Prevalence, evolution and cis-regulation of diel transcription in *Chlamydomonas reinhardtii*

Nicholas Panchy*, Guangxi Wu†, Linsey Newton‡, Chia-Hong Tsai‡,§, Jin Chen*§,**, Christoph Benning††, Eva M. Farré*‡,†, Shin-Han Shiu*‡,†,†
Running Title: Cyclic expression in *C. reinhardtii*

Keywords: Green algae, diel expression, transcriptomics, evolution, gene regulation, cis-regulatory element

**Corresponding authors:**
Eva M Farre  
Department of Plant Biology  
Michigan State University,  
East Lansing, MI, 48824  
Email: farre@msu.edu  
Tel: 517-353-5215

Shin-Han Shiu  
Department of Plant Biology  
Michigan State University,  
East Lansing, MI, 48824  
Email: shius@msu.edu  
Tel: 517-353-7196
ABSTRACT

Endogenous (circadian) and exogenous (e.g. diel) biological rhythms are a prominent feature of many living systems. In green algal species, knowledge of the extent of diel rhythmicity of genome wide gene expression, its evolution, and its cis-regulatory mechanism is limited. In this study, we identified cyclically expressed genes under diel conditions in \textit{Chlamydomonas reinhardtii} and found that ~50\% of the 17,114 annotated genes exhibited cyclic expression. These cyclic expression patterns indicate a clear succession of biological processes during the course of a day. Among 237 functional categories enriched in cyclically expressed genes, >90\% were phase-specific, including photosynthesis, cell division and motility related processes. By contrasting cyclic expression between \textit{C. reinhardtii} and \textit{Arabidopsis thaliana} putative orthologs, we found significant but weak conservation in cyclic gene expression patterns. On the other hand, within \textit{C. reinhardtii} cyclic expression was preferentially maintained between duplicates and the evolution of phase between paralogs is limited to relatively minor time shifts. Finally, to better understand the \textit{cis} regulatory basis of diel expression, putative \textit{cis}-regulatory elements were identified that could predict the expression phase of a subset of the cyclic transcriptome. Our findings demonstrate both the prevalence of cycling genes as well as the complex regulatory circuitry required to control cyclic expression in a green algal model, highlighting the need to consider diel expression in studying algal molecular networks and in future biotechnological applications.
INTRODUCTION

Diel (24 hour, day/night periods) cycles dictate physiological changes at different times of day in many organisms. The timing of these physiological oscillations is regulated by a combination of environmental, metabolic and circadian signaling processes (Farre 2012; Kinmonth-Schultz et al. 2013; Song et al. 2013; Fonken and Nelson 2014). For example, circadian clock mutants lead to phase changes under entrained diel conditions (i.e. light-dark cycles) and changes in photoperiod sensitivity (Yanovsky and Kay 2002; McNabb and Truman 2008). Oscillations can be a direct adaptation to environmental cycles, for example restricting photosynthesis and protection against UV radiation to periods of light. Diel cycles also influence biotic responses such as defense mechanisms (Arimura et al. 2008; Goodspeed et al. 2012; Baldwin and Meldau 2013) and mutualistic interactions (Frund et al. 2011; Lehmann et al. 2011). Mechanistically, many of these cycling responses are regulated at the transcriptional level. For example, in the green alga Chlamydomonas reinhardtii, oscillations in starch levels are partially regulated by the cyclic expression of ADP-Glucose pyrophosphorylase (Ral et al. 2006). However, some circadian regulated processes are controlled at the post-transcriptional level (Kojima et al. 2011) and/or by the interaction between transcriptional and post-translational regulation (Kinmonth-Schultz et al. 2013; Song et al. 2013). Early transcriptome analyses of three model organisms, Arabidopsis thaliana, Drosophila melanogaster, and Mus musculus, indicated that between one and ten percent of genes exhibit circadian oscillation with periods of ~24 hr (Doherty and Kay 2010). Moreover, in photosynthetic organisms, 30-90% of genes cycle under diel conditions (Michael et al. 2008; Monnier et al. 2010; Shi et al. 2010; Filichkin et al. 2011). In land plants, about a third of the genes that cycle in light/dark are also circadian regulated (Michael et al. 2008; Filichkin et al. 2011).
Several cis-regulatory elements (CREs) necessary for circadian regulated gene expression have been identified (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008), although it remains an open question how well the identified CREs explain global cyclic expression patterns.

