PhaeoNet: A Holistic RNAseq-Based Portrait of Transcriptional Coordination in the Model Diatom Phaeodactylum tricornutum

Ouardia Ait-Mohamed, Anna M G Novák Vanclová, Nathalie Joli, Yue Liang, Xue Zhao, Auguste Genovesio, Leila Tirichine, Chris Bowler, Richard G Dorrell

frontiers in Plant Science

ORIGINAL RESEARCH
published: 16 October 2020
doi: 10.3389/fpls.2020.590949

PhaeoNet: A Holistic RNAseq-Based Portrait of Transcriptional Coordination in the Model Diatom *Phaeodactylum tricornutum*

Ouardia Ait-Mohamed[1†], Anna M. G. Novák Vanclová[1‡], Nathalie Joli[1‡], Yue Liang[2], Xue Zhao[1,3], Auguste Genovesio[1], Leila Tirichine[1,3], Chris Bowler[1]* and Richard G. Dorrell[1]*

[1] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France, [2] Department of Oceanography, Dalhousie University, Halifax, NS, Canada, [3] Université de Nantes, CNRS, UFIP, UMR 6286, Nantes, France

Transcriptional coordination is a fundamental component of prokaryotic and eukaryotic cell biology, underpinning the cell cycle, physiological transitions, and facilitating holistic responses to environmental stress, but its overall dynamics in eukaryotic algae remain poorly understood. Better understanding of transcriptional partitioning may provide key insights into the primary metabolism pathways of eukaryotic algae, which frequently depend on intricate metabolic associations between the chloroplasts and mitochondria that are not found in plants. Here, we exploit 187 publically available RNAseq datasets generated under varying nitrogen, iron and phosphate growth conditions to understand the co-regulatory principles underpinning transcription in the model diatom *Phaeodactylum tricornutum*. Using WGCNA (Weighted Gene Correlation Network Analysis), we identify 28 merged modules of co-expressed genes in the *P. tricornutum* genome, which show high connectivity and correlate well with previous microarray-based surveys of gene co-regulation in this species. We use combined functional, subcellular localization and evolutionary annotations to reveal the fundamental principles underpinning the transcriptional co-regulation of genes implicated in *P. tricornutum* chloroplast and mitochondrial metabolism, as well as the functions of diverse transcription factors underpinning this co-regulation. The resource is publically available as PhaeoNet, an advanced tool to understand diatom gene co-regulation.

Keywords: stramenopile, transcriptomics, sigma factors, aureochromes, epigenetics, chloroplast-mitochondria

OPEN ACCESS

Edited by:
Justine Marchand, Le Mans Université, France

Reviewed by:
Pannaga Pavan Jutur, International Centre for Genetic Engineering and Biotechnology, India
Yoshihisa Hirakawa, University of Tsukuba, Japan

***Correspondence:**
Leila Tirichine
Leila.Tirichine@univ-nantes.fr
Chris Bowler
cbowler@biologie.ens.fr

[†]***Present address:***
Ouardia Ait-Mohamed, Immunity and Cancer Department, Institut Curie, PSL Research University, INSERM U932, Paris, France

[‡]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Marine and Freshwater Plants, a section of the journal Frontiers in Plant Science

Received: *03 August 2020*
Accepted: *15 September 2020*
Published: *16 October 2020*

Citation:
Ait-Mohamed O, Novák Vanclová AMG, Joli N, Liang Y, Zhao X, Genovesio A, Tirichine L, Bowler C and Dorrell RG (2020) PhaeoNet: A Holistic RNAseq-Based Portrait of Transcriptional Coordination in the Model Diatom Phaeodactylum tricornutum. Front. Plant Sci. 11:590949. doi: 10.3389/fpls.2020.590949

INTRODUCTION

The biology of prokaryotic and eukaryotic cells is dependent on elaborate metabolic, regulatory and gene expression pathways, consisting of multiple interacting components. The successful operation of these pathways depend on the coordinated expression of genes that underpin them, which allow the stoichiometric assembly of their constituent components and enable discrete and

appropriate regulatory responses to changes in physiological conditions (Gasch and Eisen, 2002; Teichmann and Babu, 2002). In prokaryotes and bacteria-derived genomes (e.g., "plastids" including chloroplasts and mitochondria) gene co-regulation is possible via the co-transcription of linked genes as part of the same transcriptional operon (Teichmann and Babu, 2002). In contrast, gene order plays a limited role in eukaryotic nuclear gene co-expression (Michalak, 2008), which depends instead on the simultaneous transcription, or transcriptional stabilization, of multiple discrete genomic loci. This may occur through common transcription factors (Teichmann and Babu, 2002; Reja et al., 2015); epigenetic modifications based around characteristic histone and DNA marks (Bird, 2002; Bártová et al., 2008); co-ordinated transcript processing events (Norbury, 2010); and the activity of small and long regulatory non-coding RNAs (Tsai et al., 2010; Kim and Sung, 2012).

The degree to which gene co-regulation is shared between different species is debated, with different studies identifying shared co-regulatory trends in between 8% (Teichmann and Babu, 2002) and 70% (Snel et al., 2004) of orthologous gene pairs between *Saccharomyces cerevisiae* (yeast) and *Caenorhabditis elegans* (nematode) genomes. Nonetheless, there is substantial merit to understanding gene co-regulation patterns in novel species. Since their origins over two billion years ago, the eukaryotes have radiated into a diverse range of different lineages, many of which are unicellular; and distantly related to model organisms in the animals, fungi and plants, with different underlying cellular biology and transcriptional dynamics (Walker et al., 2011). Identifying what gene co-regulation processes occur in microbial eukaryotes may allow us to better understand the biology underpinning the base of planetary food webs; and to better model the robustness of eukaryotic communities to environmental change.

The diatoms are a major group of predominantly marine algae, believed to be responsible for nearly one-fifth of total planetary photosynthesis (Field et al., 1998). Diatoms sit within the stramenopile supergroup, and are distantly related to animals, fungi and plants. Photosynthetic members of the stramenopiles, including diatoms, possess a chloroplast acquired through secondary endosymbiosis, unlike the primary plant chloroplast, which is of primary endosymbiotic origin (Walker et al., 2011). Previous genomic and functional investigations of model diatoms, for example *Phaeodactylum tricornutum*, have identified divergent features in their cellular biology, compared to more well-understood eukaryotic groups (Bowler et al., 2010). These include intricate metabolic connections between the diatom chloroplast, mitochondria and cytoplasm (Prihoda et al., 2012; Bailleul et al., 2015); and a wide range of different histone structural modifications (Veluchamy et al., 2013, 2015), many of which have not yet been detected in more established eukaryotic models.

Previously, microarray data from over 100 different conditions, including illumination regimes and pollutant stress (Osborn and Hook, 2013; Valle et al., 2014), have been generated from *P. tricornutum*; which have been assembled into a searchable interface, DiatomPortal that divides the *P. tricornutum* genome into 500 co-regulated gene clusters (Ashworth et al., 2016). Alongside this, a suite of RNA sequencing libraries exploring cellular responses to phosphorus, iron and nitrogen limitation have now been generated (Maheswari et al., 2010; Cruz de Carvalho et al., 2016; Smith et al., 2016; McCarthy et al., 2017) and inspecting these data may allow more precise integration of quantitative differences in transcript abundance than would be possible through microarray analyses. Furthermore, co-expression networks are a powerful tool for functional prediction and annotation of unknown genes in the absence of prior knowledge, which is the case for a significant number of genes in *P. tricornutum* (Rastogi et al., 2018). Co-expression networks can furthermore enrich our understanding of the more sparse co-expression networks generated for other marine algal species with secondary chloroplasts (principally, the distantly related diatom *Thalassiosira pseudonana*, the distantly related stramenopile *Nannochloropsis oceanica* and the haptophyte *Emiliania huxleyi*; Ashworth et al., 2016; Ashworth and Ralph, 2018).

Here, we use a tool of gene co-expression network analysis, WGCNA (Weighted Gene Correlation Network Analysis (Langfelder and Horvath, 2008; Guidi et al., 2016), to build PhaeoNet, an advanced tool for transcriptional understanding of the *P. tricornutum* genome. PhaeoNet is composed of 28 co-regulated gene modules, each with different expression dynamics. Considering the repartition of genes within these modules; functional, epigenetic and localization information from the third version annotation of the *P. tricornutum* genome (Phatr3; Rastogi et al., 2018); and annotated lists of diatom transcription factors (Rayko et al., 2010), we identify core features underpinning the transcriptional partitioning of diatom primary metabolism, including probable metabolic links between the diatom mitochondria and chloroplast; and dissect the diverse ranges of different transcriptional drivers of this co-regulation, notably in the case of chloroplast-targeted sigma factors. The raw data underpinning PhaeoNet have been made publically accessible via https://osf.io/42xmp.

MATERIALS AND METHODS

Dataset Curation and Abundance Calculations

A total of 187 publically available RNA-seq datasets from *P. tricornutum*, generated from three studies exploring, respectively, phosphorus (Cruz de Carvalho et al., 2016), iron (Smith et al., 2016) and nitrogen (McCarthy et al., 2017) stress transcriptional responses, were collected from the sequence read archive (SRA)[1] (Wheeler et al., 2006). The 182 libraries that passed through quality control steps, were included in the final version of the WGCNA performed, are named per their names respective studies in **Supplementary Table 1**, sheet 1. Data provided in the phosphate and nitrogen conditions were obtained using an Illumina Genome Analyzer (Bentley et al., 2008), while the iron study used SOLiD technology sequencing (Morey et al., 2013). *P. tricornutum* transcript IDs from each study were mapped to gene models based on the

[1]http://www.ncbi.nlm.nih.gov/Traces/sra/

Phatr3 annotation of the genome (**Supplementary Table 1**, Sheet 2; Rastogi et al., 2018).

Raw data were reprocessed using FastQC version v0.11.5[2]. Low quality reads (Phred quality score below 20) were filtered-out using trim-galore version 0.5.0[3]. The remaining sequences were aligned to the reference genome with the software package STAR version 2.5.3a (Dobin et al., 2013) (STAR –outFilterMismatchNmax 2 –outFilterMultimapNmax 1000 –alignIntronMin 20 –alignIntronMax 2000). The iron data derived from the SOLiD technology were first mapped using the Life Technologies LifeScope software suitable for data from such technology. For homogeneity purposes, the reads were remapped using the pre-cited version of STAR.

Expression levels of individual genes were obtained using featureCounts version 1.6.1 (Liao et al., 2014). Quality checks of datasets were performed using methods provided in DESeq2 version 1.19.37 (Love et al., 2014), with a PCA projection and a hierarchical dendrogram using Spearman correlation between library-normalized gene counts (Glasser and Winter, 1961). These subsequent analyses and results visualizations were performed using R package version 3.4.4.

Weighted Gene Correlation Network Analysis (WGCNA) and Network Visualization

The WGCNA R package (Langfelder and Horvath, 2008) was used to identify network modules from library-normalized gene expression values. First, a signed adjacency matrix (accepting oppositely correlated gene expression values to be clustered in the same modules) was obtained by calculating the pairwise Bi-weight mid-correlation coefficient from rij (Langfelder and Horvath, 2008), that represent expression values of genes i and j. A connectivity measure (k) per gene set was calculated by summing the connection strengths with other gene sets. Subsequently, the weighted adjacency matrix was obtained by raising the absolute value of the pairwise gene expression correlations to the soft-thresholding parameter β (Zhao et al., 2010). This achieved the scale-free topology criterion for WGCNA and typical for biological networks, emphasizing high correlations and minoring low ones, in which most nodes are not connected and only a few nodes are highly connected (Barabási, 2009).

The scale-free topology of PhaeoNet was evaluated by the Scale-Free Topology Fitting Index (R^2), which was the square of the correlation between log[p(k)] and log(k). A β coefficient of 12 with R^2 of 0.9 was used during the network building from the signed weighted adjacency matrix. The weighted adjacency matrix was finally used to calculate the Topological Overlap Matrix (TOM). Subsequently, modules were detected on the basis of the Topological Overlap measure using the following parameters: minModuleSize = 40 and mergeCutHeight = 0.25.

Graphical representations of the network were performed using Cytoscape (Shannon et al., 2003). All code used for the construction of PhaeoNet and interactive diagrams of each merged module are publically available through the following link: https://osf.io/42xmp.

Biological Interpretation of Merged Modules

The distribution of *P. tricornutum* genes in each transcriptional module was compared to the distribution of orthologous gene models (Phatr2.0 genome annotation) in microarray-derived transcriptional clusters generated as part of the DiatomPortal project (Ashworth et al., 2016). Only gene models that showed a one-to-one gene mapping (i.e., gene models that were neither split or merged, but including gene models that were truncated or extended) between version 2 (Phatr2) and version 3 (Phatr3) annotations of the *P. tricornutum* genome (Bowler et al., 2008; Rastogi et al., 2018) were considered.

