

# Profiling of the Early Nitrogen Stress Response in the Diatom *Phaeodactylum tricornutum* Reveals a Novel Family of RING-Domain Transcription Factors<sup>1</sup>[OPEN]

Michiel Matthijs<sup>2</sup>, Michele Fabris<sup>3</sup>, Stefan Broos, Wim Vyverman, and Alain Goossens\*

Department of Plant Systems Biology (M.M., M.F., A.G.) and Inflammation Research Center (S.B.), VIB, B-9052 Gent, Belgium; Department of Plant Biotechnology and Bioinformatics (M.M., M.F., A.G.) and Department of Biomedical Molecular Biology (S.B.), Ghent University, B-9052 Gent, Belgium; and Department of Biology, Laboratory of Protistology and Aquatic Ecology, Ghent University, B-9000 Gent, Belgium (M.M., M.F., W.V.)

ORCID ID: 0000-0002-1599-551X (A.G.).

Diatoms often inhabit highly variable habitats where they are confronted with a wide variety of stresses, frequently including starvation of nutrients such as nitrogen. In this study, the transcriptome of the model diatom *Phaeodactylum tricornutum* was profiled during the onset of nitrogen starvation by RNA sequencing, and overrepresented motifs were determined in promoters of genes that were early and strongly up-regulated during the nitrogen stress response. One of these motifs could be bound by a nitrogen starvation-inducible RING-domain protein termed RING-GAF-Gln-containing protein (RGQ1), which was shown to act as a transcription factor and belongs to a previously uncharacterized family that is conserved in heterokont algae.

In contrast to terrestrial habitats, photosynthesis in the ocean is performed by organisms with a variable phylogenetic background (Falkowski et al., 2004). Diatoms, belonging to the heterokont superphylum, are important components of marine phytoplankton (Tozzi et al., 2004) and are believed to include more than 100,000 species (Armbrust, 2009). Diatoms are unique organisms, not only due to their origin or their silica cell wall (Tirichine and Bowler, 2011), but also because their genomes contain metabolic pathways that were unexpected for photosynthetic organisms, such as a complete urea cycle, the Entner-Doudoroff glycolytic pathway, a chimeric sterol biosynthesis pathway, and several carbon concentrating mechanisms aiding

photosynthesis (Giordano et al., 2005; Allen et al., 2011; Fabris et al., 2012, 2014).

Besides light, availability of nutrients is the main limiting factor of marine primary productivity (Elser et al., 2007). Diatoms become particularly abundant when there is an influx of nutrients, either seasonal or when the ocean is artificially fertilized with iron, resulting in large blooms (Moore et al., 2001). One possible explanation for the success of diatoms in marine habitats is their ability to tolerate long-term nutrient deprivation and rapidly assimilate available nutrients when conditions change. These features must be encoded in their genomes. To date, several diatom genomes have been sequenced, among which two are considered model species, the centricate *Thalassiosira pseudonana* and the pennate *Phaeodactylum tricornutum* (Armbrust et al., 2004; Bowler et al., 2008). Other currently sequenced diatom genomes include *Fragilariopsis cylindrus* and *Pseudo-nitzschia multiseriata* (Port et al., 2013) and *Thalassiosira oceanica* (Lommer et al., 2012); however, the latter two have not been fully assembled yet. The analysis of these genomes showed that there is a large fraction of diatom-specific genes, that diatoms have acquired genes from a variety of other organisms through horizontal gene transfer, and that often relatively little overlap exists between diatom species (Bowler et al., 2008). This likely sets the base for species-specific responses to nutrient deprivation, e.g. the lack of the ferritin and plastocyanin genes in *T. pseudonana* makes it less adapted to iron starvation compared to species that have alternative electron carriers such as *T. oceanica* and *P. tricornutum* (Lommer et al., 2012).

Nutrient depletion often stimulates lipid accumulation in algae, with nitrogen starvation as one of the most prominent elicitors (Mühlroth et al., 2013). Because of the

<sup>1</sup> This work was supported by the Agency for Innovation by Science and Technology in Flanders (Strategisch Basisonderzoek grant no. 80031 and predoctoral fellowship to M.M.).

<sup>2</sup> Present address: Algenuity, a Spicer Consulting Limited Subsidiary, Eden Laboratory, Broadmead Road, Bedford MK43 9ND, United Kingdom.

<sup>3</sup> Present address: Plant Functional Biology and Climate Change Cluster, University of Technology Sydney, 15 Broadway, Ultimo (Sydney), New South Wales 2007, Australia.

\* Address correspondence to [algoo@psb.vib-ugent.be](mailto:algoo@psb.vib-ugent.be).

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors ([www.plantphysiol.org](http://www.plantphysiol.org)) is: Alain Goossens ([algoo@psb.vib-ugent.be](mailto:algoo@psb.vib-ugent.be)).

M.M., M.F., and S.B. performed the experiments; M.M. designed the experiments and analyzed the data; W.V. and A.G. supervised the experiments; M.M., W.V., and A.G. conceived the project and wrote the article with contributions of all the authors.

[OPEN] Articles can be viewed without a subscription.

[www.plantphysiol.org/cgi/doi/10.1104/pp.15.01300](http://www.plantphysiol.org/cgi/doi/10.1104/pp.15.01300)

inherent economic potential, several studies have investigated nitrogen starvation in *P. tricornutum* and *T. pseudonana* (Hockin et al., 2012; Valenzuela et al., 2012; Yang et al., 2013; Alipanah et al., 2015; Levitan et al., 2015). The most prevalent finding from the transcriptome analyses in these studies was that the accumulation of lipids appeared to be a consequence of the remodeling of central metabolism. This is at least partially based on transcriptional reprogramming of the tricarboxylic acid and the urea cycles, which is needed to reallocate carbon and nitrogen from the halted photosynthesis and protein synthesis and the increased catabolism of amino acids, a finding which is also supported by proteomic data (Ge et al., 2014). This particular reprogramming appeared to be largely similar between the two model diatom species, suggesting conserved regulatory mechanisms, which hitherto have remained undisclosed.

Most of the studies mentioned above have focused on diatom cells that had been acclimated to the lack of nitrogen over periods longer than 1 d. However, many transcriptome changes in response to nutrient starvation will be transient. For instance, regulatory genes, which may be part of feedback or amplification loops, will only be switched on transiently during the early stages of starvation (Mangan et al., 2006). Undoubtedly some of the most important regulators of gene expression are transcription factors (TFs). Even though the numbers of different TF families can vary enormously between species, the proteins themselves are usually easily identifiable because of their conserved DNA-binding domains. Likewise, their binding motifs can be predicted to some degree (Weirauch et al., 2014). Using protein similarity as a tool, it was estimated that roughly 2% of the genomes of *P. tricornutum* and *T. pseudonana* code for TFs (Rayko et al., 2010). Nonetheless, thus far there have only been two diatom TFs characterized, and both are *P. tricornutum* bZIP proteins (Huysman et al., 2010; Ohno et al., 2012).

Here, RNA-sequencing (RNA-Seq)-based transcriptome analysis, computational promoter analysis, and yeast one-hybrid (Y1H) screenings were used to map regulatory elements in the transcriptional rewiring of *P. tricornutum* metabolism during nitrogen starvation. Therefore, we focused on the first twenty hours after nitrogen removal in order to see the transcriptome changes during the initial stages of starvation. Y1H screening was chosen to investigate potential TF binding sites because of its ability to screen a cDNA library for DNA binding. This is especially useful for the identification of previously unannotated TFs and for organisms, such as diatoms, for which a TF library is not yet available.

## RESULTS AND DISCUSSION

### Profiling of the Early Nitrogen Stress Response Transcriptome

In contrast to the published transcriptome studies listed above, which all focused on diatom cells subjected to relatively long-term (2 d or more) nitrogen starvation, the analysis presented here was centered on

earlier time points, to map the transcriptional reprogramming of diatoms during the onset of the nitrogen stress response. To this end, exponentially growing *P. tricornutum* cells were transferred to nitrogen-replete medium, and sampled 4, 8, and 20 h after medium transfer. During this period, nitrogen-starved cells started to accumulate more lipids, began bleaching, and halted their cell cycle (data not shown).

