of the mist. *Trends in Genetics*, **39**, 187–207. Available at:  
1189 [https://www.cell.com/trends/genetics/abstract/S0168-9525\(22\)00288-8](https://www.cell.com/trends/genetics/abstract/S0168-9525(22)00288-8) [Accessed July 22,  
1190 2024].
- 1191 **McCarthy, J.K., Smith, S.R., McCrow, J.P., et al.,** (2017) Nitrate Reductase Knockout  
1192 Uncouples Nitrate Transport from Nitrate Assimilation and Drives Repartitioning of  
1193 Carbon Flux in a Model Pennate Diatom. *The Plant Cell*, **29**, 2047–2070. Available at:  
1194 <https://doi.org/10.1105/tpc.16.00910> [Accessed February 28, 2022].
- 1195 **McQuaid, J.B., Kustka, A.B., Oborník, M., et al.,** (2018) Carbonate-sensitive  
1196 phytoferritin controls high-affinity iron uptake in diatoms. *Nature*, **555**, 534–537.  
1197 Available at: <https://www.nature.com/articles/nature25982> [Accessed July 22, 2024].
- 1198 **Mock, T., Otilar, R.P., Strauss, J., et al.,** (2017) Evolutionary genomics of the cold-  
1199 adapted diatom *Fragilariopsis cylindrus*. *Nature*, **541**, 536–540. Available at:  
1200 <https://www.nature.com/articles/nature20803> [Accessed July 16, 2024].
- 1201 **Morrissey, J., Sutak, R., Paz-Yepes, J., et al.,** (2015) A Novel Protein, Ubiquitous in  
1202 Marine Phytoplankton, Concentrates Iron at the Cell Surface and Facilitates Uptake.

- 1203 *Current Biology*, **25**, 364–371. Available at:  
1204 <http://www.sciencedirect.com/science/article/pii/S0960982214015632> [Accessed July 5,  
1205 2016].
- 1206 **Nagao, R., Yokono, M., Ueno, Y., et al.**, (2021) Enhancement of excitation-energy  
1207 quenching in fucoxanthin chlorophyll *a/c*-binding proteins isolated from a diatom  
1208 *Phaeodactylum tricornutum* upon excess-light illumination. *Biochimica et Biophysica Acta*  
1209 (*BBA*) - *Bioenergetics*, **1862**, 148350. Available at:  
1210 <https://www.sciencedirect.com/science/article/pii/S0005272820302000> [Accessed July 19,  
1211 2024].
- 1212 **Nef, C., Madoui, M.-A., Pelletier, É. and Bowler, C.** (2022) Whole-genome scanning  
1213 reveals environmental selection mechanisms that shape diversity in populations of the  
1214 epipelagic diatom *Chaetoceros*. *PLOS Biology*, **20**, e3001893. Available at:  
1215 <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001893> [Accessed  
1216 July 16, 2024].
- 1217 **Nemoto, M., Iwaki, S., Moriya, H., Monden, Y., Tamura, T., Inagaki, K., Mayama, S.**  
1218 **and Obuse, K.** (2020) Comparative Gene Analysis Focused on Silica Cell Wall  
1219 Formation: Identification of Diatom-Specific SET Domain Protein Methyltransferases.  
1220 *Mar Biotechnol*, **22**, 551–563. Available at: <https://doi.org/10.1007/s10126-020-09976-1>  
1221 [Accessed October 25, 2021].
- 1222 **Osuna-Cruz, C.M., Bilcke, G., Vancaester, E., et al.**, (2020) The *Seminavis robusta*  
1223 genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat*  
1224 *Commun*, **11**, 3320.
- 1225 **Oudot-Le Secq, M.-P. and Green, B.R.** (2011) Complex repeat structures and novel  
1226 features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and  
1227 *Thalassiosira pseudonana*. *Gene*, **476**, 20–26. Available at:  
1228 <https://www.sciencedirect.com/science/article/pii/S0378111911000527> [Accessed October  
1229 28, 2021].
- 1230 **Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E.V., Bowler, C. and**  
1231 **Green, B.R.** (2007) Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and  
1232 *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage.  
1233 *Molecular Genetics and Genomics*, **4**, 427–439. Available at:  
1234 [https://www.infona.pl/resource/bwmeta1.element.springer-55ad8850-d4ac-36e0-bac5-](https://www.infona.pl/resource/bwmeta1.element.springer-55ad8850-d4ac-36e0-bac5-69f48e88b369)  
1235 [69f48e88b369](https://www.infona.pl/resource/bwmeta1.element.springer-55ad8850-d4ac-36e0-bac5-69f48e88b369) [Accessed October 26, 2021].