Biological functions within the merged modules were identified using gene functional annotations from the Phatr3 annotation of the *P. tricornutum* genome (Bowler et al., 2008; Rastogi et al., 2018). These included: GO terms, using the R package TopGO (Aibar et al., 2015); PFAM domains and biological processes (Rastogi et al., 2018); probable evolutionary affinities inferred by BLAST top hit analyses (Rastogi et al., 2018); histone and DNA modifications associated with cells grown in replete media (Veluchamy et al., 2013, 2015); Polycomb group protein marks (Zhao et al., 2020); and KEGG orthology predictions, obtained with BLASTkoala, Kofamkoala and GHOSTkoala servers (Moriya et al., 2007; Kanehisa, 2017; Aramaki et al., 2019; Kanehisa and Sato, 2020). *In silico* targeting predictions were performed for all N-complete protein sequences (i.e., protein sequences inferred to start in a methionine) within the dataset, using HECTAR (Gschloessl et al., 2008); ASAFind v2.0 (Gruber et al., 2015), in conjunction with SignalP v3.0 (Bendtsen et al., 2004); MitoFates, with a threshold detection value of 0.35 (Fukasawa et al., 2015; Dorrell et al., 2017); and WolfPSort, taking the consensus best-scoring prediction using animal, fungi and plant reference datasets (Horton et al., 2007). Enrichments in each category were analyzed both qualitatively/manually and by a simple pivot table and chi-squared test. Tabulated lists of all annotations are presented in **Supplementary Table 2**.

Core chloroplast and mitochondria-associated functions were assembled from a list of 524 KEGG ortholog numbers based on previously identified chloroplast and mitochondria functions in photosynthetic eukaryotes (Dorrell et al., 2017; Nonoyama et al., 2019; Novák Vanclová et al., 2020). Where multiple candidate proteins were detected, proteins were assigned to either the chloroplast, mitochondria, or dual chloroplast/mitochondria (Gile et al., 2015; Dorrell et al., 2017) based on *in silico* targeting predictions. Where no clear targeting predictions could be obtained, proteins were identified based on BLAST similarity to orthologous chloroplast- or mitochondria-targeted proteins from other algal and stramenopile species (Dorrell et al., 2017; Río Bártulos et al., 2018). Disregarding 135 query proteins coded by organellar genomes in diatoms (Yu et al., 2018) and 17 query proteins encoded by nuclear genes with no PhaeoNet module

[2]https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3]http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

assigned, the final set comprised of 372 unique proteins targeted to the chloroplast and/or mitochondrion, encoded by nuclear genes that belong to one of the 28 merged modules. The main metabolic pathways and complexes and quantitative pathway associations, are presented in **Supplementary Table 3**.

A complete list of *P. tricornutum* transcription factors (TF) was assembled from a previous dataset (Rayko et al., 2010) and an updated list specifically of aureochromes (Banerjee et al., 2016), which were mapped to the version 3 genome annotation by BLASTp analysis. A total of 188 candidates, from 18 TF families (HSF, Myb, Zn_finger_C2H2, bZIP, Zn_finger_CCCH, bHLH, Sigma-70, Zn_finger_TAZ, CBF/NF, E2F-DP, CSF, Aureochrome, TRF, CCAAT-binding, AP2-EREBP, TAF9, CXC, Homeobox) corresponded to genes assigned to a PhaeoNet merged module (**Figure 5** and **Supplementary Table 4**). Given that the regulation of gene expression by transcription factors play a key role in the growth and progression of the cell cycle, the distribution within merged modules genes implicated in the cell cycle (cyclins) and in light perception events (e.g., phytochrome, cryptochrome) were additionally investigated, as well as genes implied in transcription and histone-related processes (**Figure 5** and **Supplementary Table 4**; Huysman et al., 2013; Annunziata et al., 2019).

Phylogenetics

A tree of sigma factor proteins from *P. tricornutum* and orthologous diatom and non-diatom sequences was constructed using a pipeline adapted from previous studies (Dorrell et al., 2017; Rastogi et al., 2018). Briefly, the complete peptide sequences of each sigma factor protein (eight total) in the Phatr3 annotation of the *P. tricornutum* genome (Rastogi et al., 2018) were searched using BLASTp against a composite library consisting of 110 diatom genomes, MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project, Keeling et al., 2014) and independent transcriptomes; and a reference set of 59 additional eukaryotic genomes, sampled from across the tree of life (**Supplementary Table 5**). Orthologs with an *e*-value of 10^{-05} or lower were extracted and searched against the complete protein sequences encoded within the Phatr3 annotation of the *P. tricornutum* genome via reciprocal BLASTp searches. Sequences which retrieved a single best hit against a *P. tricornutum* sigma factor protein were aligned using MAFFT v 8.0 (Katoh et al., 2002) under the –auto setting (BLOSUM62 matrix, gap open penalty 1.53, offset value 0) and the in-house alignment program in GeneIOUS v 10.0.9 (Kearse et al., 2012) using a more stringent set of conditions (65% similarity cost matrix, gap open penalty 12, gap extension penalty 3, two rounds of refinement). Poorly aligned or incomplete sequences were removed at each step. The 771 protein sequences retained were manually curated to retain a representative series of 86 diatom and non-diatom sequences related to each *P. tricornutum* sigma factor and trimmed using trimal with the –gt 0.5 setting (Capella-Gutiérrez et al., 2009) to yield a 453 aa alignment. The best-scoring tree topology was inferred from the alignment using RAxML v8.2, 100 bootstrap replicates and the PROTGAMMAJTT substitution model (Stamatakis, 2014); and MrBayes v3.2.7 over 600,000 generations, burnin fractions of 0.5 and the Jones amino acid substitution model (Huelsenbeck and Ronquist, 2001). Alignment and tree outputs are provided in **Supplementary Table 5**.

RESULTS AND DISCUSSION

Construction of an Optimized WGCNA Gene Expression Dataset for *P. tricornutum*

We harnessed 187 publically available RNAseq datasets derived from diverse physiological conditions and genotypes (Cruz de Carvalho et al., 2016; Smith et al., 2016; McCarthy et al., 2017) to build an integrative model of gene co-regulation for the model diatom species *P. tricornutum* (**Figure 1A** and **Supplementary Table 1**, sheet 1). We chose to build a dataset focusing on one species only, as even closely related diatom species may contain very different protein orthogroups (Parks et al., 2018; Sato et al., 2020) and even orthologous proteins may perform different physiological functions between different diatom species, with presumably different co-regulatory dynamics (Lampe et al., 2018). *P. tricornutum* was selected as a model system for this study as vastly greater amounts of gene expression data have been generated for this species than any other marine alga (Ashworth and Ralph, 2018); and as its genome annotation (currently in third version form and verified by comparison to over forty RNAseq libraries generated under varied conditions, Rastogi et al., 2018) is arguably the most complete of any alga known, allowing unprecedented insight into protein diversity, including variant protein forms generated by alternative splicing, protein sub-cellular localization and epigenetic modifications. The use of RNAseq data for this analysis allows us to advance on previous (e.g., microarray-based, Ashworth et al., 2016) analyses by allowing us to consider absolute rather than relative changes in expression levels between different datasets, and therefore exclude distorting effects caused by low absolute levels of the expression of specific genes in the *P. tricornutum* genome.

We optimized our data through several key pre-processing steps, for example removing batch effects (**Supplementary Figure 1A**) and five samples showing strong outlier effects (exemplar shown in **Supplementary Figure 1B**) prior to network construction, retaining 182 datasets for the final network construction. We also excluded genes that were found to be lowly expressed (median expression < 10 reads) in all inspected conditions, retaining 10,650/12,177 genes in the Phatr3 annotation (Rastogi et al., 2018) of the *P. tricornutum* genome (**Supplementary Table 1**, sheet 2). All pairwise gene correlations were calculated and then converted into connectivity strengths by raising their values to the power $\beta = 12$ for PhaeoNet. This power makes it possible to work in a scale-free condition and to avoid weak correlations (**Supplementary Figure 2**).

By applying the dynamic tree cut function on the dendrogram obtained by a hierarchical clustering with the method average, we identified 50 WGCNA modules with similar connection force profiles (**Figures 1B,C**). This was reduced to a subset of 28 merged modules with internal correlations above 0.75 (**Figure 1D**, **Supplementary Table 1**,

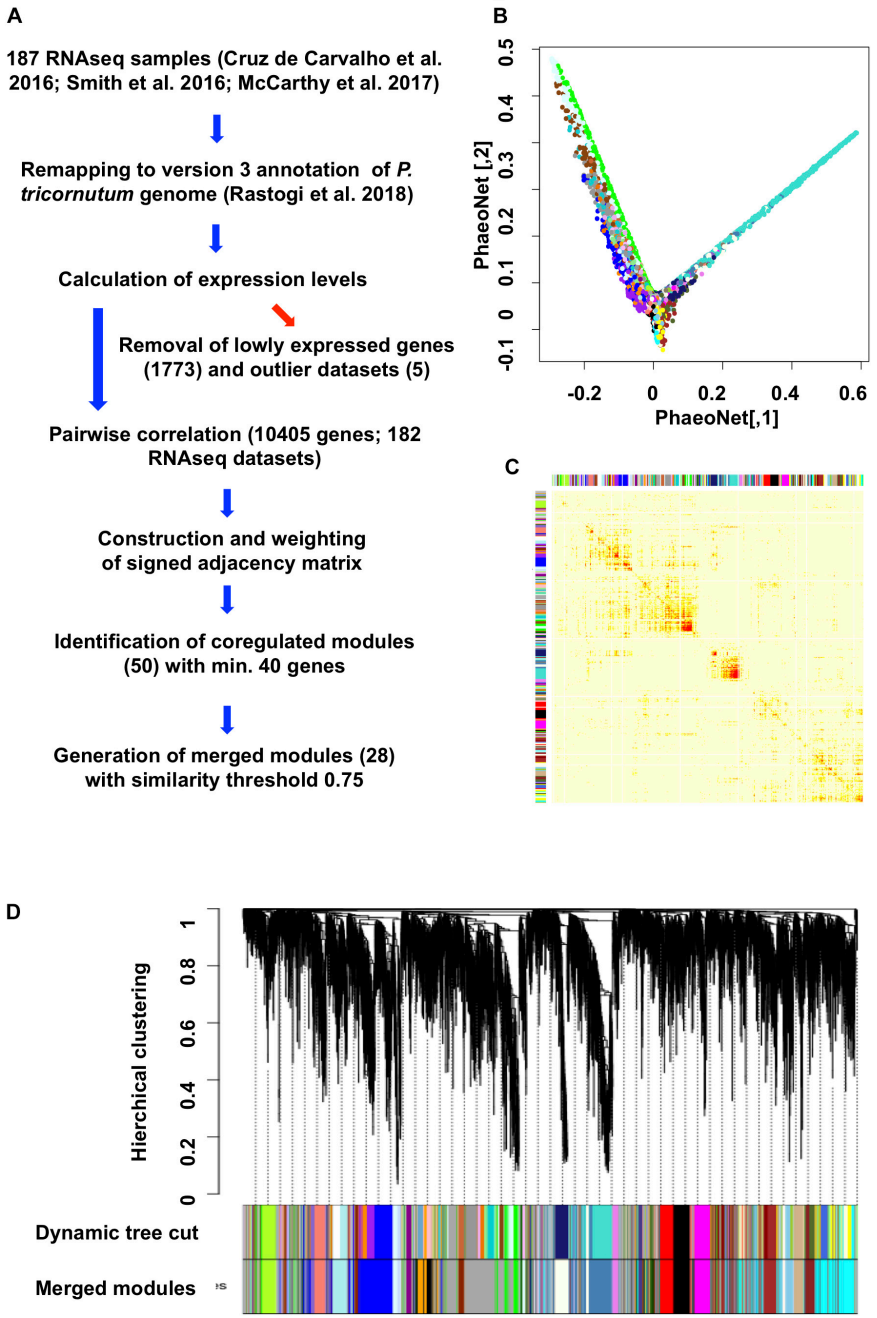

FIGURE 1 | Construction and topology of PhaeoNet. **(A)** Workflow diagram of the steps performed to construct PhaeoNet. **(B)** A Multi-Dimensional Scaling (MDS) plot of PhaeoNet. The dots correspond to genes and the colors correspond to the WGCNA modules. The tips of the plot correspond to hub genes of PhaeoNet. **(C)** Heatmap plot showing the TOM supplemented by the WGCNA module colors prior to merging. **(D)** Gene dendrogram of all incorporated PhaeoNet genes, obtained by average linkage hierarchical clustering. The first color row underneath the dendrogram shows the WGCNA module assignment obtained by the Dynamic Tree Cut method. The bottom color row shows the merged modules based on a correlation threshold of 0.75.

sheet 2; and **Supplementary Figure 3**), in accordance with other WGCNA studies (Langfelder and Horvath, 2008; Zhao et al., 2010) and following validation by cross-referencing to independently derived gene co-regulation datasets for *P. tricornutum* (described below). The final version of PhaeoNet showed good overall cohesion within the merged modules, as inferrable by multi-dimensional-scaling projection (**Figure 1C**) and correlation heatmaps of gene co-expression interconnectedness (**Figure 1B**).