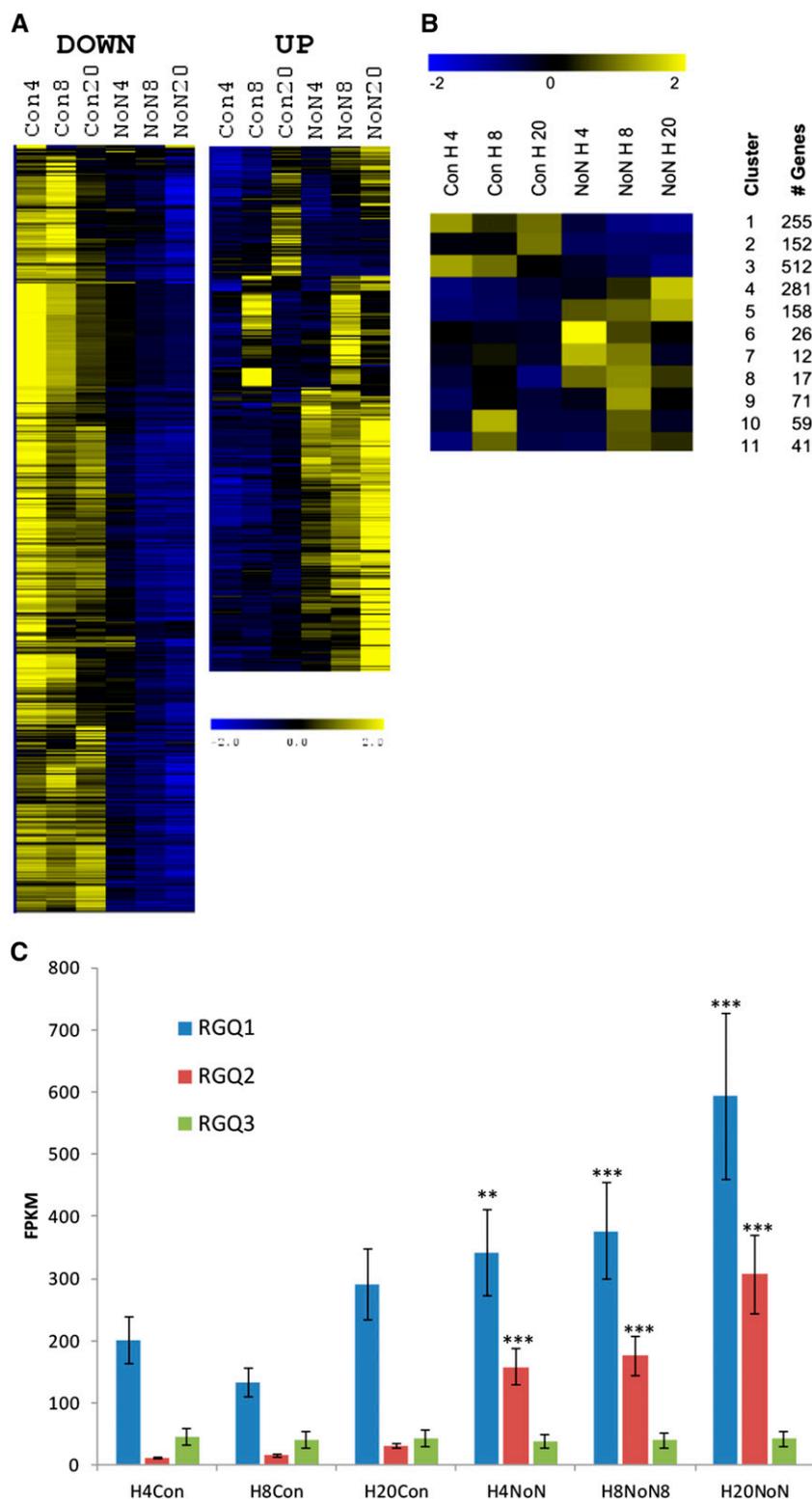
RNA-Seq was performed by Illumina HiSeq2000-based paired-end ( $2 \times 100$  bp) and indexed sequencing. In total, 4,770 genes showed statistically significant differential expression at some point during the experiment, of which 1,587 showed a log<sub>2</sub>-fold change of 1.5 or more (Supplemental Dataset S1). To ensure adequate stringency for the subsequent screens, only the latter genes were considered. These were grouped into 11 expression clusters, three consisting of down-regulated and eight consisting of up-regulated genes (Fig. 1, A and B). The down-regulated clusters contained many genes encoding proteins involved in photosynthesis, such as the oxygen enhancer complex, a photosystem stability factor, the chlorophyll biosynthesis machinery, and several fucoxanthin binding proteins. Furthermore, roughly half of all Calvin cycle genes annotated in the DiatomCyc Pathway/Genome database (Fabris et al., 2012) were down-regulated. Similarly, protein synthesis displayed an abrupt halt, as illustrated by the rapid decreased expression of most genes encoding ribosomal components and t-RNA charging molecules.

Conversely, nitrogen-starved cells increased their capacity to absorb and assimilate nitrogen, indicated by the up-regulation of nitrate transporters, nitrate, and nitrite reductases (expression maximum after 20 h of starvation, pattern of cluster 4 in Fig. 1B). Another way for cells to cope with reduced nitrogen availability is to increase turnover of nitrogen-containing metabolites (amino acids and nucleotides) and polymers, reflected by the up-regulation of numerous proteases, amidases, and aminotransferases (patterns of clusters 5 and 6 in Fig. 1B). Incorporation of the liberated ammonia may happen through the up-regulated Gln oxoglutarate aminotransferase cycle. Carbon recycling likely occurs through increased activity of the up-regulated tricarboxylic acid cycle and pyruvate dehydrogenase complex (early up-regulated and maximum expression after 20 h of starvation; pattern of cluster 5 in Fig. 1B).

Overall, these findings correlate well with those published previously and indicate that the transcriptional reprogramming of metabolism reported after 2 d or more of nitrogen starvation (Yang et al., 2013; Alipanah et al., 2015; Levitan et al., 2015) is already initiated very early in the nitrogen stress response.

### Two Motifs Are Overrepresented in Nitrogen Starvation Responsive Promoters

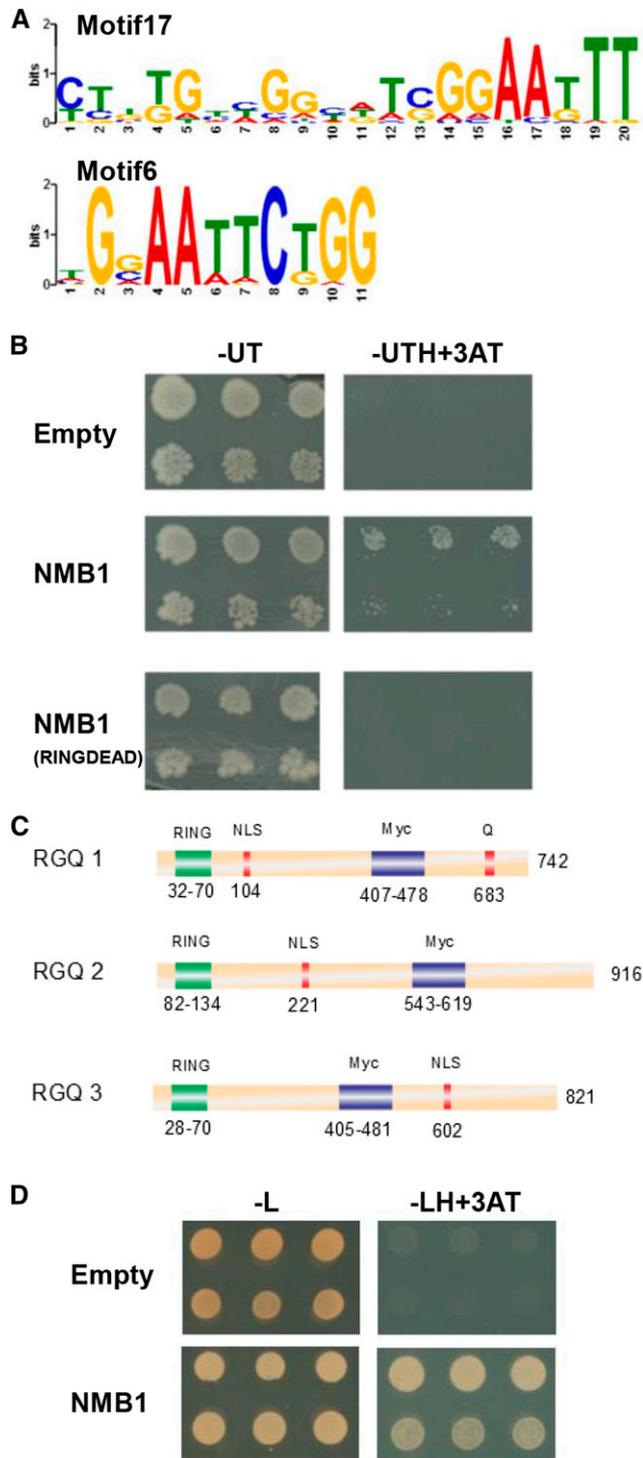
Of the 212 TFs predicted from the *P. tricornutum* genome (Rayko et al., 2010), 86 were up-regulated during nitrogen starvation (Supplemental Dataset S1). To provide a lead for functional analysis, we mined