The green alga *C. reinhardtii* has been used extensively to study physiological processes under the control of circadian and/or diel cycle (Mittag et al. 2005; Matsuo and Ishiura 2010). *C. reinhardtii*’s size, short life-cycle, and extensive genetic tool set make it an ideal model organism (Harris 2001) particularly for studies such as experimental evolution from single to multi-cellularity (Ratcliffe et al. 2013) and the genetic engineering of triacylglycerol accumulation in algae (Grossman et al. 2007; Hu et al. 2008; Siaut et al. 2011). *C. reinhardtii* has also been used to study rhythmic responses to light (Bruce 1970), ammonium (Byrne et al. 1992) and nitrogen availability (Pajuelo et al. 1995). However, studies of cyclic expression in *C. reinhardtii* have been limited to single (Mittag et al. 2005; Matsuo and Ishiura 2010) or relatively small sets of genes (Kuchroo et al. 2005). Despite the large evolutionary distance, there are some conserved elements between both the circadian (Corello et al. 2009; Matsuo and Ishiura 2010) and the photoperiodic (Romero and Valverde 2009) oscillators of flowering plants and green algae, raising the question whether and to what extent cyclic expression is conserved. Therefore, a genome wide analysis of cyclic expression in *C. reinhardtii* can provide insight not only into cyclic physiological behavior in green algae, but also how this behavior has evolved in divergent lineages of the *Plantae*. Such an analysis will also be relevant to economically important processes in algae such as oil production.

In this study, we examined gene expression patterns under diel conditions in *C. reinhardtii*. We characterized the prevalence of cycling gene expression in the *C. reinhardtii*
genome and observed that genes involved in distinct biological processes are consistently expressed at certain times during the day/night cycle. We also investigated the conservation of cyclic expression patterns between orthologs in *C. reinhardtii* and *Arabidopsis thaliana*, which diverged ~650-800 million years ago (Sanderson et al. 2004) and the evolution of cycling paralogous genes. Finally, to understand the cis-regulatory basis of diel expression, we identified putative CREs (pCREs) associated with cyclic expression at different phases and investigated how these pCREs can be used to predict cycling gene expression.

**MATERIAL AND METHODS**

*Growth of Chlamydomonas reinhardtii Cultures*

*C. reinhardtii* dw15.1 was grown in TAP (Tris-Acetate-Phosphate) media in flasks without aeration, shaken at 100 rpm, at 22 °C. While the acetate present in this media provides an alternative source of carbon, allowing for *C. reinhardtii* to grow in the dark, prior studies have shown that that the cell cycle (Voigt and Munzner 1987; Davies and Grossman 1994) and other metabolic cycles (Ral et al. 2006) are still synchronized in *C. reinhardtii* grown in acetate-containing media under light/dark cycles. Additionally, the amplitude and phase of cell cycle gene expression in our study and in previous studies where cultures were grown under autotrophic conditions (Bisova et al. 2005) are similar (Figure S1). An initial 200 mL culture was grown to a density of 25 million cells mL⁻¹ in constant light (50 μmol s⁻¹ m⁻²) and used to set up 50 mL cultures of 0.5 million cells mL⁻¹ that were transferred to 12 hr light (50 μmol s⁻¹ m⁻²) and 12 hr dark conditions for 48 hours prior to sampling. Two biological replicates were collected every 3 hours between ZT (Zeitgeber Time, hours since last dawn) 0 and ZT 21. Each sample originated from an independent 50 mL culture. Samples collected during the light to dark
or dark to light transition were taken just prior to change of conditions. For collection, 2 mL of the culture was placed in a 2 mL tube and centrifuged at max speed in at 4°C for 10 min. Amber tubes were used for samples collected during the dark period and the supernatant was removed under weak green light. The pellets were snap frozen in liquid nitrogen. The frozen samples were ground using the Qiagen tissue lyser for RNA extraction.