We present an exemplar merged module output (paleturquoise) in **Figure 2**. A density heatmap, divided vertically by condition and horizontally by gene expression

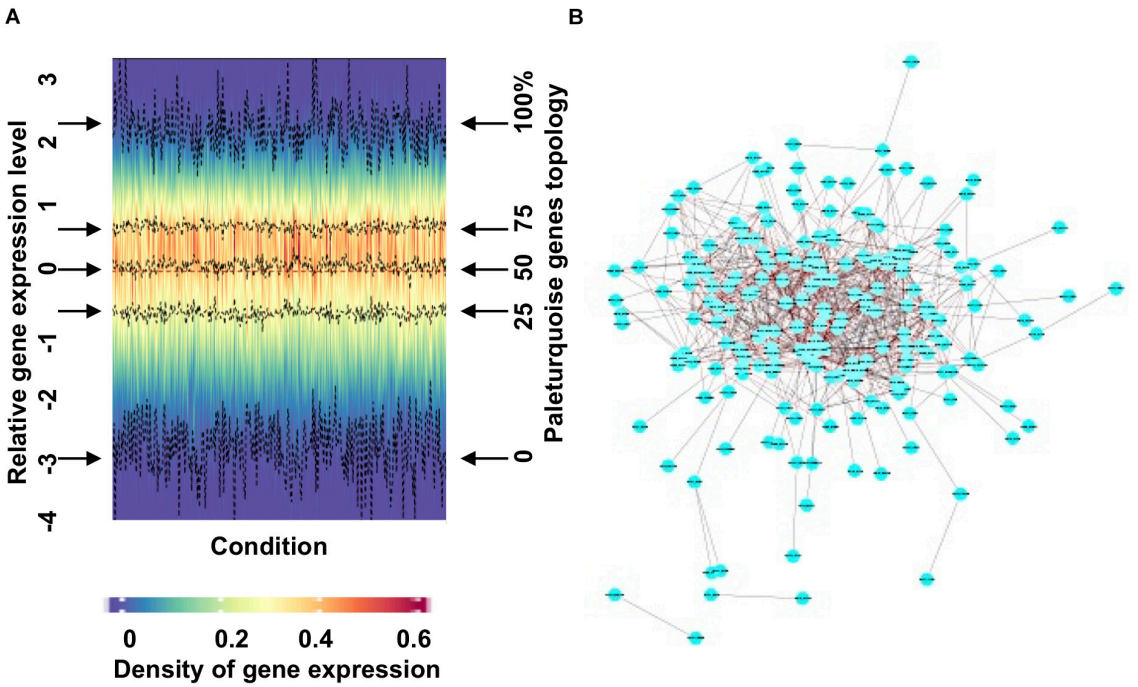

FIGURE 2 | Visualization and analysis of an exemplar PhaeoNet merged module (paleturquoise). **(A)** Density heatmap of all genes assigned to the paleturquoise merged module. The y-coordinate positions in the graph relate to density distribution of gene expression in each sample (shown on the left *y*-axis); the four middle dashed lines (indicated by horizontal arrows on either side of the graph) correspond to the median, first and third quantiles (shown on the right *y*-axis). The majority of the genes in this specific module show limited variation in expression profiles over different conditions samples. **(B)** A topological representation of connectedness within the paleturquoise merged module, visualized with Cytoscape (version 3.6.1). Each node represents a gene. Edges represent pairwise correlations between genes. The network shows all the paleturquoise module genes with a correlation value over a threshold of 0.20.

profile, shows a cohesive module as illustrated by stable values of first quantile, median, and third quantile values (**Figure 2A**) and is defined by high levels of expression across the majority of the conditions explored (**Supplementary Figure 3B**). Cytoscape (Shannon et al., 2003) visualization of the network with a correlation threshold of 0.2 (**Figure 2B**) demonstrates that the paleturquoise merged module is highly connected, showing a cluster of hub genes with high connectivity located in the central part of the network and only a small number of genes with limited connectivity. We provide detailed expression and Cytoscape data for each PhaeoNet merged module via https://osf.io/42xmp.

PhaeoNet Merged Modules Show Concordance With Microarray Co-regulation Data

We tested the reproducibility of our assignations, which may be considered as an independent measure of their robustness, by comparing the repartition of all *P. tricornutum* genes assigned to a PhaeoNet merged module with their corresponding distributions in 500 co-regulated clusters previously assembled from microarray data within the DiatomPortal server (Ashworth et al., 2016; **Supplementary Figure 4A** and **Supplementary Table 2**). Across 7,751 assessable genes with both PhaeoNet and DiatomPortal assignations, we identified 4,127 (53%) that occurred in the same PhaeoNet merged module as another gene with the same DiatomPortal cluster assignation; and 2,751 genes (35%) that occurred within the single PhaeoNet merged module incorporating the greatest number of genes from the same DiatomPortal cluster. Both of these frequencies were judged to be significantly greater than expected through a random distribution ($P = 0$, one-tailed chi-squared test), suggesting strong concordance between both datasets.

From the 461 (83%) DiatomPortal clusters for which we could identify corresponding PhaeoNet merged modules, 369 (80%) were preferentially distributed in one PhaeoNet merged module only, with the greatest number of clusters associated with the darkgray merged module (79 clusters), blue (46 clusters) and cyan (44 clusters) merged modules, reflecting the greater size of each merged module (**Supplementary Figure 4A** and **Supplementary Table 2**). No DiatomPortal clusters were found to be incorporated preferentially into the bisque4, darkmagenta, greenyellow, gray, lightsteelblue1 and mediumpurple3 PhaeoNet merged modules. It is possible that these merged modules represent transcriptional networks not visualized within DiatomPortal due to the different source datasets, generated using different techniques (e.g., microarray versus RNAseq data, assembled with hierarchical clustering versus WGCNA; Ashworth et al., 2016), which may influence what genes are inferred to be coexpressed using each analysis.

We also verified the number of associations independently found between pairs of genes in DiatomPortal clusters and PhaeoNet modules generated with independent merging thresholds, as an independent test of the appropriateness of our selected 0.75 merging threshold (**Supplementary Figure 4B**). We found greater concordance between DiatomPortal and PhaeoNet modules generated with a 0.75 merging threshold, as in our methodology, than in unmerged WGCNA modules, or modules merged with higher (0.8) or lower (0.7) threshold values (**Supplementary Figure 4B**).

Different PhaeoNet Merged Modules Perform Different Biological Activities in the *P. tricornutum* Cell

Next, we profiled the predominant biological activities associated with each merged module by calculating enrichment scores for different functional, subcellular targeting and evolutionary annotations across the *P. tricornutum* genome (Rastogi et al., 2018; **Figure 3** and **Supplementary Figure 5**). A full set of protein annotations *P. tricornutum*, including PhaeoNet module assignations, inferred functions, predicted localization and inferred evolutionary origin, is provided for user exploration in **Supplementary Table 2**.

We identified seven major subsets of merged modules with different biological properties. The first subset consists of merged modules (blue, lightcyan1, lightsteelblue1 and salmon) associated with the chloroplast [either genes encoding chloroplast-targeted proteins, inferred with ASAFind (Gruber et al., 2015) or HECTAR (Gschloessl et al., 2008), or of inferred red algal origin in a previous BLAST top hit analysis of the *P. tricornutum* genome (Rastogi et al., 2018)]. We included proteins of red algal origin as an independent estimator of chloroplastic origin, as the vast majority of red algal protein in *P. tricornutum* likely derive from the diatom chloroplast endosymbiont (Dorrell et al., 2017) and to allow us to detect chloroplast-associated proteins that elude *in silico* targeting prediction (Nonoyama et al., 2019; Schober et al., 2019). These merged modules were also enriched (as inferred with KEGG analysis (Kanehisa, 2017) in genes related to photosynthesis, carbon-fixation and core biosynthetic pathways (e.g., amino acid and pigment biosynthesis) associated with diatom chloroplasts (**Figure 3** and **Supplementary Figure 5**; Nonoyama et al., 2019). Nearly all of the merged modules within this subset were enriched in activating histone marks (e.g., H3K9Ac and H3K14Ac) and depleted in repressive marks (e.g., H3K9me2 and H3K27me3) in cultures grown under replete media conditions (**Figure 3** and **Supplementary Figure 5**; Veluchamy et al., 2015; Zhao et al., 2020), consistent with high levels of expression. Each of the chloroplast-enriched modules contained enrichments in different KEGG functions (discussed below), although only one of these modules (blue) was enriched in proteins containing at least one KEGG annotation (**Supplementary Figure 5**); and, in any case, all merged modules contain substantial numbers (between 18%, bisque4; and 54%, violet).

A second parallel set of merged modules (floralwhite, magenta, mediumpurple3, and orangered4), which was also found to be enriched in activating histone marks, was enriched in genes encoding mitochondria-targeted proteins (inferred with MitoFates, HECTAR and WolfPSort (Horton et al., 2007; Gschloessl et al., 2008; Fukasawa et al., 2015) and mitochondria-associated functions (e.g., oxidative phosphorylation and pyruvate metabolism; **Figure 3** and **Supplementary Figure 5**). Of note, the paleturquoise merged module was uniquely enriched in genes encoding both chloroplast and mitochondria-targeted proteins, suggesting a probable hub between both organelle functions (**Figure 3**).

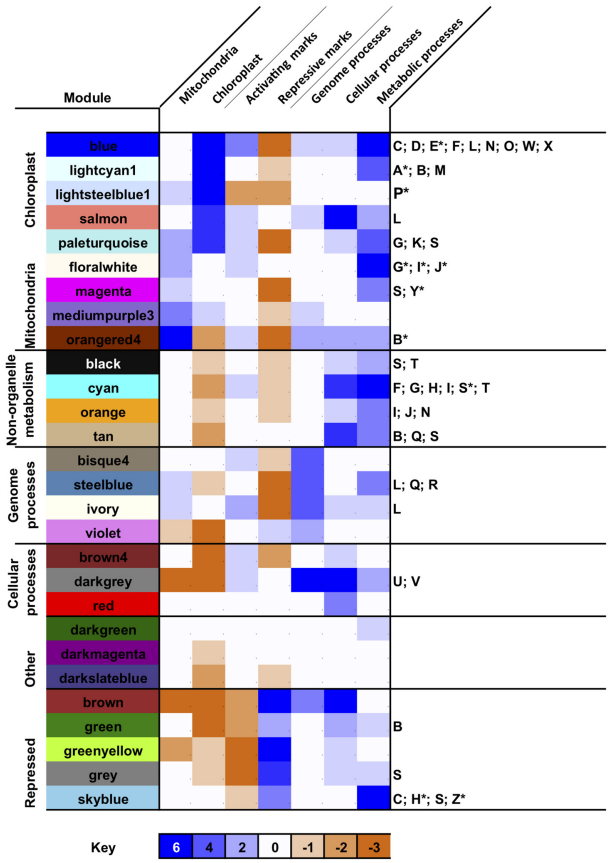

FIGURE 3 | Biological properties associated with PhaeoNet merged modules. This Figure provides an overview of enrichments of different organelle targeting (Horton et al., 2007; Gschloessl et al., 2008; Fukasawa et al., 2015; Gruber et al., 2015), epigenetic (Veluchamy et al., 2013, 2015; Zhao et al., 2020), evolutionary (Rastogi et al., 2018) and KEGG pathway annotations (Kanehisa, 2017) enriched in merged modules. The first seven (shaded) columns provide a score for different conditions, aggregated from chi-squared P-values of multiple enrichment predictors (defined beneath): enrichments in each condition carry a score of +1 if significant to $P < 0.05$ and +2 if significant to $P < 10^{-05}$; and depletions in each condition carry a score of –1 if significant to $P < 0.05$ and –2 if significant to $P < 10^{-05}$, assessed by chi-squared test against a null hypothesis of a random distribution of these features across all genes assigned to a PhaeoNet merged module. The final column lists all metabolic pathways enriched to $P < 0.05$, or $P < 10^{-05}$ (asterisked) for each merged module, assessed by chi-squared test as above. Verbose outputs for each set of conditions are provided in **Supplementary Figure 5**. Additional annotations, e.g., enrichments in inferred evolutionary origins of each merged module, are provided for user exploration in **Supplementary Table S2**.