**Figure 1.** Overview of the early nitrogen stress transcriptome of *P. tricornutum*. **A**, Overview of differentially expressed genes with a log<sub>2</sub>-fold change of 1.5 or more from the relevant control sample at the same time point. Shown are normalized FPKM values, at three time points in nutrient-replete (Con) or nitrogen-starved (NoN) cells, ordered using hierarchical clustering as implemented in the TMEV4 tool. **B**, Self-organizing tree of the same data. Cluster numbers and the number of genes in each cluster are indicated at the right. **C**, Expression pattern of *RGQ1*, *RGQ2*, and *RGQ3*, as reflected from the FPKM values in the RNA-Seq dataset (indicated in the y axis). Error bars represent standard deviations of three biological replicates. *P* values (\*\**P* < 0.01, \*\*\**P* < 0.001) were generated by Cuffdiff using the standard settings (Trapnell et al., 2013).

for overrepresented motifs in promoters of nitrogen starvation-responsive genes, and these were subsequently screened for TFs able to bind these DNA sequences. From the RNA-Seq dataset, 861 genes were selected with the highest fold induction during nitrogen starvation and a median expression higher than 20

fragments per kilobase per million (FPKM). The intergenic sequences upstream of the coding sequences were selected and used as input for the MEME Suite program (Bailey et al., 2009) to identify motifs in these promoters. The putative promoters were selected as being 500 bp upstream of the start codon. It should be noted that the



**Figure 2.** RGQ1 binds and transactivates Motif17. **A**, Overrepresented motifs in promoters of *P. tricornutum* genes up-regulated after nitrogen starvation. **B**, Y1H analysis showing that transactivation of the pMotif17::HIS3 selectable marker by RGQ1-GAL4AD only occurs when the RGQ1 RING domain is functional. Empty vectors were used as negative control. Transformed yeast cultures dropped in serial dilutions (10- and 100-fold) were grown for 6 d on control (–UT) and selective medium (–UTH+3AT). **C**, Domain organization of and position in the RGQ1, RGQ2, and RGQ3 proteins. Myc, bHLH-MYC\_N or GAF-like

*P. tricornutum* genome is gene dense, and the selected interval often overlapped with the promoter or untranslated region of the previous gene. These short intergenic spaces made statistical analysis and motif discovery difficult, as two genes with different expression profiles were often assigned overlapping intergenic sequences. Using a  $\chi^2$ -based approach, only the motifs that were significantly enriched in the test set were kept (Fig. 2A). Motifs that showed a high repetition were excluded. Using this approach, only two motifs were retained, Motif6 ( $P$  value  $6.32 \times 10^{-03}$ ) and Motif17 ( $P$  value  $6.07 \times 10^{-04}$ ), with Motif6 containing the TTC core typical of heat shock factor-type TFs (Åkerfelt et al., 2010).

#### Y1H Screening Identifies a RING-Domain-Containing Protein Capable of Binding Motif17

A custom-made *P. tricornutum* cDNA library (Huysman et al., 2013) was used to screen for proteins capable of binding the two identified motifs in a Y1H screen. No hits could be obtained with Motif6, but of the 13 colonies selected from the screen with Motif17, 10 contained the open reading frame of *Phatr50304*, which was subsequently renamed *RING-GAF-Gln-containing protein* (RGQ1) for the three domains it contains (see below). Retransformation of the Motif17 bait strain with a full-length RGQ1-GAL4AD plasmid, containing a fusion between RGQ1 and the GAL4 activating domain, confirmed the interaction (Fig. 2B).

Notably, the coding sequence of RGQ1 does not contain any known DNA-binding domain and was not listed in the TFs inferred from the genome sequence (Rayko et al., 2010). In contrast, the National Center for Biotechnology Information conserved domain database (Marchler-Bauer et al., 2015) and InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan5/>) searches indicated the presence of a RING/FYVE/PHD-type zinc finger domain (IPR013083) and a domain called bHLH-MYC\_N (PF14215) or GAF-like (IPR029016) in the RGQ1 amino acid sequence (Fig. 2C). The bHLH-MYC\_N domain is associated with several plant TFs, such as MYC2, but it does not contain the DNA-binding domain of these proteins. Instead, this sequence resembles a GAF domain, which is commonly associated with a sensor-type function and can sense a variety of signals, such as light, redox status, or a range of small molecules (Auldridge and Forest, 2011; Uden et al., 2013). Finally, the RGQ1 protein is also predicted to have a nuclear localization signal (Kosugi et al., 2009) and a C-terminal stretch of Gln residues that is commonly associated with transcriptional activation

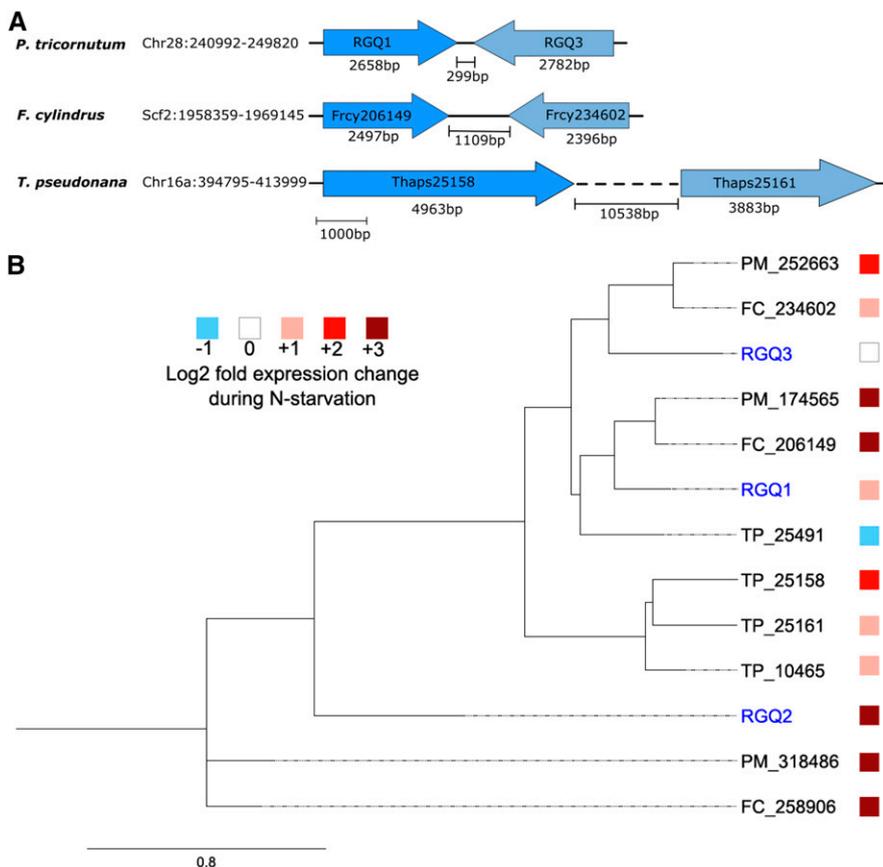
domain; NLS, nuclear localization signal; Q, Gln-rich stretch; RING, RING/FYVE/PHD-type Zinc finger. **D**, RGQ1 fused to the GAL4-DBD can transactivate the pUAS::HIS3 reporter gene in yeast. Transformed yeast cultures dropped in serial dilutions (10- and 100-fold) were grown for 6 d on control (–L) and selective medium (–LH+3AT).

(Courey and Tjian, 1988; Titz et al., 2006; Fig. 2C), thus fitting with a potential role as TF. To test this assumption, the RGQ1 sequence was fused to that of the GAL4 DNA-binding domain (GAL4DBD) and introduced into yeast. The RGQ1-GAL4DBD fusion protein was found to be able to drive transcription of a UAS::HIS3 selectable marker in *Saccharomyces cerevisiae* (Fig. 2D), confirming that RGQ1 can act as a transcriptional activator. Generally, on average 5% of all proteins seem to be capable of autoactivating *HIS* transcription after fusion with the GAL4 DNA-binding domain (Van Criekinge and Beyaert, 1999). No known activation domains are found in the RGQ sequences. However, this finding is not unexpected as activation domains can often not be identified on the basis of protein homology, and proteins rich in Gln, such as the RGQs, are known to cause transcriptional activation in eukaryotes (Titz et al., 2006). As outlined in the next section, these Glns are evolutionary conserved and likely essential for the functioning of the protein.

RING domains are usually found in E3 Ubiquitin (Ub)-ligases (Stone et al., 2005). The most conserved Cys and His residues in the putative RING domain were defined by aligning the amino acid sequence of *P. tricornutum* RGQ1 with those of homologs from other diatoms (Supplemental Fig. S1). Out of the 48 diatom

RGQ sequences, 40 contained the RING-like domain. It should be noted though that due to the origin of the data (see the "Results" section below), some of these amino acid sequences may be incomplete. RING-type proteins are usually classified according to the spacing of these residues. The consensus of the RGQ RING-like domain is C-X<sub>2</sub>-C-X<sub>5</sub>-D-X<sub>5</sub>-C-X<sub>3-4</sub>-C-X<sub>2</sub>-H-X-R-C, which did not fit any of the known plant E3 Ub-ligase classes (Stone et al., 2005). Especially striking is that even though half of the RGQ proteins have a His residue within two amino acids of the first conserved Cys, only the His adjacent to the fourth Cys is conserved over all RGQ sequences. Correspondingly, no auto-ubiquitination activity could be detected in in vitro assays with recombinant RGQ1 protein purified from *Escherichia coli* (data not shown), indicating RGQ1 may not be functioning as an Ub-ligase.