**RNA-sequencing**

RNA was extracted using the Omega eZNA Plant RNA kit. The RNA was eluted in 50 μL DEPC-H₂O and the concentration was measured using a Nanodrop (Thermo-Fisher). A portion of the RNA was diluted to 1 ng uL⁻¹ to check the RNA Integrity Number (RIN) with a Bioanalyzer (Agilent). All samples had a RIN equal to or greater than 7. Library preparation and sequencing was performed at the MSU-Research Technology Support Facility using the Illumina Tru-Seq Stranded kit with an Illumina HiSeq 2500. Eight samples were sequenced in each lane using a custom bar-coding, but the two biological replicates from the same time point were run in separate lanes. The average number of RNA-Seq reads per sample was 1.81e7 and they ranged between 7.07e6 and 2.58e7. The reads from each of 16 samples (8 time points, 2 samples each time point) were mapped to the *C. reinhardtii* genome (version 4.3 from Phytozome) using Tophat (TRAPPELL et al. 2009) with default parameters except for intron length (min 13, max 8712) and max-multi-hits (1). Gene models on non-chromosomal fragments were not considered. FPKM (Fragments Per Kilobase of transcript per Million mapped reads) per gene was calculated using Cufflinks (TRAPPELL et al. 2010) with parameter –I 8712. A high percentage of reads mapped to the genome: the least mapped sample had 82% of reads mapped and the average of all samples was 85%. Upper quartile normalization was applied to all samples to correct for technical variation as recommend in (BULLARD et al. 2010). The two biological replicates were
appended and used as two consecutive days for downstream analysis. Raw read data is available through the NCBI SRA, BioProject accession [PRJNA264777].

**Identification of Cycling Genes**

Two programs were used to identify cyclic patterns of expression in FPKM data, COSPOT (which is described in (PANDA et al. 2002)) and an application of the discrete Fourier transform (DFT). The DFT has previously been applied to the analysis of cyclic expression using RNA-Seq data (RODRIGUEZ et al. 2013), but our method is based primarily on PRIISM (ROSA et al. 2012). We chose to use both COSPOT and the DFT in conjunction because we found that the combination of methods had superior coverage of known cycling genes without a substantial increase in the expected false positive rate (see Supporting Information).

In our application, we take the discrete Fourier transform of each gene expression vector in the *C. reinhardtii* FPKM data set, converting a set of ‘N’ FPKM values (x) in terms of expression vs. time to new values (y) in terms of expression vs. frequency such that:

$$y_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}$$  \hspace{1cm} (1)

Where $x_n$ is the FKPM value at the nth time point and $y_k$ is the kth frequency component with period $T/k$ where $T$ is the time period spanned by the expression vector. The set of frequency components represents the power spectra of the associated expression vector, that is, the contribution of each periodic cycle to the overall data. In calculating the power spectra of the expression data, we employed a non-windowed application of Welch’s Method (too few data points were present tolerate the loss of information involved in windowing) to average the power spectra over subsets of the expression vector with $T = 24$ hr. This was done to reduce bias in the calculation of the power spectra that might be induced by a particular subset of the expression data at the cost of reducing the overall resolution of the power spectra (though this loss of
resolution was primarily at the extreme ends of the spectra and should not affect our results). Furthermore, the coefficients of each power spectra were normalized prior to averaging using the following equation:

\[
y^*_k = \frac{y_k - y_{\min}}{y_{\max} - y_{\min}}
\]

(2)