We identified three further subsets of merged modules that were enriched in cytoplasmic or nuclear processes involved in metabolism (black, cyan, orange, and tan); genome-associated processes including transcription, translation and genome repair (bisque4, steelblue, ivory and violet); or cellular processes including protein modification, protein trafficking and the cell cycle (brown4, darkgray, and red; **Figure 3** and **Supplementary Figure 5**). Certain merged modules contained a mixture of genes encoding both metabolic and non-metabolic proteins: amongst other examples, the steelblue merged module was found to be enriched both in genes encoding proteins associated with ribosome and tRNA biogenesis and also in genes encoding enzymes involved in purine and pyrimidine metabolism, suggesting a probable transcriptional coordination of nucleotide biosynthesis to translational activity in *P. tricornutum* cells (**Figure 3** and **Supplementary Figure 5**). A sixth subset of merged modules (darkgreen, darkmagenta and darkslateblue) showed no obvious enrichment in any KEGG function or organelle localization, except for a possible enrichment in peroxisomal functions in the darkgreen merged module (Davis et al., 2017).

The final merged module subset (brown, green, greenyellow, gray, and skyblue) was uniquely enriched in repressive histone marks and depleted in activating histone marks, in cultures grown on replete media (**Figure 3**; Veluchamy et al., 2015). These merged modules may either be constitutively repressed in *P. tricornutum* cells, or might lose their repressive histone marks and be expressed in alternative conditions to the replete culture conditions in which the epigenetic datasets were collected (Zhao et al., 2020). We noted that the greenyellow merged module, for example, was enriched in proteins with at least one KEGG annotation; and the skyblue merged module was found to be significantly enriched in genes encoding proteins involved in carbon fixation, the TCA cycle and propionate metabolism (**Figure 3** and **Supplementary Figure 5**). Further studies of the epigenetic marks associated with these modules, including under physiological conditions in which they are most highly expressed (**Supplementary Figure 3**) will be necessary to determine under what circumstances the genes they contain make significant contributions to *P. tricornutum* biology.

PhaeoNet Merged Modules Reveal Transcriptional Co-regulation in *P. tricornutum* Chloroplast and Mitochondrial Metabolism

Having noticed specific biases in the distribution of mitochondria- and chloroplast-targeted proteins within our dataset and given the distinctive organelle metabolism noted in diatoms compared to plants (Kroth et al., 2008; Nonoyama et al., 2019; Smith et al., 2019), we wished to identify which key chloroplast and mitochondrial functions are revealed by PhaeoNet to be transcriptionally coordinated with one another. We searched the distribution of 372 manually curated nuclear-encoded proteins with known chloroplast- and mitochondria-associated functions and localizations (**Figure 4**, **Supplementary Figure 6**, and **Supplementary Table 3**). At least one gene encoding one such protein of each merged module was present in this set, however, only 12 merged modules contained more than 10 genes and amounted to 83% of the set.

The most abundantly represented merged module (blue, 69 chloroplast or mitochondrial occurrences) was clearly associated with genes encoding chloroplast anabolic reactions, containing enzymes associated with the Calvin–Benson–Bassham (CBB) cycle, chloroplast-targeted glycolysis/gluconeogenesis (Kroth et al., 2008) and fatty acid synthesis (Maréchal and Lupette, 2020), along with theta-class carbonic anhydrases that mediate biophysical carbon concentrating mechanisms in diatom chloroplasts (**Figure 4** and **Supplementary Figure 6**; Kikutani et al., 2016; Nonoyama et al., 2019). The blue merged module additionally contained genes encoding chloroplast-targeted proteins implicated in photoprotection, including the diatom xanthophyll cycle (e.g., Phatr3_J51703 encoding violaxanthin de-epoxidase; Frommolt et al., 2008; Dautermann and Lohr, 2017), tocopherol synthesis (e.g., Phatr3_J20470, encoding tocopherol cyclase; Dłużewska et al., 2016; Nonoyama et al., 2019) and two genes (Phatr3_J27278 and Phatr3_J44733) encoding LhcX-class chlorophyll-binding proteins, associated with high- and low-light adaptation responses in diatoms (**Supplementary Tables 2**, **3**; Taddei et al., 2016; Buck et al., 2019).

Genes encoding photosynthetic metabolism enzymes were concentrated in the lightcyan1 (41 occurrences) and lightsteelblue1 merged modules (19 occurrences). The lightcyan1 merged module included genes encoding LhcF-, LhcR-, and chlorophyll a/b-binding proteins, which are typically considered not to be involved in light stress responses (Gundermann et al., 2013; Büchel, 2015) and nucleus-encoded subunits of photosystems I, II and cytochrome c_6 (Grouneva et al., 2011; Roncel et al., 2016); whereas the lightsteeblue1 merged module contained the majority of genes involved in diatom chlorophyll and isoprenoid synthesis (Bertrand, 2010; Cihlar et al., 2016). We noted the presence of two genes encoding enzymes involved in pigment biosynthesis, respectively carotenoids (Phatr3_J21829, encoding 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase) and chlorophyll (Phatr3_J30690, encoding 3,8-divinyl protochlorophyllide-a 8-vinyl reductase (Wang et al., 2010) in the lightcyan1 merged module (**Supplementary Figure 6**). We also noted the presence of the gene Phatr3_J47674 encoding the iron stress-induced protein ISIP3 within the lightcyan1 merged module, which may point to a functional role for this protein in chloroplast photosystem assembly (**Supplementary Figure 6**; Allen et al., 2008; Chappell et al., 2015).

Genes encoding mitochondrial respiratory chain proteins were concentrated toward the orangered4 merged module (27 occurrences), whereas, genes encoding TCA cycle enzymes were concentrated toward the cyan merged module (34 occurrences). The orangered4 merged module also contained large numbers of genes encoding mitochondrial ribosomal proteins, which may relate to redox-state dependent regulation of mitochondrial biogenesis pathways (Allen, 2003). In contrast, most genes encoding chloroplast biogenesis-related proteins were identified in separate PhaeoNet merged modules to genes encoding proteins of the photosystem core, with significant enrichments

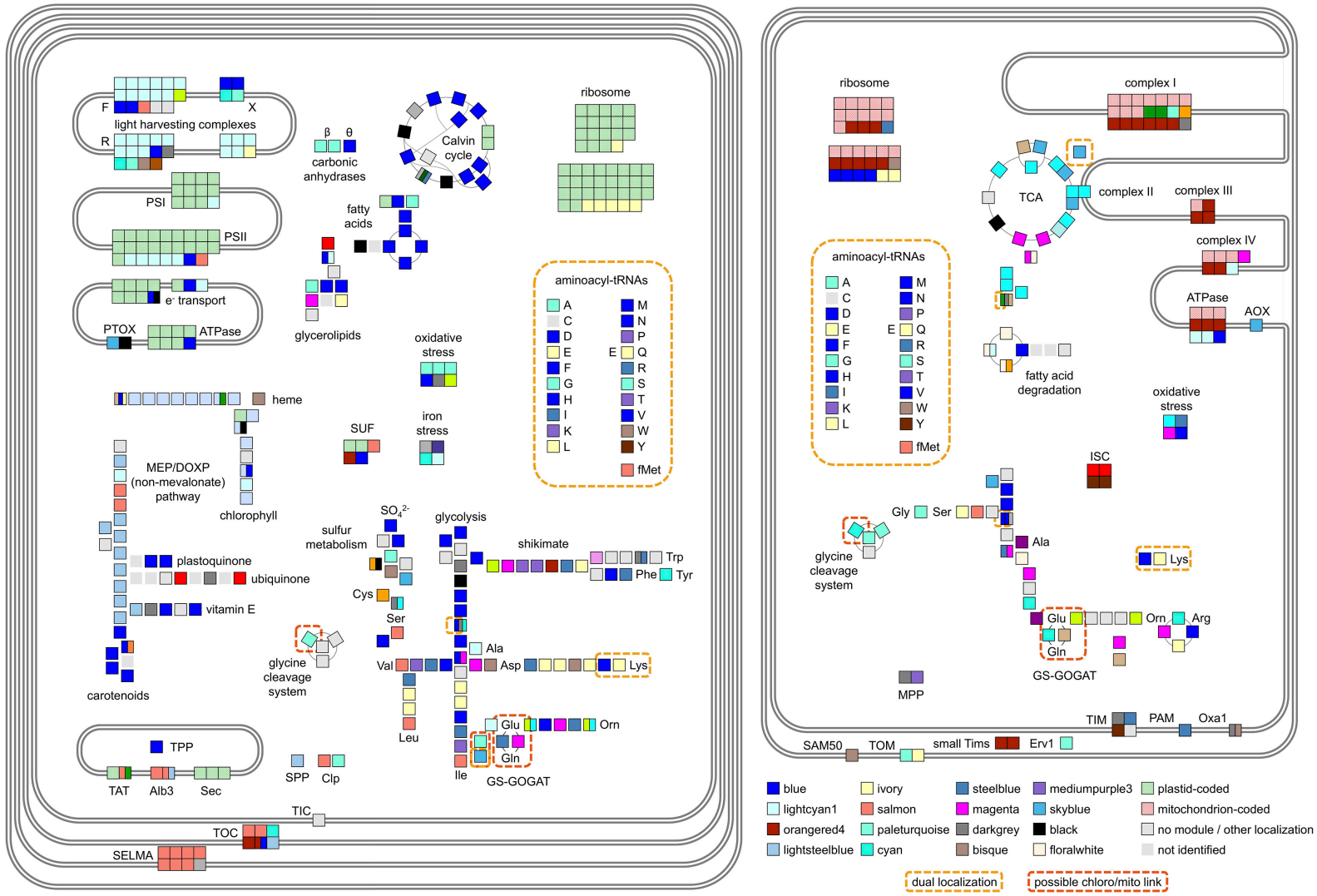

FIGURE 4 | Main metabolic pathways and functional complexes of *P. tricornutum* plastid **(left)** and mitochondrion **(right)** and their composition in regard to PhaeoNet merged modules. Each square represents a gene encoding a protein identified either from N-terminal targeting predictions to function in the chloroplast or mitochondrion. Clusters of adjacent squares pertain to genes encoding different components of a specific multi-unit enzyme or complex; and split squares pertain to genes encoding functional homologs of one specific protein. The assigned merged modules are indicated as their respective colors, with the 16 most abundant merged modules shown in the legend. Additionally, proteins coded in organellar genomes (Oudot-Le Secq et al., 2007; Oudot-Le Secq and Green, 2011; Yu et al., 2018) are shown as dotted green or red; proteins for which chloroplast- or mitochondria-targeted isoforms or merged modules could not be assigned are shown as light gray; and enzymatic steps not identified in the genome are shown as light gray squares without borders. Dual-localized proteins (Gile et al., 2015; Dorrell et al., 2017) are marked by checkered yellow boxes; while orange boxes highlight potential connection points between the two organelles. Abbreviations are as follows: CAs, carbonic anhydrases; MEP/DOXP, mevalonate and non-mevalonate pathways for isoprenoid biosynthesis; SUF, iron-sulfur complex assembly; MPP,/TPP/SPP, mitochondrial, thylakoid and stromal processing peptidases; TAT, twin-arginine-dependent thylakoid protein import pathways; AOX/PTOX, mitochondrial and chloroplast alternative oxidases; TCA, Citric Acid cycle; Orn, ornithine; GCS, glycine shuttle; GS-GOGAT, glutamine synthetase/glutamate synthase shuttle. Detailed enzyme distributions for each pathway are shown in **Supplementary Figure 6**.

of genes encoding chloroplast ribosomal proteins in the ivory merged module (otherwise enriched in chloroplast branched-chain amino acid and lysine biosynthesis (Bromke, 2013). It remains to be determined to what extent the expression of the chloroplast- and mitochondrial-genomes of *P. tricornutum* are regulated in response to the redox state, versus metabolic fluxes experienced in both organelles.