Other plant RING-type zinc finger proteins have been reported to possess DNA-binding activity mediated by the RING domain (Eklund et al., 2010; Malhotra and Sowdhamini, 2013). Therefore, a "RING-dead" version of the RGQ1-GAL4AD fusion was designed, in which conserved Cys and His residues in the RING domain (Supplemental Fig. S1) were mutated to Ala, and introduced into the yeast Motif17 bait strain. The RING-dead version of RGQ1 was not capable of restoring



**Figure 3.** The RGQ1 protein belongs to a small gene family that is conserved in diatoms. A, Scaled genome organization of the RGQ1/RGQ3 loci in *P. tricornutum*, *F. cylindrus* (Frcy), and *T. pseudonana* (Thaps). Genome positions are indicated at the left. B, Phylogenetic tree of RGQ homologs in four sequenced diatoms, generated with Phylogeny.fr with standard settings (Dereeper et al., 2008). Branches are proportional to branch support lengths. Expression data from this study and GEO project GSE56132 (<http://www.ncbi.nlm.nih.gov/geo/>) were mapped on the tree. All expression values were significant at  $P < 0.05$ . FC, *F. cylindrus*; PM, *P. multiseriis*; TP, *T. pseudonana*.

growth on selective medium, indicating that the RING domain is crucial for DNA binding (Fig. 2B).

#### RGQ1 Belongs to a Small Protein Family Conserved in Heterokonts

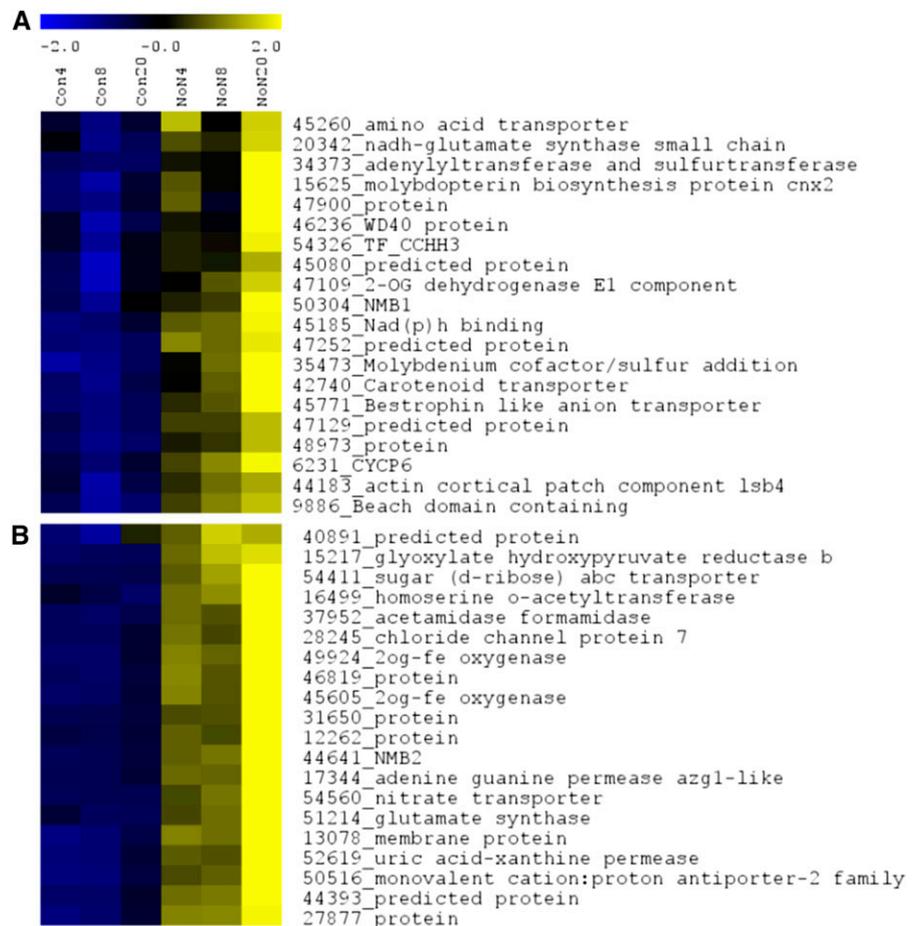
A BLASTp homology search showed that there are two other proteins with a high sequence homology present in *P. tricornutum*, Phatr44641 (RGQ2) and Phatr50305 (RGQ3), which have similar domain organization and 31% and 41% identity to RGQ1, respectively (Figs. 2C and 3; Supplemental Fig. S2; Supplemental Table S1). *RGQ1* and *RGQ3* lie adjacent to each other in the *P. tricornutum* genome and share the same terminator region (Fig. 3A).

Besides *P. tricornutum*, two other diatoms also have fully assembled and publicly available genome sequences, namely *T. pseudonana* (Armbrust et al., 2004) and *F. cylindrus* (Port et al., 2013). These two species contain four and three RGQ-like proteins, respectively. A near identical gene organization as for *P. tricornutum* *RGQ1/RGQ3* is also seen in *F. cylindrus*, whereas in *T. pseudonana*, two *RGQ* genes are located in a tandem repeat, separated by another gene coding for a putative branched chain amino acid ABC transporter (Thaps25159; Fig. 3A).

The sparse diatom genome data were supplemented by transcriptomes assembled by the Marine Microbial Eukaryote Transcriptome Sequencing Project (<http://marinemicroeukaryotes.org/>), which has sequenced and assembled mRNA from a wide variety of diatoms. The transcriptomes of 20 diatom species, covering pennate and centricates, were searched for RGQ-like sequences using BLASTp. After removing partial sequences, 48 RGQ-like proteins were identified. This number is likely to be an underestimate as most transcriptomes were de novo assembled using a limited set of RNA-Seq experiments carried out on diatoms grown under nitrogen-replete conditions. Under these conditions, *RGQ* transcripts are likely to be expressed at low levels and thus not represented in the RNA-Seq dataset at sufficient levels to allow assembly. The 48 sequences were grouped in a phylogenetic tree (Fig. 3B; Supplemental Fig. S2), which indicated that the *RGQ2* clade groups separately from and appears to be more divergent between pennate and centricate diatoms than the *RGQ1/RGQ3* clade.

Out of the 48 sequences investigated, the N-terminal region was recognized as a RING domain or the related PHD domain 11 times by the National Center for Biotechnology Information conserved domain search. However, the alignment (Supplemental Fig. S1) shows

**Figure 4.** The coexpression neighborhood of *RGQ1* and *RGQ2*. Coexpression neighborhood of *RGQ1* (A) and *RGQ2* (B) generated using the CummRbund Package and visualized using TMEV4. The color scale indicates normalized FPKM values.



high conservation in this region, with a similar spacing of His and Cys residues, which supports their involvement in coordinating zinc binding. A similar situation is obtained when scoring the presence of the GAF domain. Domain searches failed to detect the GAF domain in most diatom RGQ sequences, but the alignment again shows high conservation in this region (Supplemental Fig. S3).

Finally, most diatom RGQ-like proteins also contain unusually long stretches of Gln residues. While these regions are C terminal, their positions are not highly conserved across the aligned species. To determine whether this feature is common in diatom proteins, a random subset of 10,000 proteins was searched for a stretch of four Gln residues and compared to the RGQ sequences used to construct the phylogenetic tree (Supplemental Fig. S2; Supplemental Table S1). In the 10,000 random proteins, a stretch of four uninterrupted Glns occurred 232 times, while the 48 RGQ sequences contained it in 71 instances. These Gln stretches have been reported to be involved in transcriptional activation (Courey and Tjian, 1988).

RGQ-like genes were also found to be present in nondiatom stramenopile algae. For example, the multicellular brown alga *Ectocarpus siliculosus* has two similar genes (*Esi\_0137\_0082* and *Esi\_0368\_0011*), whereas the eustigmatophyte *Nannochloropsis gaditana* contains only one RGQ-like gene in its genome (*Naga\_100020g79*). However, the pelagophyte *Aureococcus anophagefferens* lacks RGQ genes. Likewise, oomycetes generally have several genes encoding proteins with a domain that resembles the N-terminal RING-like domain, but these are much shorter and lack the GAF domain and the Gln stretch. In green algae no significant BLAST hits are found for the RING-like domain or the full-length protein. Using the MYC-like domain as a query, only proteins that are more similar to the plant-like bHLH-type TF MYC2 could be found, while no orthologs of RGQ could be identified. This bioinformatics analysis indicates that the RGQ gene family is unique for stramenopiles and that it has expanded in diatoms.