Where \(y_{\min}\) is the smallest coefficient of the power spectra and \(y_{\max}\) is the largest. As such, the normalized values, \(y^*_k\), are on the interval \([0,1]\), further reducing the affect that a single subset can have on the average power spectra. The “cyclic score” of each gene is defined as the normalized value of the 24 hour frequency component. This score is equated to a \(p\)-value by randomizing the order of values in each expression vector and scoring the vectors in this random population. For this study, we tested cyclic score thresholds equal to the 5th, 2nd, and 1st percentiles of the score distribution of the randomized data (equal to 0.745, 0.808, and 0.841 respectively) and chose the 2nd percentile as our cutoff for calling cycling genes (equivalent to a \(p\)-value of 0.02). In comparison, the 5th percentile of cyclic score for the set of predicted cycling genes in \(C.\ reinhardtii\) was 1. Additional information about how these thresholds were determined as well as a comparison to COPSPOT can be found in the Supporting Information.

**Clustering Cycling Genes According to Phase**

Cycling genes in \(C.\ reinhardtii\) were first divided by their phase of expression, that is, the Zeitgeber Time (ZT) at which peak expression occurred in the FPKM data set. Within each phase cluster, genes were ordered using hierarchical clustering implemented in R for display purposes. Phase clusters were further broken down using two-rounds of k-means clustering, implemented using custom Python scripts. K-means clustering involves initially selecting “k” random centers in parameter space and assigning genes to clusters based on their distance to the nearest center. The mean of each cluster is then used to define new centers which in turn are used
to redefine clusters; this process is repeated until the clusters converge or the amount change per iteration falls below a specified threshold. The final clusters used for pCRE identification contained 10 to 90 genes. Enrichment of Gene Ontology (GO) terms and pCREs in phase groups was done using the Fisher Exact Test, and the resulting $p$-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg 1995).

**Conservation of cyclic expression and phase of expression amongst duplicate genes**

Gene trees in *C. reinhardtii* were defined using the pipeline described in Zou et al. (2009) using a set of protein domains defined using PFAM (Punta et al. 2012). These domains were extracted from protein sequences and aligned using MAFFT (Katoh et al. 2002), and a phylogeny was inferred using RAxML (Stamatakis 2006) with parameters `-f d -m PROTGAMMAJTT`. Large domain families were divided by building neighbor joining trees with PHYLIP (Felsenstein 2005) and cutting a at distance to root $\geq 0.05$ to create sub-clusters between 4 and 300 genes in size. Domains were mapped back to *C. reinhardtii* genes to infer gene trees. The gene trees, including the divided trees for large domain families, were reconciled with an existing species tree (Moreau et al. 2012) using NOTUNG (Chen et al. 2000). An archive of these gene trees in Nexus (.nex) format has been included as File S2. Branches containing *A. thaliana* and *C. reinhardtii* genes were extracted from the overall tree. The significance of the retention rate of cyclic expression and the phase of cyclic expression was determined by randomly pairing genes in the set of duplicates 100,000 times and comparing retention among actual duplicates to the random population.

**Modeling cycling state divergence of duplicate genes**

The divergence of duplicate genes was modeled using a system of three difference equations with a common rate ‘d’ for the divergence of both cycling and non-cycling duplicates.
and a common rate ‘s’ for the reversion of diverged duplicates back to an identical state.
Duplicate gene pairs were binned according to $K_s$ (width = 0.3), and we assume that the initial frequency of duplicates was the same within each bin (If the initial conditions were significantly different, we would expect to see deviation from the observed frequencies in the model predictions, which was not the case). We then solved for values of ‘d’ and ‘s’ using the observed change between consecutive bins, arriving at a solution with the same qualitative behavior as the observed data. A detailed description of the model can be found in the Supporting Information.