Finally, we considered the repartition of functionally uncharacterized, but conserved domains across chloroplast-targeted proteins in our dataset, focusing on DUFs (Domains of Unknown Function). We found 30 chloroplast-targeted proteins containing at least one DUF and 8 DUFs assigned to at least two chloroplast-targeted proteins (**Supplementary Table 3**, Sheet 2). Amongst these recurrent chloroplast-associated DUFs were two examples (DUF1995 and DUF3493), which have previously been implicated to function in photosystem assembly within thylakoid membranes (Chi et al., 2012; Bohne et al., 2016; Li et al., 2019). Both of these DUFs were found amongst chloroplast-targeted proteins in the paleturquoise merged module (Phatr3_J38149, Phatr3_J40136, and Phatr3_J46926, containing DUF1995; and Phatr3_EG02444, containing DUF3493); in the blue module (Phatr3_J44212, containing DUF1995; and Phatr3_J45569, containing DUF3494); and DUF1995 furthermore occurred in chloroplast-targeted proteins in the lightcyan1 (Phatr3_J44529) and lightsteelblue (Phatr3_J40199) modules (**Supplementary Table 3**, Sheet 2). Each of these modules are enriched in different chloroplast-targeted metabolism pathways (**Figures 3**, **4**), suggesting complex connections between the regulation of chloroplast anabolism and photosystem assembly.

Amongst the other DUFs associated with more than one chloroplast-targeted protein were DUF814, which is implicated in RNA quality control and amongst *P. tricornutum* chloroplast-targeted protein includes one (Phatr3_J45207, within the paleturquoise module) with some structural homology to a ferrous iron transporter (Maxwell Burroughs and Aravind, 2014); and DUF563, which contains a carbohydrate-active domain (Park et al., 2010) and includes at least one chloroplast-targeted protein (Phatr3_EG00581) within the blue module, otherwise implicated in chloroplast carbon metabolism (**Figures 3**, **4**). It remains to be determined if either of these proteins has novel functions, e.g., respectively in iron status sensing or in the diversification of carbohydrate metabolism in the *P. tricornutum* chloroplast, via the generating and phenotyping of mutant lines.

PhaeoNet Merged Modules Identify Complex Crosstalk between the Chloroplast and Mitochondrion in *P. tricornutum*

Previously, intricate metabolic connections have been observed between *P. tricornutum* chloroplasts and mitochondria, which are distinctive to those found in plants (Prihoda et al., 2012; Bailleul et al., 2015; Broddrick et al., 2019; Murik et al., 2019). We wished to determine which of these connections were visible within our data, noting multiple, transcriptionally independent connections between the predicted proteomes of chloroplasts and mitochondria in PhaeoNet data (highlighted in **Supplementary Figure 6**). These included the presence of genes encoding chloroplast-targeted protein import subunits (e.g., Phatr3_J32195 encoding Tic20, Phatr3_EG02421 encoding Tic21) within the otherwise predominantly mitochondrial orangered4 merged module and the presence of large numbers of amino-acyl tRNA synthetase genes (which are typically dual-targeted to the chloroplasts and mitochondria in diatoms (Gile et al., 2015; Dorrell et al., 2017, 2019) in the otherwise chloroplast-associated blue merged module.

We furthermore noted the presence of multiple chloroplast-targeted proteins associated with chloroplast division (e.g., Phatr3_J34093, Phatr3_J42361, and Phatr3_J14995, encoding FtsZ-type division proteins) in the blue module, potentially linking the synthesis of chloroplast and mitochondrial tRNAs to chloroplast replication. A further two proteins implicated in chloroplast replication (e.g., Phatr3_J21455, encoding a dynamin-related DRPB85-class protein and Phatr3_J14426, encoding a further FtsZ protein)—were found in the darkgray module, which was also populated by proteins involved in mitochondrial protein import (MPP, TIM, OXA1; **Supplementary Table 3**), suggesting probable links between chloroplast and mitochondrial biogenesis. Of note, at least two of the FtsZ proteins (Phatr3_J34093, within the blue module and Phatr3_J14426, within the darkgray module) were inferred to possess both chloroplast and mitochondrial-targeting sequences, underpinning the likely coordination of biogenesis of both organelles (**Supplementary Table 3**). This coordination may underpin the close topological associations and synchronized division cycles observed between the *P. tricornutum* mitochondrion and chloroplast observed *in vivo* (Tanaka et al., 2015; Dorrell and Bowler, 2017).

Alongside these more general links, we identified specific points of co-regulation between each organelle. The paleturquoise merged module, as the only merged module found to be enriched in both chloroplast and mitochondria functions (**Figure 3**) was of particular interest and contained genes encoding enzymes participating in several different chloroplast and mitochondria metabolic pathways. These included genes for mitochondria-targeted glycine dehydrogenase (Phatr3_J22187) and serine hydroxymethyltransferase (Phatr3_J32847) and a gene for a chloroplast-targeted dihydrolipoamide dehydrogenase (Phatr3_J30113), which participate (as part of the glycine shuttle) in metabolic recycling of 2-P-glycolate produced through photosynthesis (**Supplementary Figure 6**; Zheng et al., 2013; Davis et al., 2017). The paleturquoise merged module additionally contains a gene encoding mitochondria-targeted malate dehydrogenase (Phatr3_J54082), which may additionally participate in the photorespiratory metabolism of glycolate by allowing the recycling of mitochondrial serine (via pyruvate) in the TCA cycle (Davis et al., 2017; Broddrick et al., 2019). Genes encoding at least three further plastidial oxidative stress-related proteins (Phatr3_J12583, encoding Fe-Mn family superoxide dismutase; Phatr3_J45252, encoding a plastidial thioredoxin; and Phatr3_J31436, encoding a plastidial ortholog of peroxisomal membrane protein 2, Davis et al., 2017; Dorrell et al., 2017) belong to the paleturquoise merged module, underlining its importance in oxidative stress responses.

Genes encoding both glutamine synthase (GS) and glutamate synthase/glutamine oxoglutarate aminotransferase (GOGAT), which have distinct plastidial and mitochondrial homologs in *P. tricornutum* (Broddrick et al., 2019; Smith et al., 2019), belong to different PhaeoNet merged modules (cyan, tan, steelblue and magenta), suggesting a relatively complex regulation of this hub. The plastid-localized GS (encoded by Phatr3_J51092) belongs to the magenta merged module, which also contains the subsequent genes encoding enzymes mediating the entry of GS-produced NH_3 into the mitochondrial ornithine-urea cycle (Phatr3_J42398 encoding malate dehydrogenase; Phatr3_J30145 encoding citrate synthase; Phatr3_J22913 encoding pyruvate kinase), suggesting this co-regulated pathway may have roles in recycling excess NH_3 produced in the chloroplast, in accordance with previous studies (Levering et al., 2016; Broddrick et al., 2019; Smith et al., 2019).

Finally, we noted the presence of genes encoding chloroplast-targeted plastoquinol terminal oxidase (Phatr3_J4283) and mitochondria-targeted alternative oxidase (Phatr3_EG02359), which are both associated with the photoprotective removal of excess metabolic reducing potential in the skyblue merged module (Bailleul et al., 2015; Murik et al., 2019). This merged module, as discussed above, contains genes encoding three successive enzymes associated with the TCA cycle (Phatr3_J40430 encoding α-ketoglutaryl dehydrogenase; Phatr3_J42015 encoding succinyl-CoA synthetase and Phatr3_J41812 encoding succinate dehydrogenase; Kroth et al., 2008), along with methylmalonyl-CoA mutase (Phatr3_J51830), which may allow excess succinyl-CoA to be diverted into lipid synthesis via propionyl-CoA (Helliwell et al., 2011; Valenzuela et al., 2012). This co-regulation underlines the importance of the succinate hub, and presumably both the glyoxylate cycle and ornithine shunt (as sources of mitochondrial α-ketoglutarate), as routes for the mitochondrial dissipation of excess chloroplast reducing potential (Bailleul et al., 2015; Broddrick et al., 2019).

Transcriptional Regulators of Chloroplast-Targeted Proteins Show Separate Expression Dynamics, Informed by Evolutionary History

Finally, given the complex transcriptional partitioning of genes encoding components of chloroplast and mitochondrial metabolism pathways across PhaeoNet data, we investigated what transcriptional drivers might be implicated in the co-regulation of different metabolism-enriched pathway clusters. First, we considered the repartition of a manually curated list of genes encoding proteins implied in histone and transcription-related processes (including transcription factors, TFs; Rayko et al., 2010; Banerjee et al., 2016) across all merged modules (**Supplementary Figure 7** and **Supplementary Table 4**). These genes were most frequently observed (>5% of total merged module genes) in the darkgray, brown and steelblue merged modules (**Supplementary Figure 7** and **Figure 3**). The brown and darkgray merged modules were additionally enriched in KEGG merged modules related to cytoskeleton proteins (**Supplementary Figure 4**), pointing to close links between cytoskeletal organization and transcriptional regulation in diatoms (for example, within organization of the cell cycle (Huysman et al., 2013; Tanaka et al., 2015). The single most abundant TF family, heat shock factor family (HSF) proteins (Rayko et al., 2010), were most frequently detected in the brown, brown4, cyan and skyblue merged modules (> 5 HSFs each, **Supplementary Figure 7**). Notably, both the brown and brown4 merged modules are also enriched in KEGG functions associated with stress responses (protein ubiquitinylation, autophagy and membrane trafficking) (**Supplementary Figure 5**), consistent with previously inferred functions of specific *P. tricornutum* HSFs in the maintenance of cellular fitness (Chen et al., 2014; Egue et al., 2015).

We also found specific repartitions of genes encoding proteins implicated in light- and circadian-dependent transcriptional responses in *P. tricornutum*, e.g., aureochromes and cryptochromes (Takahashi et al., 2007; Banerjee et al., 2016). These proteins typically have cytoplasmic localizations, but through the perception of light and translocation to the nucleus can regulate the expression of core chloroplast metabolic pathways (Kroth et al., 2017). The circadian-regulated Aureochrome 1c (Phatr3_J12346; Banerjee et al., 2016; Kroth et al., 2017) and a cryptochrome-like blue light receptor (Phatr3_J34592) were both found in the blue merged module, implicated in anabolic metabolism; and the light-regulated Aureochrome 1b (Phatr3_J15977) and the blue-light-dependent protochlorophyllide reductase 1 (Phatr3_J12155; Hunsperger et al., 2016; Mann et al., 2017) were both found in the lightsteelblue1 merged module, alongside the majority of genes encoding other pigment biosynthesis enzymes. In contrast, the gene encoding Aureochrome 1a (Phatr3_J49116), which is essential for high light acclimation but appears to be under exclusively circadian (light-independent) regulation, falls within the lightcyan1 merged module of core photosystem-associated genes (**Supplementary Table 4** and **Supplementary Figure 7**; Banerjee et al., 2016; Mann et al., 2017); while RITMO1 (Phatr3_J44962), associated with the *P. tricornutum* circadian clock, falls within the skyblue merged module, which contains limited chloroplast-related functions except for alternative electron flow pathways (**Supplementary Table 4** and **Supplementary Figure 7**; Annunziata et al., 2019). The separate distributions of light- and circadian-regulated chloroplast regulators might reflect a circadian-entrained synthesis of the core photosynthetic machinery (via Aureochrome 1a), independent of light status, with chloroplast biosynthesis pathways upregulated both by circadian signaling (via Aureochrome 1c) and as a function of light availability (via Aureochrome 1b). This is reminiscent of circadian gene expression patterns visualized in plant and other algal lineages (e.g., the green alga *Ostreococcus* and the dinoflagellate *Lingulodinium*), in which photosynthesis and plastid biogenesis proteins are either expressed at separate times of the day, or show different regulatory responses to circadian and light signals (Wang et al., 2005; Monnier et al., 2010; Noordally et al., 2013). Finally, the gene encoding the Aureochrome 2 protein (Phatr3_J8113), which lacks the conserved flavin-binding domain required for light perception (Takahashi et al., 2007; Kroth et al., 2017), falls within the greenyellow merged module of generally transcriptionally

repressed proteins (**Figure 3**), underlining its independence of chloroplast functions.

Finally, we wished to consider within our dataset what transcriptional dynamics within the nuclear genome may underpin chloroplast gene expression in *P. tricornutum*. Chloroplast transcription in *P. tricornutum*, as in other diatoms, is performed by a plastid-encoded RNA polymerase, unlike the situation in plants in which both plastid- and nuclear-encoded and plastid-targeted polymerases participate (Oudot-Le Secq et al., 2007; Yu et al., 2018). Plastid-encoded RNA polymerases in plants typically interact with nucleus-encoded sigma factors, which may direct them to specific target genes, in response to different regulatory and physiological signals (Shimizu et al., 2010; Noordally et al., 2013). Eight genes are annotated in the *P. tricornutum* nuclear genome to encode sigma factor related proteins (Rayko et al., 2010; **Supplementary Table 4**), but the functions of each protein with regard to the expression of the chloroplast genome remain unclear.