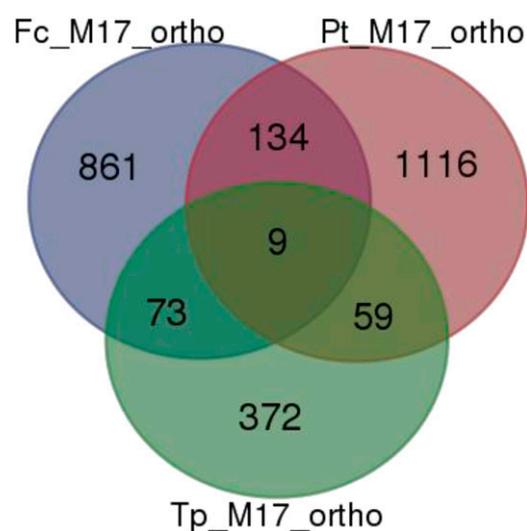
### RGQ1/2 Expression Is Induced by Nitrogen Starvation

Mining of the RNA-Seq data indicated that both *RGQ1* and *RGQ2*, but not *RGQ3*, are induced early during nitrogen starvation (Fig. 1C). *RGQ3* has an average expression of 24 FPKM, which is close to the median of expression of our entire dataset and is not induced at any point during the time course. Notably, *RGQ1* is also transcribed in exponentially growing cells in nutrient-replete conditions, whereas *RGQ2* has lower transcript levels in this condition but is induced much stronger upon nitrogen starvation, i.e. approximately 15-fold 4 h after the removal of nitrogen versus approximately 2-fold for *RGQ1* at that time point (Fig. 1C). The expression pattern was confirmed in an independent quantitative real-time PCR (qRT-PCR) experiment, in which 16 h after nitrogen removal from the

medium, *RGQ1* and *RGQ2* were 3- and 8-fold up-regulated, respectively (Supplemental Fig. S4).

Publicly available RNA-Seq data (Armbrust et al., 2004; <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56132>) indicated that RGQ expression was also induced by nitrogen starvation in other diatom species, both pennates and centricates (Fig. 3B). In other pennates, such as *F. cylindrus* and *P. multiseriis*, all RGQ gene copies appeared nitrogen starvation inducible, including the *RGQ3* homologs. In the centricate *T. pseudonana*, three of the RGQ homologs were induced by nitrogen starvation, whereas the closest homolog of *P. tricornutum* *RGQ1* was repressed by nitrogen starvation (Fig. 3B). This indicates that increased RGQ expression constitutes a conserved element in the nitrogen starvation response in diatoms.

To assess the potentially conserved role of RGQ proteins in the diatom nitrogen starvation response further, we assessed prevalence of Motif17 in the three diatom genomes with fully assembled genomes in more detail. In *P. tricornutum*, the frequency of Motif17 in the promoter regions was determined to be 16% or a total of 1,743 genes (Supplemental Dataset S2). To verify that Motif17 was not an artifact, its prevalence was also checked in a reshuffled promoterome maintaining the same dinucleotide frequency. In this random set, 1,417 sequences contained the motif, but only 190 genes were shared between both sets, indicating that it is likely not an artifact. Furthermore, as indicated above, Motif17 was enriched in nitrogen starvation-responsive promoters: about 25% of the genes up-regulated after nitrogen starvation and coregulated with *RGQ1* or *RGQ2* contain Motif17 in the first 500 bp upstream of their start codon. Mining of the genes with the highest expression correlation with *RGQ1* and *RGQ2* indicated that both cluster with genes encoding proteins involved



**Figure 5.** Prevalence of Motif17 in promoters of diatom genes. Venn diagram showing the overlap between orthologous genes from *P. tricornutum* (Pt), *T. pseudonana* (Tp), and *F. cylindrus* (Fc) containing Motif17 in their promoter (see Supplemental Dataset S2 for gene details).

in the assimilation of nitrogen and to a lesser degree genes involved in redox reactions (Fig. 4). Recently it was shown that these two processes were linked since many nitrogen assimilating proteins have oxidation-sensitive thiol groups (Rosenwasser et al., 2014). Furthermore, the *RGQ1* cluster contains two genes related to the biosynthesis of molybdopterin, a cofactor required for nitrate reductase function. In the *RGQ2* cluster, also several genes related with the synthesis of amino acids and transport of nitrogen-containing compounds were encountered.

Next, the prevalence of Motif17 was also checked in the intergenic sequences of the pennate diatom *F. cylindrus*. In total 4,367 genes contained the motif in their putative promoter (Supplemental Dataset S2), which corresponds to 16% of all gene models, comparable to what was found in *P. tricornutum*. Using reciprocal best BLAST hits as the selection criterion, 1,077 orthologous *P. tricornutum* genes could be identified for these genes, of which 143 also contain Motif17 in their promoter (Fig. 5). Among these genes is *Fracy206149*, the most likely candidate for an ortholog of *RGQ1* (Supplemental Dataset S2).

The same analysis was performed on the distantly related centric diatom *T. pseudonana*, in which 1,155 genes contain the motif in their putative promoter or approximately 10% of the genome (Supplemental Dataset S2). Retaining only orthologs with *P. tricornutum*, 68 *P. tricornutum* genes with Motif17 remain (Fig. 5). In contrast to *F. cylindrus*, no *RGQ* sequences are found in this list, but instead several genes essential for nitrogen assimilation are present. Examples are a Glu synthase (*Phatr51214/Thaps29861*), which is present in the *RGQ2* coexpression cluster (Fig. 4B), and one gene encoding a nitrate reductase (*Phatr54983/Thaps25299*), which groups in cluster 4 of the nitrogen starvation-inducible genes together with *RGQ1* (Fig. 1B). Together, our in silico expression and promoter analysis suggests that the role of the *RGQ* family is conserved throughout the pennate and centric diatom lineages.

## CONCLUSION

We propose that the *RGQ* proteins represent a hitherto unknown class of TFs that form an integral and conserved part of the nitrogen stress response in diatoms. Our findings also illustrate that the predicted number of diatom TFs is likely to be an underestimate because poorly sampled branches of the evolutionary tree often contain yet unknown families of DNA-binding proteins.

## MATERIALS AND METHODS

### Diatom Culturing

All experiments were carried out with the *Phaeodactylum tricornutum* (Pt1) Bohlin Strain 8.6 obtained from the diatom culture collection available at Ghent University. Cells were grown in 500-mL Erlenmeyer flasks with ESAW medium containing 7.5 mg sodium nitrate per liter and other nutrients, according to the

standard recipe (Berges et al., 2001). Flasks were placed on a shaking platform at 120 rpm, and light was provided using continuous overhead fluorescent lighting (18W/865; Osram) with an average lighting intensity of  $100 \mu\text{E m}^{-2} \text{s}^{-1}$  in a temperature controlled room at 21°C. For stress treatments, the precultured cells were diluted 2-fold daily to maintain exponential growth. Cells were harvested by centrifugation for 30 min at 6,000g and washed twice with a 20 g/L saline solution. This starter culture was split into equal parts and used to inoculate ESAW with and without added sodium nitrate.

### Expression Profiling

For each condition and time point, cells from three biological repeats were captured on a 3  $\mu\text{m}$  pore size PVDF membrane (Millipore) by filtering 20 mL of culture as rapidly as possible using a mild vacuum. Filters or algal pellets were transferred to 2-mL Eppendorf tubes and flash frozen in liquid nitrogen. For RNA extraction, the pellets were resuspended with 1.5 mL of TRI Reagent (Molecular Research) and processed according to the manufacturer's instructions up until the RNA precipitation step. The aqueous phase was mixed with an equal volume of 70% (v/v) ethanol and transferred to an RNeasy mini spin column (Qiagen). On column, DNase digest was performed according to the manufacturer's instructions using RQ1 DNase (Promega). RNA was eluted in RNase-free water, and quality control was performed on Nanodrop and 1% agarose gel. RNA-Seq was performed by Illumina HiSeq2000-based paired-end ( $2 \times 100$  bp) and indexed sequencing of the RNA-Seq libraries at Genomics Core UZ KULeuven (<http://gc.uzleuven.be/>).

For RNA-Seq data analysis, using a cutoff value of 20, FASTQ files were quality filtered, and adaptor sequences were removed using Cutadapt (Martin, 2011). Only the sequences that were still paired were retained for further analysis. Mapping and counting the reads was done with Tophat/Cufflinks/Cuffdiff using the GFF filtered models and genome files available on the Joint Genome Institute Web site (<http://genome.jgi-psf.org/>; Bowler et al., 2008; Trapnell et al., 2013). The expression of significantly expressed genes was log2 transformed and visualized using the TM4 MeV package (Saeed et al., 2006).