- 1236 **Patel, H., Ewels, P., Peltzer, A., et al.,** (2021) *nf-core/rnaseq: nf-core/rnaseq v3.5 -*  
1237 *Copper Chameleon*, Zenodo. Available at: <https://zenodo.org/record/5789421> [Accessed  
1238 February 28, 2022].
- 1239 **Poulsen, N. and Kröger, N.** (2023) *Thalassiosira pseudonana* (*Cyclotella nana*) (Hustedt)  
1240 Hasle et Heimdal (Bacillariophyceae): A genetically tractable model organism for studying  
1241 diatom biology, including biological silica formation. *Journal of Phycology*, **n/a**. Available  
1242 at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jpy.13362> [Accessed October 5, 2023].
- 1243 **Rastogi, A., Maheswari, U., Dorrell, R.G., et al.,** (2018) Integrative analysis of large  
1244 scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum*  
1245 genome and evolutionary origin of diatoms. *Sci Rep*, **8**, 4834. Available at:  
1246 <https://www.nature.com/articles/s41598-018-23106-x> [Accessed October 28, 2021].
- 1247 **Rastogi, A., Vieira, F.R.J., Deton-Cabanillas, A.-F., et al.,** (2020) A genomics approach  
1248 reveals the global genetic polymorphism, structure, and functional diversity of ten  
1249 accessions of the marine model diatom *Phaeodactylum tricornutum*. *ISME J*, **14**, 347–363.  
1250 Available at: <https://www.nature.com/articles/s41396-019-0528-3> [Accessed February 28,  
1251 2022].
- 1252 **Rogato, A., Amato, A., Iudicone, D., Chiurazzi, M., Ferrante, M.I. and Alcalà, M.R.**  
1253 **d'** (2015) The diatom molecular toolkit to handle nitrogen uptake. *Marine Genomics*, **24**,  
1254 95–108. Available at:  
1255 <https://www.sciencedirect.com/science/article/pii/S1874778715000951> [Accessed October  
1256 28, 2021].
- 1257 **Rogato, A., Richard, H., Sarazin, A., et al.,** (2014) The diversity of small non-coding  
1258 RNAs in the diatom *Phaeodactylum tricornutum*. *BMC Genomics*, **15**, 698. Available at:  
1259 <https://doi.org/10.1186/1471-2164-15-698> [Accessed March 3, 2022].
- 1260 **Russo, M.T., Rogato, A., Jaubert, M., Karas, B.J. and Falciatore, A.** (2023)  
1261 *Phaeodactylum tricornutum*: An established model species for diatom molecular research  
1262 and an emerging chassis for algal synthetic biology. *Journal of Phycology*, **59**, 1114–1122.  
1263 Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jpy.13400> [Accessed July 19,  
1264 2024].
- 1265 **Sato, S., Nanjappa, D., Dorrell, R.G., et al.,** (2020) Genome-enabled phylogenetic and  
1266 functional reconstruction of an araphid pennate diatom *Plagiosiriata* sp. CCMP470,  
1267 previously assigned as a radial centric diatom, and its bacterial commensal. *Sci Rep*, **10**,