We investigated the functions of *P. tricornutum* sigma factors by combining the repartition of each sigma factor in PhaeoNet with predicted *in silico* localizations of *P. tricornutum* proteins and their closest homologs from other diatom species, as resolved with a single-gene (RAxML) tree (**Figure 5**). Three of the sigma factor genes in *P. tricornutum* possess chloroplast-targeting sequences, as inferred by *in silico* prediction with HECTAR and ASAFind (Gschloessl et al., 2008; Gruber et al., 2015). One of these proteins (Phatr3_J14599, SIGMA1a) falls within the paleturquoise merged module, which is otherwise enriched in chloroplast-related functions pertaining to carbon concentration and the glycine shunt (**Figures 3–5**); while the two remaining chloroplast-targeted proteins (Phatr3_J3388, SIGMA1b; Phatr3_J17029, SIGMA3) fall within the steelblue module, which otherwise lacks obvious enrichments in chloroplast-targeted functions and instead seems to be most closely connected to nucleotide metabolism (**Figures 3**, **5**). Phylogenetic analysis of these three sigma factors indicate that many of their closest

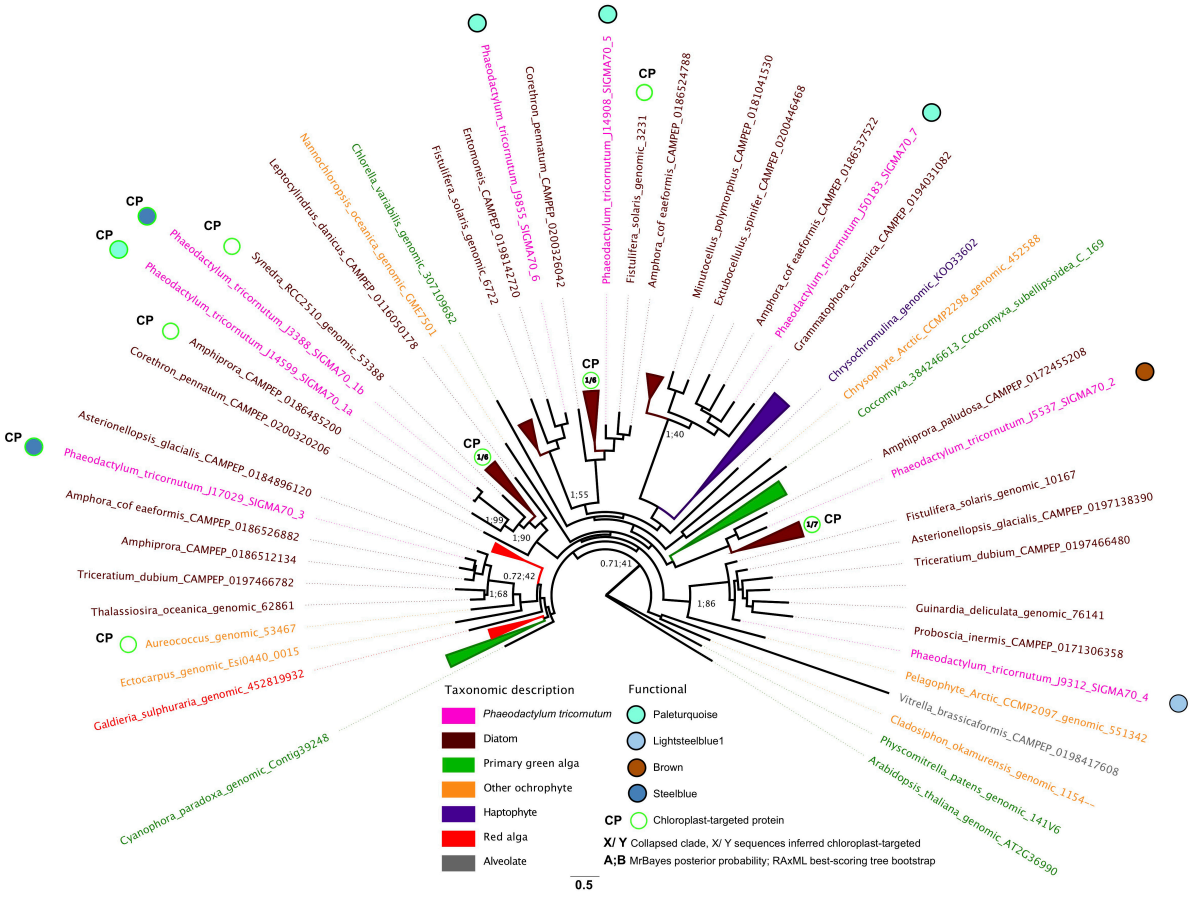

FIGURE 5 | Phylogenetic and transcriptional dynamics of *P. tricornutum* sigma factors. This Figure shows an unrooted best-scoring tree topology for an 86 taxa x 453 aa alignment of subsampled diatom and non-diatom sigma factors and realized using MrBayes v 3.2.7a with the Jones substitution matrix, 600,000 generations, two start chains and 0.5 burnin thresholds (Huelsenbeck and Ronquist, 2001); and RAxML v 8.2 with the PROTGAMMAJTT substitution model with 300 bootstrap replicates (Stamatakis, 2014). Chloroplast-targeting predictions were performed using ASAFind with SignalP v 3.0 (Gruber et al., 2015); and HECTAR (Gschloessl et al., 2008) under default conditions. Branches are colored by phylogenetic affiliation and bootstrap values of nodes recovered with > 40% support are shown. Eight *P. tricornutum* sigma factors are labeled with PhaeoNet merged module repartition and chloroplast targeting sequences were predicted by HECTAR or ASAFind (Gruber et al., 2015).

relatives are sequences with chloroplast-targeting signals from other diatoms, and indeed SIGMA1a and SIGMA1b appear to be recently derived paralogs of one another (**Figure 5**), indicating that they are likely to be conserved parts of the diatom chloroplast transcriptional machinery. The repartition of SIGMA1b and SIGMA3 within a transcriptional module that is largely related to non-chloroplast processes may allow hierarchical control of chloroplast transcription in response to non-chloroplast signals in *P. tricornutum* (e.g., coordination with circadian or cell cycles, Noordally et al., 2013; Tanaka et al., 2015).

The remaining five *P. tricornutum* sigma factors were not predicted to be targeted to the chloroplast and phylogenetic analysis indicated that their closest diatom relatives primarily also lacked chloroplast-targeting signals (**Figure 5**). One of these non-chloroplast-associated sigma factors (Phatr3_J5537, SIGMA2) fell within the largely chloroplast-independent brown module, suggesting that it has non-chloroplastic functions. In contrast, the remaining non-chloroplast targeted sigma factors fell within modules otherwise enriched in chloroplast-associated functions; either lightsteelblue (Phatr3_J9312, SIGMA4), or paleturquoise (Phatr3_J14908, Phatr3_J9855, Phatr3_J50183; SIGMA 5-7; **Figures 3**, **5**). It remains to be determined whether these sigma factors are targeted to the *P. tricornutum* chloroplast, but using alternative methods to those recognized by HECTAR or ASAFind, as per certain other diatom proteins (Kazamia et al., 2018; Schober et al., 2019); function in compartments other than the chloroplast, but participate indirectly in the regulation, e.g., of nucleus-encoded proteins implicated in chloroplast metabolism; or have functions independent of the chloroplast, as has been documented for some other eukaryotic sigma factors (Shadel and Clayton, 1995; Beardslee et al., 2002). These different possibilities may be best discriminated by the experimental characterization, e.g., through mutagenesis and functional phenotyping, of individual *P. tricornutum* sigma factor genes.

CONCLUDING REMARKS

In this project, we have used WGCNA to build an integrated network of *P. tricornutum* gene co-regulation, which we name "PhaeoNet." Our model is able to retrieve well established biological pathways (e.g., chloroplast photosynthetic, anabolic metabolism; and mitochondrial respiration, **Figure 4**) and compares favorably to existing (e.g., microarray-based; Ashworth et al., 2016) studies of gene co-regulation for this species (Ashworth et al., 2016; **Figures 1**, **2** and **Supplementary Figures 1–4**). Moreover, our dataset carries the advantage of decomposing the *P. tricornutum* genome into a smaller number (28) of functionally distinct modules than produced by DiatomPortal. We have integrated these data into previously generated functional, targeting and evolutionary analyses of the *P. tricornutum* genome, allowing us to gain holistic insights into the processes underpinning the gene co-regulation of specific biological processes and organelle metabolic pathways pertinent to diatom biology (**Figure 3** and **Supplementary Figure 5**).

Through a deeper inspection of genes encoding chloroplast and mitochondria-targeted proteins within these data, we identify PhaeoNet merged modules underpinning anabolic (blue), photosynthetic (lightsteelblue 1, lightcyan1) and respiratory (orangered4, cyan) metabolism, and identify multiple metabolic connections between the chloroplast and mitochondria. These include the glycine shunt within the paleturquoise merged module; the ornithine-urea cycle within the magenta merged module; and coordinated chloroplast and mitochondrial alternative oxidase activities in the skyblue merged module; **Figure 4** and **Supplementary Figure 6**. Finally, considering the repartition of transcription-related proteins within our data, we identify probable cognate regulators for different co-ordinated metabolic pathways (**Figure 5** and **Supplementary Figure 7**), demonstrating different associations of aureochrome transcription factors with different chloroplast metabolic pathways. We notably identify hidden diversity in the range of sigma factor genes in the *P. tricornutum* genome, some of which are likely to be involved in the transcriptional regulation of different chloroplast-encoded genes in response to different physiological signals, while others are likely to have different functions to chloroplast gene expression.

The PhaeoNet dataset may be usable as a predictive tool for the characterization of poorly understood proteins, either directly in *P. tricornutum*, as a well-studied model diatom species, or in other diatom or microalgal species for which homologs of *P. tricornutum* proteins are known either from genome or transcriptome datasets (e.g., Keeling et al., 2014; Carradec et al., 2018; Sato et al., 2020). We stress that biological processes elucidated in this species may not necessarily be directly extrapolatable to other algal species; and examples are already known of proteins (e.g., proteins involved in iron-stress tolerance and C4 photosynthesis) that may have different physiological functions even between different diatoms (Kustka et al., 2014; Lampe et al., 2018). Cross-comparisons between PhaeoNet and other data, e.g., gene coregulation datasets erected in other, less well-studied species (Ashworth et al., 2016; Ashworth and Ralph, 2018); environmental expression trends (Carradec et al., 2018); and the phenotypes of a wider range of mutant lines generated in *P. tricornutum* will be essential to understanding the diversity of functions performed by understudied proteins in diatoms and other algae. Nonetheless, insights from our data, delivering actors and signatures of metabolic co-regulation in diatoms, will provide a useful community resource for subsequent directed experimental investigation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/42xmp/.

AUTHOR CONTRIBUTIONS

OA-M was responsible for the design and construction of PhaeoNet. AMGNV and NJ performed the functional analysis of the PhaeoNet modules. YL and XZ participated in the construction of the data used for functional analysis. AG, LT, and CB were responsible for the supervision of the construction of

PhaeoNet. RGD was responsible for the supervision of functional analysis. OA-M and RGD wrote the manuscript, with input from all other co-authors. All authors contributed to the article and approved the submitted version.

FUNDING

RGD and AMGNV acknowledge funding from a CNRS Momentum Fellowship (awarded to RGD, 2019-2021). CB acknowledges funding from the European Research Council for an Advanced Award (grant ERC 835067-DIATOMIC), grants from the French Agence Nationale de la Recherche (MEMOLIFE, ref. ANR10-LABX-54, OCEANOMICS, ref. ANR-11-BTBR-0008 and BrownCut, ref. ANR-19-CE20-0020) and Research Grant "Green Life in the Dark" (RGP0003/2016) from the Human Frontier Science Program. LT acknowledges funding from the CNRS, the region of Pays de la Loire (ConnecTalent EPIALG project) and Nantes métropoles. XZ acknowledges a Ph.D. fellowship from the Chinese Scholarship Council (CSC-201604910722) and funding from Région Pays de la Loire (Awarded to LT).