For qRT-PCR, cDNA was generated with iScript kit (Bio-Rad) from RNA extracted as described above. qRT-PCR was carried out with a Lightcycler 480 (Roche) and SYBR Green QPCR Master Mix (Stratagene). *PLA* (*Phatr3\_J16787*), *VTC4* (*Phatr3\_J29004*), and *RP3A* (*Phatr3\_J13566*) were used as reference genes with primers listed in Supplemental Table S2. Analysis was performed using the  $\Delta\Delta\text{Ct}$  method as described (Schmittgen and Livak, 2008).

### Molecular Cloning

Gateway technology was used for subcloning and the generation of expression clones according to the manufacturer's instructions (Invitrogen Life Technologies). Models were manually checked for completeness and correspondence with RNA-Seq coverage using the Integrative Genomics Viewer browser (<https://www.broadinstitute.org/igv/>). All destination vectors used were generated previously (Siaut et al., 2007), and amplification of *P. tricornutum* gene sequences was performed with Primestart DNA polymerase (TaKaRa Biosciences).

### Y1H Analysis

Y1H was performed as described (Deplancke et al., 2006) and using pDEST22-*RGQ1* transformed into yeast strain YM427152.

### Phylogenetic Analysis

The consensus coding sequences of nine diatoms were downloaded from the Marine Microbial Eukaryote Transcriptome Sequencing Project Web site (<http://marinemicroeukaryotes.org/>) and used to construct a local BLAST database. The tBLASTn program was used to find homologous sequences, using only the highest scoring hit for each sequence. Alignments were performed using MUSCLE as implemented in Geneious 8.1 (Biomatters; <http://assets.geneious.com/manual/8.1/GeneiousManual.html>) using standard settings. Visualization of the alignments was done with the same program. For tree construction, the default method was chosen on phylogeny.fr (Dereeper et al., 2008). Briefly, alignment was performed using MUSCLE 3.7 set at highest accuracy. Poorly aligned regions and gaps were removed using Gblocks v0.91b by eliminating all nonconserved positions longer than eight amino acids with a minimum remaining block length of 10, not allowing any gaps in the final alignment and ensuring at least 85% of the sequences present in any flanking regions. The maximum likelihood method was used as implemented in the

PhyML program v3.0 aRtL using the WAG substitution model assuming an estimated proportion of invariant sites and four gamma-distributed rate categories with an estimated parameter of 1.293. Branch reliability was tested using the aLRT test. Graphics were generated with FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) and Geneious.

## In Silico Promoter Analysis

The intergenic sequences between two *P. tricornutum* genes were extracted and trimmed to a maximum length of 500 bp. The input sequences were examined for possible motifs using the MEME Suite program (v4.9.1) with a maximum number of 20 motifs; other settings were left at default (Bailey et al., 2009). Using a custom script, a  $\chi^2$  analysis was performed on the obtained motif in the test set versus a random list of intergenic sequences of the same length. Motif occurrences were determined with the Find Individual Motif Occurrences tool on the MEME Suite (Bailey et al., 2009) and the original MEME motif file as input. The same procedure was followed for Motif17 prevalence in the other diatoms whose genomes were obtained from the Joint Genome Institute Web site (<http://genome.jgi-psf.org/>). For the occurrence of Motif17 in a random set, the intergenic sequences of *P. tricornutum* were reshuffled while maintaining dinucleotide frequency using the "shuffle" program of the HMMER package (v3.1b1; <http://hmmer.janelia.org/>) with default settings.

## Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession number E-MTAB-4096.

## Supplemental Data

The following supplemental materials are available.

**Supplemental Figure S1.** Protein alignment of the RING-like domain in the RGQ protein family.

**Supplemental Figure S2.** Phylogenetic tree of the RGQ protein.

**Supplemental Figure S3.** Protein alignment of the GAF domain in the RGQ protein family.

**Supplemental Figure S4.** *RGQ1* and *RGQ2* expression analysis by qRT-PCR.

**Supplemental Table S1.** *RGQ* genes identified in heterokont genomes.

**Supplemental Table S2.** Primers used for qRT-PCR in this study.

**Supplemental Dataset S1.** FPKM values of expressed *P. tricornutum* genes.

**Supplemental Dataset S2.** Prevalence of Motif17 in promoters of diatom genes.

## ACKNOWLEDGMENTS

We thank Dr. Gino Baart for helpful discussions, Sophie Carbonelle and Robin Vanden Bossche for technical assistance, and Annick Bleys for help in preparing the manuscript.

Received August 19, 2015; accepted November 13, 2015; published November 18, 2015.

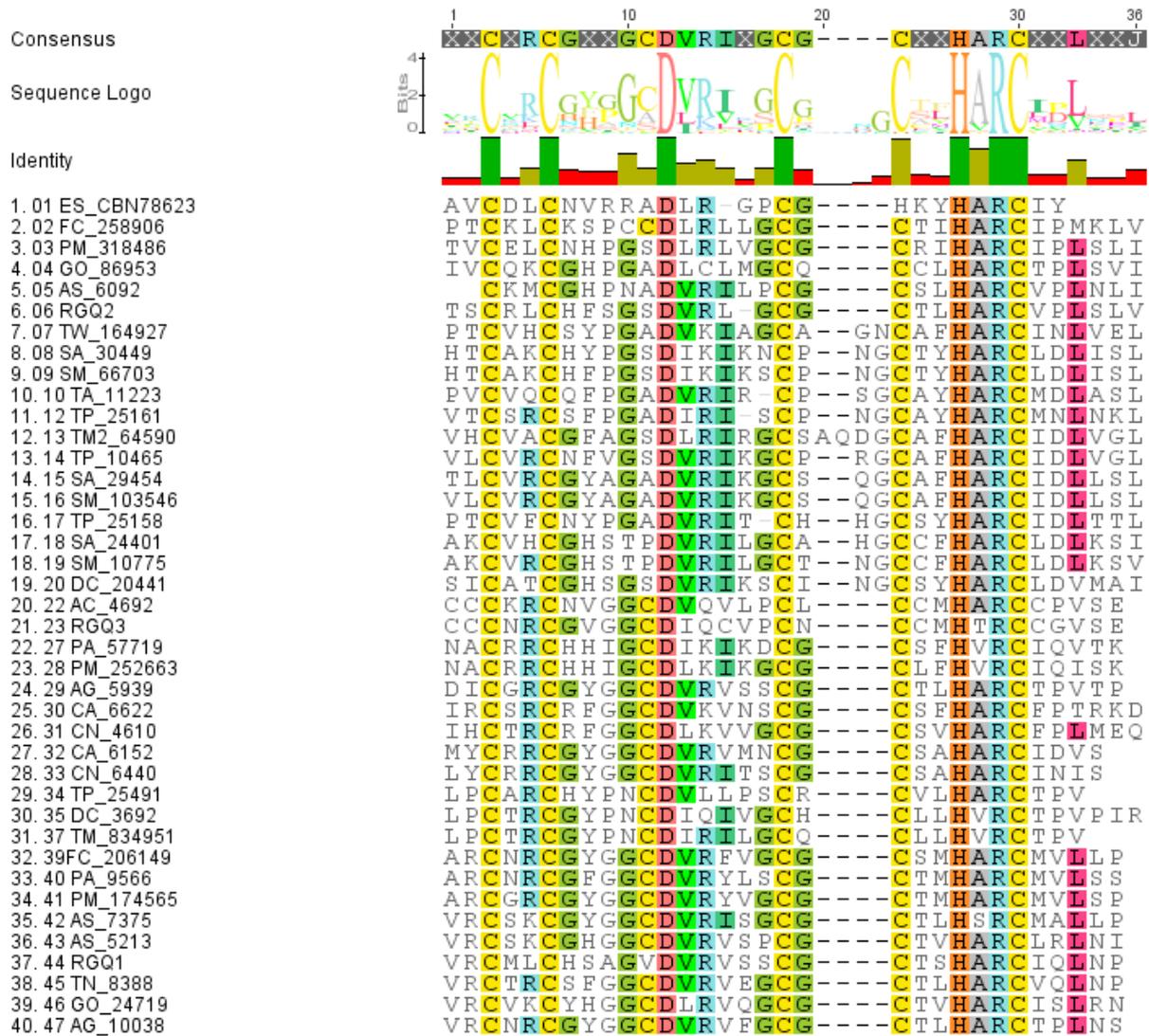
## LITERATURE CITED

- Åkerfelt M, Morimoto RI, Sistonen L (2010) Heat shock factors: integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol* **11**: 545–555
- Alipanah L, Rohloff J, Winge P, Bones AM, Brembu T (2015) Whole-cell response to nitrogen deprivation in the diatom *Phaeodactylum tricornutum*. *J Exp Bot* **66**: 6281–6296
- Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, et al (2011) Evolution and