- 1268 9449. Available at: <https://www.nature.com/articles/s41598-020-65941-x> [Accessed May  
1269 17, 2023].
- 1270 **Sétif, P.** (2001) Ferredoxin and flavodoxin reduction by photosystem I. *Biochimica et*  
1271 *Biophysica Acta (BBA) - Bioenergetics*, **1507**, 161–179. Available at:  
1272 <https://www.sciencedirect.com/science/article/pii/S0005272801002055> [Accessed July 22,  
1273 2024].
- 1274 **Shen, C., Dupont, C.L. and Hopkinson, B.M.** (2017) The diversity of CO<sub>2</sub>-concentrating  
1275 mechanisms in marine diatoms as inferred from their genetic content. *Journal of*  
1276 *Experimental Botany*, **68**, 3937–3948. Available at: <https://doi.org/10.1093/jxb/erx163>  
1277 [Accessed October 28, 2021].
- 1278 **Singer, D., Seppey, C.V.W., Lentendu, G., et al.,** (2021) Protist taxonomic and  
1279 functional diversity in soil, freshwater and marine ecosystems. *Environment International*,  
1280 **146**, 106262. Available at:  
1281 <https://www.sciencedirect.com/science/article/pii/S0160412020322170> [Accessed May 17,  
1282 2023].
- 1283 **Smith, S.R., Dupont, C.L., McCarthy, J.K., et al.,** (2019) Evolution and regulation of  
1284 nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nat*  
1285 *Commun*, **10**, 4552. Available at: <https://www.nature.com/articles/s41467-019-12407-y>  
1286 [Accessed February 28, 2022].
- 1287 **Smith, S.R., Gillard, J.T.F., Kustka, A.B., et al.,** (2016) Transcriptional Orchestration of  
1288 the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron  
1289 Limitation. *PLOS Genetics*, **12**, e1006490. Available at:  
1290 <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006490> [Accessed  
1291 February 28, 2022].
- 1292 **Song, J., Zhao, H., Zhang, L., Li, Z., Han, J., Zhou, C., Xu, J., Li, X. and Yan, X.**  
1293 (2023) The Heat Shock Transcription Factor PtHSF1 Mediates Triacylglycerol and  
1294 Fucoxanthin Synthesis by Regulating the Expression of GPAT3 and DXS in  
1295 *Phaeodactylum tricornutum*. *Plant Cell Physiol*, **64**, 622–636.
- 1296 **Statello, L., Guo, C.-J., Chen, L.-L. and Huarte, M.** (2021) Gene regulation by long  
1297 non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*, **22**, 96–118.  
1298 Available at: <https://www.nature.com/articles/s41580-020-00315-9> [Accessed July 22,  
1299 2024].