ACKNOWLEDGMENTS

We would like to thank Dr. Justin Ashworth (University of Technology Sydney) and Dr. Serdar Turkarslan (Institute for Systems Biology, Seattle) for the kind provision of analogous cluster composition data from the DiatomPortal project and Prof. Angela Falciatore (Institut de Biologie Physico-Chimique, Paris) for assistance with the annotation of *P. tricornutum* transcription factors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.590949/full#supplementary-material

REFERENCES

Aibar, S., Fontanillo, C., Droste, C., and De Las Rivas, J. (2015). Functional gene networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* 31, 1686–1688. doi: 10.1093/bioinformatics/btu864

Allen, A. E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., et al. (2008). Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10438–10443. doi: 10.1073/pnas.0711370105

Allen, J. F. (2003). The function of genomes in bioenergetic organelles. *Phil. Trans. R. Soc. Biol.* 358, 19–37. doi: 10.1098/rstb.2002.1191

Annunziata, R., Ritter, A., Fortunato, A. E., Manzotti, A., Cheminant-Navarro, S., Agier, N., et al. (2019). bHLH-PAS protein RITMO1 regulates diel biological rhythms in the marine diatom P. *tricornutum*. *Proc. Natl. Acad. Sci. U.S.A.* 116, 13137–13142. doi: 10.1073/pnas.1819660116

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., et al. (2019). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 36, 2251–2252. doi: 10.1093/bioinformatics/btz859

Ashworth, J., and Ralph, P. J. (2018). An explorable public transcriptomics compendium for eukaryotic microalgae. *bioRxiv[Preprint]*. doi: 10.1101/403063

Ashworth, J., Turkarslan, S., Harris, M., Orellana, M. V., and Baliga, N. S. (2016). Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Mar. Genom.* 26, 21–28. doi: 10.1016/j.margen.2015.10.011

Bailleul, B., Berne, N., Murik, O., Petroutsos, D., Prihoda, J., and Tanaka, A. (2015). Energetic coupling between plastids and mitochondria drives CO2 assimilation in diatoms. *Nature* 524, 366–371. doi: 10.1038/nature14599

Banerjee, A., Herman, E., Serif, M., Maestre-Reyna, M., Hepp, S., Pokorny, R., et al. (2016). Allosteric communication between DNA-binding and light-responsive domains of diatom class I aureochromes. *Nucl. Acids Res.* 44, 5957–5970. doi: 10.1093/nar/gkw420

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science* 325, 412–413. doi: 10.1126/science.1173299

Bártová, E., Krejčí, J., Harničarová, A., Galiová, G., and Kozubek, S. (2008). Histone modifications and nuclear architecture: a review. *J. Histochem. Cytochem* 56, 711–721. doi: 10.1369/jhc.2008.951251

Beardslee, T. A., Roy-Chowdhury, S., Jaiswal, P., Buhot, L., Lerbs-Mache, S., Stern, D. B., et al. (2002). A nucleus-encoded maize protein with sigma factor activity accumulates in mitochondria and chloroplasts. *Plant. J.* 31, 199–209. doi: 10.1046/j.1365-313x.2002.01344.x

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol* 340, 783–795. doi: 10.1016/j.jmb.2004.05.028

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Bertrand, M. (2010). Carotenoid biosynthesis in diatoms. *Photosynthesis Res.* 106, 89–102. doi: 10.1007/s11120-010-9589-x

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21. doi: 10.1101/gad.947102

Bohne, A.-V., Schwenkert, S., Grimm, B., and Nickelsen, J. (2016). Roles of tetratricopeptide repeat proteins in biogenesis of the photosynthetic apparatus. *Int. Rev. Cell Mol. Biol.* 324, 187–227. doi: 10.1016/bs.ircmb.2016.01.005

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., et al. (2008). The P. *tricornutum* genome reveals the evolutionary history of diatom genomes. *Nature* 456, 239–244. doi: 10.1038/nature07410

Bowler, C., Vardi, A., and Allen, A. E. (2010). Oceanographic and biogeochemical insights from diatom genomes. *Ann. Rev. Mar. Sci.* 2, 333–365. doi: 10.1146/annurev-marine-120308-081051

Broddrick, J. T., Du, N., Smith, S. R., Tsuji, Y., Jallet, D., and Ware, M. A. (2019). Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom *Phaeodactylum tricornutum*. *New Phytol.* 222, 1364–1379. doi: 10.1111/nph.15685

Bromke, M. A. (2013). Amino Acid biosynthesis pathways in diatoms. *Metabolites* 3, 294–311. doi: 10.3390/metabo3020294

Büchel, C. (2015). Evolution and function of light harvesting proteins. *J. Plant. Physiol.* 172, 62–75. doi: 10.1016/j.jplph.2014.04.018

Buck, J. M., Sherman, J., Río Bártulos, C., Serif, M., Halder, M., and Henkel, J. (2019). Lhcx proteins provide photoprotection via thermal dissipation of absorbed light in the diatom *Phaeodactylum tricornutum*. *Nat. Commun.* 10:4167. doi: 10.1038/s41467-019-12043-6

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., et al. (2018). A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9:373. doi: 10.1038/s41467-017-02342-1

Chappell, P. D., Whitney, L. P., Wallace, J. R., Darer, A. I., Jean-Charles, S., and Jenkins, B. D. (2015). Genetic indicators of iron limitation in wild populations

of Thalassiosira oceanica from the northeast Pacific Ocean. *ISME J.* 9, 592–602. doi: 10.1038/ismej.2014.171

Chen, Z., Yang, M. K., Li, C. Y., Wang, Y., Zhang, J., Wang, D. B., et al. (2014). Phosphoproteomic analysis provides novel insights into stress responses in *Phaeodactylum tricornutum*, a model diatom. *J Proteom. Res.* 13, 2511–2523. doi: 10.1021/pr401290u

Chi, W., Ma, J., and Zhang, L. (2012). Regulatory factors for the assembly of thylakoid membrane protein complexes. *Phil. Trans. R. Soc. B* 367, 3420–3429. doi: 10.1098/rstb.2012.0065

Cihlar, J., Fussy, Z., Horak, A., and Obornik, M. (2016). Evolution of the tetrapyrrole biosynthetic pathway in secondary algae: conservation, redundancy and replacement. *PLoS One* 11:0166338. doi: 10.1371/journal.pone.0166338

Cruz de Carvalho, M. H., Sun, H. X., Bowler, C., and Chua, N. H. (2016). Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New Phytol.* 210, 497–510. doi: 10.1111/nph.13787

Dautermann, O., and Lohr, M. (2017). A functional zeaxanthin epoxidase from red algae shedding light on the evolution of light-harvesting carotenoids and the xanthophyll cycle in photosynthetic eukaryotes. *Plant J.* 92, 879–891. doi: 10.1111/tpj.13725

Davis, A., Abbriano, R., Smith, S. R., and Hildebrand, M. (2017). Clarification of photorespiratory processes and the role of Malic Enzyme in diatoms. *Protist* 168, 134–153. doi: 10.1016/j.protis.2016.10.005

Dłużewska, J., Szymanìska, R., Gabruk, M., Koìs, P. B., Nowicka, B., and Kruk, J. (2016). Tocopherol cyclases- substrate specificity and phylogenetic relations. *PLoS One* 11:0159629. doi: 10.1371/journal.pone.0159629

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformat.* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dorrell, R. G., Azuma, T., Nomura, M., Audren de Kerdrel, G., Paoli, L., Yang, S., et al. (2019). Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. *Proc. Natl. Acad. Sci. U.S.A.* 116, 6914–6923. doi: 10.1073/pnas.1819976116

Dorrell, R. G., and Bowler, C. (2017). Secondary plastids of stramenopiles. *Adv. Bot. Res.* 84, 59–103.

Dorrell, R. G., Gile, G., McCallum, G., Méheust, R., Bapteste, E. P., Klinger, C. M., et al. (2017). Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* 6, 23717. doi: 10.7554/eLife.23717

Egue, F., Chenais, B., Tastard, E., Marchand, J., Hiard, S., Gateau, H., et al. (2015). Expression of the retrotransposons Surcouf and Blackbeard in the marine diatom *Phaeodactylum tricornutum* under thermal stress. *Phycologia* 54, 617-627.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237–240. doi: 10.1126/science.281.5374.237

Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, S., et al. (2008). Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol. Biol. Evol.* 25, 2653–2667. doi: 10.1093/molbev/msn206

Fukasawa, Y., Tsuji, J., Fu, S. C., Tomii, K., Horton, P., Imai, K., et al. (2015). MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteom.* 14, 1113–1126. doi: 10.1074/mcp.M114.043083

Gasch, A. P., and Eisen, M. B. (2002). Exploring the conditional co-regulation of yeast gene expression through fuzzy k-means clustering. *Genom. Biol.* 3:0059. doi: 10.1186/gb-2002-3-11-research0059

Gile, G. H., Moog, D., Slamovits, C. H., Maier, U. G., and Archibald, J. M. (2015). Dual organellar targeting of aminoacyl-tRNA synthetases in diatoms and cryptophytes. *Genom. Biol. Evol.* 7, 1728–1742. doi: 10.1093/gbe/evv095

Glasser, G. J., and Winter, R. F. (1961). Critical values of the coefficient of rank correlation for testing the hypothesis of independence. *Biometrika* 48, 444–448. doi: 10.2307/2332767

Grouneva, I., Rokka, A., and Aro, E. M. (2011). The thylakoid membrane proteome of two marine diatoms outlines both diatom-specific and species-specific features of the photosynthetic machinery. *J. Proteom. Res.* 10, 5338–5353. doi: 10.1021/pr200600f

Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V., and Mock, T. (2015). Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J.* 81, 519–528. doi: 10.1111/tpj.12734

Gschloessl, B., Guermeur, Y., and Cock, J. M. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformat.* 9:393. doi: 10.1186/1471-2105-9-393

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465–470. doi: 10.1038/nature16942

Gundermann, K., Schmidt, M., Weisheit, W., Mittag, M., and Büchel, C. (2013). Identification of several sub-populations in the pool of light harvesting proteins in the pennate diatom *Phaeodactylum tricornutum*. *Biochim. Biophys. Acta.* 1827, 303–310. doi: 10.1016/j.bbabio.2012.10.017

Helliwell, K. E., Wheeler, G. L., Leptos, K. C., Goldstein, R. E., and Smith, A. G. (2011). Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes. *Mol. Biol. Evol.* 28, 2921–2933. doi: 10.1093/molbev/msr124

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucl. Acids Res.* 35, 585–587. doi: 10.1093/nar/gkm259

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Hunsperger, H. M., Ford, C. J., Miller, J. S., and Cattolico, R. A. (2016). Differential regulation of duplicate light-dependent protochlorophyllide oxidoreductases in the diatom *Phaeodactylum tricornutum*. *PLoS One* 11:0158614. doi: 10.1371/journal.pone.0158614

Huysman, M. J., Fortunato, A. E., Matthijs, M., Costa, B. S., Vanderhaeghen, R., Van den Daele, H., 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. doi: 10.1105/tpc.112.106377

Kanehisa, M. (2017). Enzyme Annotation and Metabolic Reconstruction Using KEGG. *Methods Mol. Biol.* 1611, 135–145. doi: 10.1007/978-1-4939-7015-5_11

Kanehisa, M., and Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.* 29, 28–35. doi: 10.1002/pro.3711

Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kazamia, E., Sutak, R., Paz-Yepes, J., Dorrell, R. G., Vieira, F. R. J., Mach, J., et al. (2018). Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms. *Sci. Adv.* 16:5. doi: 10.1126/sciadv.aar4536

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., et al. (2014). The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 10:1001889. doi: 10.1073/pnas.1603112113

Kikutani, S., Nakajima, K., Nagasato, C., Tsuji, Y., Miyatake, A., and Matsuda, Y. (2016). Thylakoid luminal theta-carbonic anhydrase critical for growth and photosynthesis in the marine diatom *Phaeodactylum tricornutum*. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9828–9833. doi: 10.1073/pnas.1603112113

Kim, E.-D., and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.* 17, 16–21. doi: 10.1016/j.tplants.2011.10.008

Kroth, P. G., Chiovitti, A., Gruber, A., Martin-Jezequel, V., Mock, T., and Parker, M. S. (2008). A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PLoS One* 3:e1426. doi: 10.1371/journal.pone.0001426

Kroth, P. G., Wilhelm, C., and Kottke, T. (2017). An update on aureochromes: phylogeny - mechanism - function. *J. Plant Physiol.* 217, 20–26. doi: 10.1016/j.jplph.2017.06.010

Kustka, A., Milligan, A. J., Zheng, H., New, A. M., Gates, C., Bidle, K. D., et al. (2014). Low CO2 results in a rearrangement of carbon metabolism to support C4 photosynthetic carbon assimilation in Thalassiosira pseudonana. *New Phytol.* 204, 507–520. doi: 10.1111/nph.12926

Lampe, R. H., Mann, E. L., Cohen, N. R., Till, C. P., Thamatrakoln, K., Brzezinski, M. A., et al. (2018). Different iron storage strategies among bloom-forming

diatoms. *Proc. Natl. Acad. Sci. U.S.A* 115, 12275–12284. doi: 10.1073/pnas.1805243115

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformat.* 9:559. doi: 10.1186/1471-2105-9-559