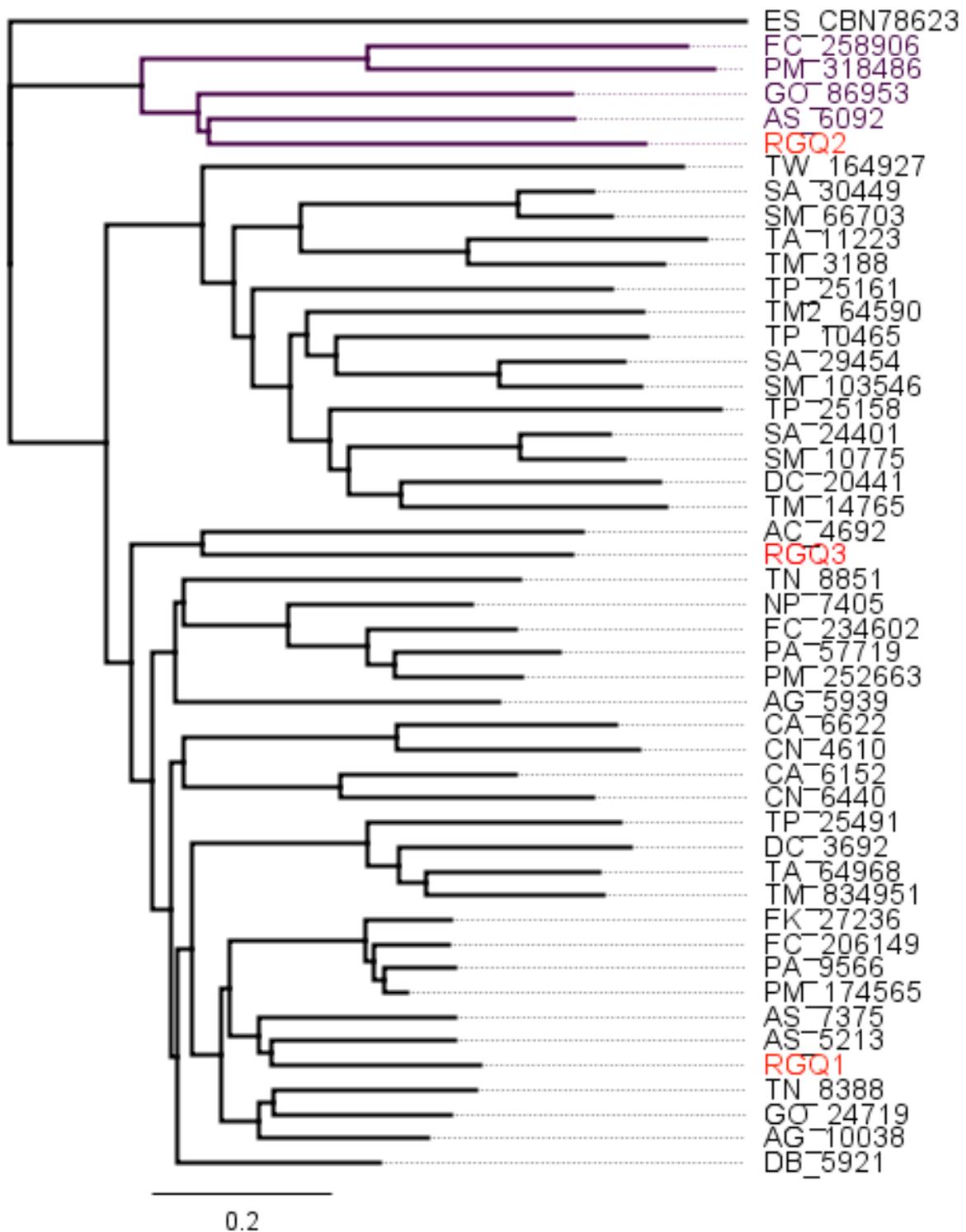
- metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**: 203–207
- Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* **459**: 185–192
- Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79–86
- Auldrige ME, Forest KT (2011) Bacterial phytochromes: more than meets the light. *Crit Rev Biochem Mol Biol* **46**: 67–88
- Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208
- Berges JA, Franklin DJ, Harrison PJ (2001) Evolution of an artificial seawater medium: improvements in enriched seawater, artificial water over the last two decades. *J Phycol* **37**: 1138–1145
- Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Ollilar RP, et al (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–244
- Courey AJ, Tjian R (1988) Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55**: 887–898
- Deplancke B, Vermeirssen V, Arda HE, Martinez NJ, Walkout AJM (2006) Gateway-compatible yeast one-hybrid screens. *Cold Spring Harb Protoc* **2006**: pdb.prot4590
- Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, et al (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: W465–W469
- Eklund DM, Ståldal V, Valsecchi I, Cierlik I, Eriksson C, Hiratsu K, Ohme-Takagi M, Sundström JF, Thelander M, Ezcurra I, et al (2010) The *Arabidopsis thaliana* STYLISH1 protein acts as a transcriptional activator regulating auxin biosynthesis. *Plant Cell* **22**: 349–363
- Elser JJ, Bracken MES, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, Ngai JT, Seabloom EW, Shurin JB, Smith JE (2007) Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol Lett* **10**: 1135–1142
- Fabris M, Matthijs M, Carbonelle S, Moses T, Pollier J, Dasseville R, Baart GJE, Vyverman W, Goossens A (2014) Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*. *New Phytol* **204**: 521–535
- Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE (2012) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J* **70**: 1004–1014
- Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR (2004) The evolution of modern eukaryotic phytoplankton. *Science* **305**: 354–360
- Ge F, Huang W, Chen Z, Zhang C, Xiong Q, Bowler C, Yang J, Xu J, Hu H (2014) Methylcrotonyl-CoA carboxylase regulates triacylglycerol accumulation in the model diatom *Phaeodactylum tricornutum*. *Plant Cell* **26**: 1681–1697
- Giordano M, Beardall J, Raven JA (2005) CO<sub>2</sub> concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu Rev Plant Biol* **56**: 99–131
- Hockin NL, Mock T, Mulholland F, Kopriva S, Malin G (2012) The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. *Plant Physiol* **158**: 299–312
- Huysman MJJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inzé D, Bowler C, Kroth PG, et al (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin *dsCYC2* controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*). *Plant Cell* **25**: 215–228
- Huysman MJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler C, Inzé D, Van de Peer Y, De Veylder L, Vyverman W (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol* **11**: R17
- Kosugi S, Hasebe M, Tomita M, Yanagawa H (2009) Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc Natl Acad Sci USA* **106**: 10171–10176
- Levitani O, Dinamarca J, Zelzion E, Lun DS, Guerra LT, Kim MK, Kim J, Van Mooy BAS, Bhattacharya D, Falkowski PG (2015) Remodeling of

- intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress. *Proc Natl Acad Sci USA* **112**: 412–417
- Lommer M, Specht M, Roy A-S, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, et al** (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol* **13**: R66
- Malhotra S, Sowdhamini R** (2013) Genome-wide survey of DNA-binding proteins in *Arabidopsis thaliana*: analysis of distribution and functions. *Nucleic Acids Res* **41**: 7212–7219
- Mangan S, Itzkovitz S, Zaslaver A, Alon U** (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *J Mol Biol* **356**: 1073–1081
- Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al** (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**: D222–D226
- Martin M** (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**: 10–12
- Moore JK, Doney SC, Glover DM, Fung IY** (2001) Iron cycling and nutrient-limitation patterns in surface waters of the World Ocean. *Deep Sea Res Part II Top Stud Oceanogr* **49**: 463–507
- Mühlroth A, Li K, Røkke G, Winge P, Olsen Y, Hohmann-Marriott MF, Vadstein O, Bones AM** (2013) Pathways of lipid metabolism in marine algae, co-expression network, bottlenecks and candidate genes for enhanced production of EPA and DHA in species of Chromista. *Mar Drugs* **11**: 4662–4697
- Ohno N, Inoue T, Yamashiki R, Nakajima K, Kitahara Y, Ishibashi M, Matsuda Y** (2012) CO<sub>2</sub>-cAMP-responsive cis-elements targeted by a transcription factor with CREB/ATF-like basic zipper domain in the marine diatom *Phaeodactylum tricornutum*. *Plant Physiol* **158**: 499–513
- Port JA, Parker MS, Kodner RB, Wallace JC, Armbrust EV, Faustman EM** (2013) Identification of G protein-coupled receptor signaling pathway proteins in marine diatoms using comparative genomics. *BMC Genomics* **14**: 503
- Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C** (2010) Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol* **188**: 52–66
- Rosenwasser S, Graff van Creveld S, Schatz D, Malitsky S, Tzfadia O, Aharoni A, Levin Y, Gabashvili A, Feldmesser E, Vardi A** (2014) Mapping the diatom redox-sensitive proteome provides insight into response to nitrogen stress in the marine environment. *Proc Natl Acad Sci USA* **111**: 2740–2745
- Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J** (2006) TM4 microarray software suite. *Methods Enzymol* **411**: 134–193
- Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C** (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* **406**: 23–35
- Schmittgen TD, Livak KJ** (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* **3**: 1101–1108
- Stone SL, Hauksdóttir H, Troy A, Herschleb J, Kraft E, Callis J** (2005) Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. *Plant Physiol* **137**: 13–30
- Tirichine L, Bowler C** (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J* **66**: 45–57
- Titz B, Thomas S, Rajagopala SV, Chiba T, Ito T, Uetz P** (2006) Transcriptional activators in yeast. *Nucleic Acids Res* **34**: 955–967
- Tozzi S, Schofield O, Falkowski P** (2004) Historical climate change and ocean turbulence as selective agents for two key phytoplankton functional groups. *Mar Ecol Prog Ser* **274**: 123–132
- Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L** (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46–53
- Unden G, Nilkens S, Singenstreu M** (2013) Bacterial sensor kinases using Fe-S cluster binding PAS or GAF domains for O<sub>2</sub> sensing. *Dalton Trans* **42**: 3082–3087
- Valenzuela J, Mazurie A, Carlson RP, Gerlach R, Cooksey KE, Peyton BM, Fields MW** (2012) Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum*. *Biotechnol Biofuels* **5**: 40
- Van Criekinge W, Beyaert R** (1999) Yeast two-hybrid: state of the art. *Biol Proced Online* **2**: 1–38
- Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al** (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431–1443
- Yang Z-K, Niu Y-F, Ma Y-H, Xue J, Zhang M-H, Yang W-D, Liu J-S, Lu S-H, Guan Y, Li H-Y** (2013) Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation. *Biotechnol Biofuels* **6**: 67