- 1300 **Tréguer, P.J., Sutton, J.N., Brzezinski, M., et al.,** (2021) Reviews and syntheses: The  
1301 biogeochemical cycle of silicon in the modern ocean. *Biogeosciences*, **18**, 1269–1289.  
1302 Available at: <https://bg.copernicus.org/articles/18/1269/2021/> [Accessed October 28,  
1303 2021].
- 1304 **Turnšek, J., Brunson, J.K., Viedma, M. del P.M., Deerinck, T.J., Horák, A., Oborník,**  
1305 **M., Bielinski, V.A. and Allen, A.E.** (2021) Proximity proteomics in a marine diatom  
1306 reveals a putative cell surface-to-chloroplast iron trafficking pathway C. S. Hardtke, J.  
1307 Kleine-Vehn, and C. Brownlee, eds. *eLife*, **10**, e52770. Available at:  
1308 <https://doi.org/10.7554/eLife.52770> [Accessed July 22, 2024].
- 1309 **Vancaester, E., Depuydt, T., Osuna-Cruz, C.M. and Vandepoele, K.** (2020)  
1310 Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms.  
1311 *Molecular Biology and Evolution*, **37**, 3243–3257. Available at:  
1312 <https://doi.org/10.1093/molbev/msaa182> [Accessed July 16, 2024].
- 1313 **Vandepoele, K., Van Bel, M., Richard, G., Van Landeghem, S., Verhelst, B., Moreau,**  
1314 **H., Van de Peer, Y., Grimsley, N. and Piganeau, G.** (2013) pico-PLAZA, a genome  
1315 database of microbial photosynthetic eukaryotes. *Environmental Microbiology*, **15**, 2147–  
1316 2153. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.12174>  
1317 [Accessed February 28, 2022].
- 1318 **Vanormelingen, P., Verleyen, E. and Vyverman, W.** (2009) The diversity and  
1319 distribution of diatoms: from cosmopolitanism to narrow endemism. In W. Foissner and D.  
1320 L. Hawksworth, eds. *Protist Diversity and Geographical Distribution*. Topics in  
1321 Biodiversity and Conservation. Dordrecht: Springer Netherlands, pp. 159–171. Available  
1322 at: [https://doi.org/10.1007/978-90-481-2801-3\\_12](https://doi.org/10.1007/978-90-481-2801-3_12) [Accessed October 28, 2021].
- 1323 **Veluchamy, A., Lin, X., Maumus, F., et al.,** (2013) Insights into the role of DNA  
1324 methylation in diatoms by genome-wide profiling in *Phaeodactylum tricorutum*. *Nat*  
1325 *Commun*, **4**, 2091. Available at: <https://www.nature.com/articles/ncomms3091> [Accessed  
1326 October 28, 2021].
- 1327 **Veluchamy, A., Rastogi, A., Lin, X., et al.,** (2015) An integrative analysis of post-  
1328 translational histone modifications in the marine diatom *Phaeodactylum tricorutum*.  
1329 *Genome Biology*, **16**, 102. Available at: <https://doi.org/10.1186/s13059-015-0671-8>  
1330 [Accessed October 28, 2021].
- 1331 **Vernette, C., Lecubin, J., Sánchez, P., et al.,** (2022) The Ocean Gene Atlas v2.0: online  
1332 exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Research*,

1333 **50**, W516–W526. Available at: <https://doi.org/10.1093/nar/gkac420> [Accessed July 19,  
1334 2024].

1335 **Wu, Y., Chaumier, T., Manirakiza, E., Veluchamy, A. and Tirichine, L.** (2023)  
1336 PhaeoEpiView: an epigenome browser of the newly assembled genome of the model  
1337 diatom *Phaeodactylum tricornutum*. *Sci Rep*, **13**, 8320. Available at:  
1338 <https://www.nature.com/articles/s41598-023-35403-1> [Accessed July 16, 2024].

1339 **Yang, M., Lin, X., Liu, X., Zhang, J. and Ge, F.** (2018) Genome Annotation of a Model  
1340 Diatom *Phaeodactylum tricornutum* Using an Integrated Proteogenomic Pipeline.  
1341 *Molecular Plant*, **11**, 1292–1307. Available at:  
1342 <https://www.sciencedirect.com/science/article/pii/S1674205218302508> [Accessed  
1343 February 28, 2022].

1344 **Zhao, X., Rastogi, A., Deton Cabanillas, A.F., et al.,** (2021) Genome wide natural  
1345 variation of H3K27me3 selectively marks genes predicted to be important for cell  
1346 differentiation in *Phaeodactylum tricornutum*. *New Phytologist*, **229**, 3208–3220. Available  
1347 at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.17129> [Accessed May 17, 2023].

1348 **Zhou, L., Gao, S., Yang, W., Wu, S., Huan, L., Xie, X., Wang, X., Lin, S. and Wang,**  
1349 **G.** (2022) Transcriptomic and metabolic signatures of diatom plasticity to light  
1350 fluctuations. *Plant Physiology*, **190**, 2295–2314. Available at:  
1351 <https://doi.org/10.1093/plphys/kiac455> [Accessed July 22, 2024].  
1352