Levering, J., Broddrick, J., Dupont, C. L., Peers, G., Beeri, K., Mayers, J., et al. (2016). Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One* 11:0155038. doi: 10.1371/journal.pone.0155038

Li, X., Patena, W., Fauser, F., Jinkerson, R. E., Saroussi, S., Meyer, M. T., et al. (2019). A genome-wide algal mutant library reveals a global view of genes required for eukaryotic photosynthesis. *Nat. Genet.* 1, 627–635. doi: 10.1038/s41588-019-0370-6

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genom. Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Maheswari, U., Jabbari, K., Petit, J.-L., Porcel, B. M., Allen, A. E., Cadoret, J.-P., et al. (2010). Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genom. Biol.* 11:85. doi: 10.1186/gb-2010-11-8-r85

Mann, M., Serif, M., Jakob, T., Kroth, P. G., and Wilhelm, C. (2017). PtAUREO1a and PtAUREO1b knockout mutants of the diatom *Phaeodactylum tricornutum* are blocked in photoacclimation to blue light. *J. Plant Physiol.* 217, 44–48. doi: 10.1016/j.jplph.2017.05.020

Maréchal, E., and Lupette, J. (2020). Relationship between acyl-lipid and sterol metabolisms in diatoms. *Biochimie* 169, 3–11. doi: 10.1016/j.biochi.2019.07.005

Maxwell Burroughs, A., and Aravind, L. (2014). A highly conserved family of domains related to the DNA-glycosylase fold helps predict multiple novel pathways for RNA modifications. *RNA Biol.* 11, 360–372. doi: 10.4161/rna.28302

McCarthy, J. K., Smith, S. R., McCrow, J. P., Tan, M., Zheng, H., Beeri, K., et al. (2017). Nitrate reductase knockout uncouples nitrate transport from nitrate assimilation and drives repartitioning of carbon flux in a model pennate diatom. *Plant Cell* 29, 2047–2070. doi: 10.1105/tpc.16.00910

Michalak, P. (2008). Coexpression, co-regulation and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* 91, 243–248. doi: 10.1016/j.ygeno.2007.11.002

Monnier, A., Liverani, S., Bouvet, R., Jesson, B., Smith, J. Q., Mosser, J., et al. (2010). Orchestrated transcription of biological processes in the marine picoeukaryote Ostreococcus exposed to light/dark cycles. *BMC Genom* 11:192. doi: 10.1186/1471-2164-11-192

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J., Couce, M. L., and Cocho, J. A. (2013). A glimpse into past, present and future DNA sequencing. *Mol. Genet. Metab.* 110, 3–24. doi: 10.1016/j.ymgme.2013.04.024

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucl. Acids Res.* 35, 182–185. doi: 10.1093/nar/gkm321

Murik, O., Tirichine, L., Prihoda, J., Thomas, Y., Araújo, W. L., Allen, A. E., et al. (2019). Downregulation of mitochondrial alternative oxidase affects chloroplast function, redox status and stress response in a marine diatom. *New Phytol.* 221, 1303–1316. doi: 10.1111/nph.15479

Nonoyama, T., Kazamia, E., Nawaly, H., Gao, X., Tsuji, Y., Matsuda, Y., et al. (2019). Metabolic innovations underpinning the origin and diversification of the diatom chloroplast. *Biomolecules* 9:322. doi: 10.3390/biom9080322

Noordally, Z. B., Ishii, K., Atkins, K. A., Wetherill, S. J., Kusakina, J., Walton, E. J., et al. (2013). Circadian control of chloroplast transcription by a nuclear-encoded timing signal. *Science* 339, 1316–1319. doi: 10.1126/science.1230397

Norbury, C. J. (2010). 3' uridylation and the regulation of RNA function in the cytoplasm. *Biochem. Soc. Trans.* 38, 1150–1153. doi: 10.1042/bst0381150

Novák Vanclová, A. M. G., Zoltner, M., Kelly, S., Soukal, P., Záhonová, K., Füssy, Z., et al. (2020). Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid. *New Phytol.* 225, 1578–1592. doi: 10.1111/nph.16237

Osborn, H. L., and Hook, S. E. (2013). Using transcriptomic profiles in the diatom *Phaeodactylum tricornutum* to identify and prioritize stressors. *Aquat. Toxicol.* 13, 12–25. doi: 10.1016/j.aquatox.2013.04.002

Oudot-Le Secq, M. P., and Green, B. R. (2011). Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and Thalassiosira pseudonana. *Gene* 476, 20–26. doi: 10.1016/j.gene.2011.02.001

Oudot-Le Secq, M. P., Grimwood, J., Shapiro, H., Armbrust, E. V., Bowler, C., and Green, B. R. (2007). Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. *Mol. Genet. Genom.* 277, 427–439. doi: 10.1007/s00438-006-0199-4

Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R., and Uberbacher, E. C. (2010). CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* 20, 1574–1584. doi: 10.1093/glycob/cwq106

Parks, M. B., Nakov, T., Ruck, E. C., Wickett, N. J., and Alverson, A. J. (2018). Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *Am. J. Bot.* 105, 330–347. doi: 10.1002/ajb2.1056

Prihoda, J., Tanaka, A., de Paula, W. B. M., Allen, J. F., Tirichine, L., and Bowler, C. (2012). Chloroplast-mitochondria cross-talk in diatoms. *J. Exp. Bot.* 63, 1543–1557. doi: 10.1093/jxb/err441

Rastogi, A., Maheswari, U., Dorrell, R. G., Vieira, F. R. J., Maumus, F., Kustka, A., et al. (2018). Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci. Rep.* 8:4834. doi: 10.1038/s41598-018-23106-x

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., and Bowler, C. (2010). Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* 188, 52–66. doi: 10.1111/j.1469-8137.2010.03371.x

Reja, R., Vinayachandran, V., Ghosh, S., and Pugh, B. F. (2015). Molecular mechanisms of ribosomal protein gene co-regulation. *Genes Dev.* 29, 1942–1954. doi: 10.1101/gad.268896.115

Río Bártulos, C., Rogers, M. B., Williams, T. A., Gentekaki, E., Brinkmann, H., Cerff, R., et al. (2018). Mitochondrial glycolysis in a major lineage of eukaryotes. *Genom. Biol. Evol.* 10, 2310–2325. doi: 10.1093/gbe/evy164

Roncel, M., Gonzalez-Rodriguez, A. A., Naranjo, B., Bernal-Bayard, P., Lindahl, A. M., Hervas, M., et al. (2016). Iron deficiency induces a partial inhibition of the photosynthetic electron transport and a high sensitivity to light in the diatom *P. tricornutum*. *Front. Plant Sci.* 7:14. doi: 10.3389/fpls.2016.01050

Sato, S., Nanjappa, D., Dorrell, R. G., Vieira, F. R. J., Kazamia, E., Tirichine, L., et al. (2020). Genome-enabled phylogenetic and functional reconstruction of an araphid pennate diatom Plagiostriata sp. *CCMP*470, previously assigned as a radial centric diatom and its bacterial commensal. *Sci. Rep.* 10:9449. doi: 10.1038/s41598-020-65941-x

Schober, A., Río Bártulos, C., Bischoff, A., Lepetit, B., Gruber, A., and Kroth, P. G. (2019). Organelle studies and proteome analyses of mitochondria and plastids fractions from the diatom Thalassiosira pseudonana. *Plant Cell Physiol.* 60, 1811–1828. . https://doi.org/10.1093/pcp/pcz097, doi: 10.1093/pcp/pcz097

Shadel, G. S., and Clayton, D. A. (1995). A Saccharomyces cerevisiae mitochondrial transcription factor, sc-mtTFB, shares features with sigma factors but is functionally distinct. *Mol. Cell Biol.* 15, 2101–2108. doi: 10.1128/mcb.15.4.2101

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genom. Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shimizu, M., Kato, H., Ogawa, T., Kurachi, A., Nakagawa, Y., and Kobayashi, H. (2010). Sigma factor phosphorylation in the photosynthetic control of photosystem stoichiometry. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10760–10764. doi: 10.1073/pnas.0911692107

Smith, S. R., Dupont, C. L., McCarthy, J. K., Broddrick, J. T., Oborník, M., Horák, A., et al. (2019). Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nat. Commun.* 10:4552. doi: 10.1038/s41467-019-12407-y

Smith, S. R., Gillard, J. T., Kustka, A. B., McCrow, J. B., Badger, J. H., Zheng, H., et al. (2016). Transcriptional orchestration of the global cellular response of a model pennate diatom to diel light cycling under iron limitation.". *PLoS Genet* 12:e1006490. doi: 10.1371/journal.pgen.1006490

Snel, B., van Noort, V., and Huynen, M. A. (2004). Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucl. Acids Res.* 32, 4725–4731. doi: 10.1093/nar/gkh815

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Taddei, L., Stella, G. R., Rogato, A., Bailleul, B., Fortunato, A. E., Annunziata, R., et al. (2016). Multisignal control of expression of the LHCX protein family in the marine diatom *Phaeodactylum tricornutum*. *J. Exp. Bot.* 67, 3939–3951. doi: 10.1093/jxb/erw198

Takahashi, F., Yamagata, D., Ishikawa, M., Fukamatsu, Y., Ogura, Y., Kasahara, M., et al. (2007). AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19625–19630. doi: 10.1073/pnas.0707692104

Tanaka, A., De Martino, A., Amato, A., Montsant, A., Mathieu, B., Rostaing, P., et al. (2015). Ultrastructure and membrane traffic during cell division in the marine pennate diatom *Phaeodactylum tricornutum*. *Protist* 166, 506–521. doi: 10.1016/j.protis.2015.07.005

Teichmann, S. A., and Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* 20, 407–410. doi: 10.1016/s0167-7799(02)02032-2

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., et al. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693. doi: 10.1126/science.1192002

Valenzuela, J., Mazurie, A., Carlson, R. P., Gerlach, R., Cooksey, K. E., Peyton, B. M., et al. (2012). Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum*. *Biotechnol. Biofuels* 5:40. doi: 10.1186/1754-6834-5-40

Valle, K. C., Nymark, M., Aamot, I., Hancke, K., Winge, P., Andresen, K., et al. (2014). System responses to equal doses of photosynthetically usable radiation of blue, green and red light in the marine diatom *Phaeodactylum tricornutum*. *PLoS One* 9:e114211. doi: 10.1371/journal.pone.0114211

Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., Creasy, T., et al. (2013). Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat. Commun.* 4:2091. doi: 10.1038/ncomms3091

Veluchamy, A., Rastogi, A., Lin, X., Lombard, B., Murik, O., Thomas, Y., et al. (2015). An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*. *Genom. Biol.* 16:102. doi: 10.1186/s13059-015-0671-8

Walker, G., Dorrell, R. G., Schlacht, A., and Dacks, J. B. (2011). Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* 138, 1638–1663. doi: 10.1017/S0031182010001708

Wang, P., Gao, J., Wan, C., Zhang, F., Xu, Z., Huang, X., et al. (2010). Divinyl chlorophyll(ide) a can be converted to monovinyl chlorophyll(ide) a by a divinyl reductase in rice. *Plant Physiol.* 153, 994–1003. doi: 10.1104/pp.110.158477

Wang, Y., Jensen, L., Højrup, P., and Morse, D. (2005). Synthesis and degradation of dinoflagellate plastid-encoded psbA proteins are light-regulated, not circadian-regulated. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2844–2849. doi: 10.1073/pnas.0406522102

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2006). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* 34, 173–180. doi: 10.1093/nar/gkj158

Yu, M., Ashworth, M. P., Hajrah, N. H., Khiyami, M. A., Sabir, M. J., Alhebshi, A. M., et al. (2018). "Evolution of the plastid genomes in diatoms," in *Plastid Genome Evolution*, Vol. 83, eds S. M. Chaw and R. K. Jansen (Amsterdam: Elsevier), 129–155. doi: 10.1016/bs.abr.2017.11.009

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., and Horvath, S. (2010). Weighted gene coexpression network analysis: state of the art. *J. Biopharm. Stat.* 20, 281–300. doi: 10.1080/10543400903572753

Zhao, X., Deton Cabanillas, A.-F., Veluchamy, A., Bowler, C., Vieira, F. R. J., and Tirichine, L. (2020). Probing the diversity of Polycomb and Trithorax proteins in cultured and environmentally sampled microalgae. *Front. Mar. Sci.* 7:3389. doi: 10.3389/fmars.2020.00189

Zheng, Y. T., Quinn, A. H., and Sriram, G. (2013). Experimental evidence and isotopomer analysis of mixotrophic glucose metabolism in the marine diatom *Phaeodactylum tricornutum*. *Microb. Cell Fact.* 12:109. doi: 10.1186/1475-2859-12-109

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.