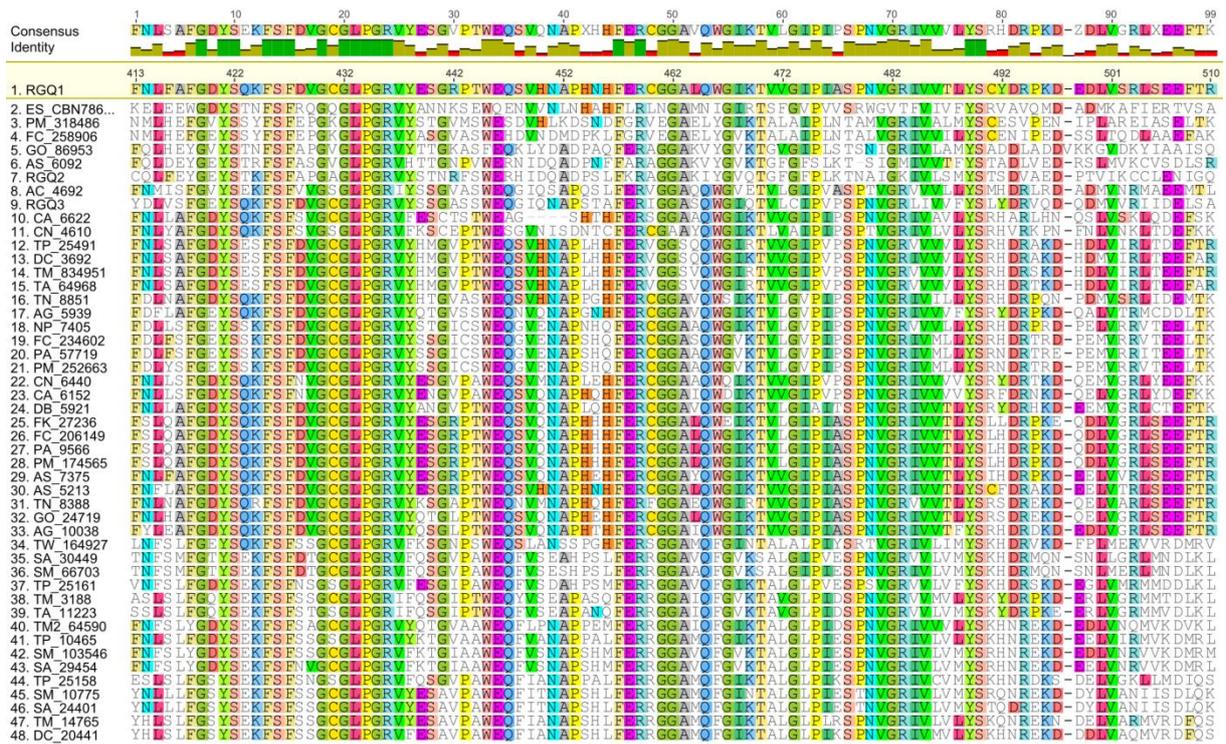
## SUPPLEMENTAL DATA



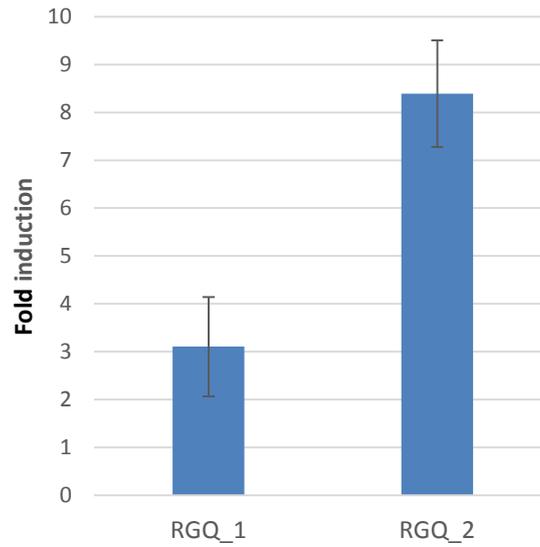
**Figure S1. Protein alignment of the RING like domain in the RGQ protein family.** Numbers represent MMETSP coding sequences (see Supplemental Table S1 for details). Residues mutated to alanine for the RING-dead version are boxed.



**Figure S2. Phylogenetic tree of RGQ protein.** Forty-seven diatom proteins homologous to RGQ were aligned together with one *E. silicosis* RGQ-like protein. Numbers represent MMETSP coding sequences (see Supplemental Table S1 for details). Amino acid identity is proportional to the indicated scale.



**Figure S3. Protein alignment of the GAF domain in the RGQ protein family.** Numbers represent MMETSP coding sequences (see Supplemental Table S1 for details).



**Figure S4. *RGQ1* and *RGQ2* expression analysis by qRT-PCR.** Cells growing in exponential phase were harvested, washed once and subsequently resuspended in  $\text{NaNO}_3$  free medium. Cells were harvested by centrifugation after 16 hours. Y-axis represents fold increase in *RGQ* transcript levels in nitrogen-starved relative to nitrogen-replete cells.

**Table S1. *RGQ* genes identified in heterokont genomes.**

*See separate .xlsx file*

**Table S2. Primers used for qRT-PCR in this study.** FW, forward primer; RV, reverse primer.

Name	Sequence	Target (Phatr number)
<b>Reference Genes</b>		
FW_VTC4	5' -GGCACATTTTCGGCTACGATT-3'	29004
RV_VTC4	5' -TATCCTCGTGTACGTGTCCG-3'	
FW_PUA	5' -CTGGATTGACAAACCCTGGC-3'	16787
RV_PUA	5' -TCCCCCTCACTTCCAAACTG-3'	
FW_RP3A	5' -AAAGAGCATGCCAAGTGGTG-3'	13566
RV_RP3A	5' -TCTACAGCTCGAATGTCCCC-3'	
<b>Target Genes</b>		
FW_Q_RGQ1	5' -TGCGTCTCCTTCTTCTTCGT-3'	50304
RV_Q_RGQ1	5' -CAACCCTGTGTTTGTCTGTTG-3'	
FW_Q_RGQ2	5' -TCCCGATCAAAGCAAATCTC-3'	44641
RV_Q_RGQ2	5' -CAAACCTCGTCCACAATGCTG-3'	

**Supplemental Dataset 1. FPKM values of expressed *P. tricornutum* genes.** FPKM values were grouped per time point for nutrient-replete (CON) or nitrogen-starved (NoN) cells. The log<sub>2</sub>-fold change between the two different conditions at a particular time point and the p-value indicating statistically significant differential expression, as determined using default Cuffdiff settings, are included.

*See separate .xlsx file*

**Supplemental Dataset 2. Prevalence of Motif17 in promoters of diatom genes.** The different sheets indicate prevalence of Motif17 in the intergenic sequences upstream of the coding sequences in *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, and *Fragilariopsis cylindrus*, respectively. The last sheet indicates prevalence of Motif17 in randomized versions of the *P. tricornutum* promoter sequences. The putative promoters were selected as being 500 bp upstream of the start codon. The different columns indicate sequence names, start position, stop position, strand, score, p-value, q-value, and matched sequence, respectively. The *T. pseudonana* and *F. cylindrus* sheets display an additional column with the orthologous gene from *P. tricornutum* identified by reciprocal BLAST. #N/A indicates that no unambiguous orthologue could be identified assigned.

*See separate .xlsx file*