**Identification of putative cis-regulatory elements and phase prediction**

Identification of pCREs in the promoter regions of *C. reinhardtii* genes followed the pipeline described in Zou et al. (2011). Cycling genes were clustered according to phase and expression profile as previously described. For each cycling gene the promoter region, defined as the first 1kb upstream of the transcription start site less any bases which overlap with another gene, was isolated. Six motif finders, AlignAce, MDscan MEME, Motif Sampler, Weeder, and YMF, were used to identify motifs enriched in the promoter region of each phase cluster compared to the promoters of all cycling genes. The resulting motifs were merged using UPGMA to reduce the number of motifs and remove redundant motifs. Merged motifs were mapped back to the *C. reinhardtii* genome using a threshold $p$-value of 1e-05.

The presence or absence of pCREs was used to predict the phase of expression of cycling genes using a Support Vector Machine (SVM) implemented in Weka (HALL et al. 2009). Given a test-set of positive and negative examples defined using n-variables (in this case, presence or absence of pCREs), SVM seeks to define a linear classifier (i.e. a hyperplane in variable space), which best divides positive and negative examples. This classifier is then used to assign subsequent data points to either the positive or negative set. A grid search of two parameters, the
minimum distance between positive and negative groups (C) and the ratio of negative to positive examples in the training set (R), were used to optimize separation and pick the best classifier. The tested range of each parameter was as follows: \( C = (0.01, 0.1, 0.5, 1, 1.5, 2.0) \) and \( R = (0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4) \). Results were validated using 10-fold cross validation, which involved dividing positive and negative examples for each phase into training test sets using stratified random sampling. Each of the 10 test sets was classified by an independent SVM run and the average of the 10 runs was used to score the performance of the parameter set.

**Identifying groups of genes with common expression or common function**

Cyclic genes with common expression were defined using k-means clustering as described above. Cyclic genes with common function were defined as those which shared the same GO annotation. For the purpose of predicting cyclic expression, we used only those GO annotations over-enriched in at least one phase of cyclic expression and where at least 8 annotated genes were over-enriched in the same phase.

**RESULTS**

**Cycling gene expression is extensive in the *C. reinhardtii* genome**

To characterize cyclic expression in *C. reinhardtii*, the expression profiles of 17,114 annotated *C. reinhardtii* genes were defined from samples taken at three hour intervals over two 24-hour time courses (see **Methods**). A gene was defined as cyclically expressed if it exhibited statistically significant, non-random variation at a regular period as identified by either COSPOT or DFT (see **Methods**). The union of predictions for both methods covered 8072 cyclically expressed genes (47.2% of the *C. reinhardtii* genome), which we hereafter refer to as “cycling genes”. Both approaches generated cyclic expression models that correlated with the original
expression data, with an average Pearson correlation coefficient of 0.987 for COSPOT and 0.880 for DFT. The correlation for COSPOT models is higher compared to that of DFT because COSPOT models are fit directly to the overall pattern of the data while the DFT models are based only on variations which occur at a period of 24-hours. Taken together, cyclic variation in gene expression represented the predominant form of non-linear variation in RNA content at both the genome wide and individual gene level.

Cyclic variation can be described using three parameters: period, amplitude, and phase (Figure 1A). Using the fitted models, we inferred the period, amplitude and phase of all cycling genes in the *C. reinhardtii* genome. The distribution of period for our set of cycling genes was centered around 24 hours (+/- 1.10 hours, 95% confidence interval) (Figure 1B, S2A). The amplitude of cyclic expression was highly correlated with mean expression level ($r^2 > 0.7$) and, on average, was only half the size of the mean, indicating that most cycling genes are expressed at some constitutive level even during the trough of the cycle (Figure 1C, S2B). The phase distribution of cycling genes was bimodal with one peak at around ZT 0 (20.6% of cycling genes) and a second around ZT 12 (16.4% of cycling genes), corresponding to the night-to-day and the day-to-night transitions, respectively (Figure 1D, S2C). Our finding concurs with the phase distribution reported for *A. thaliana* and other plant species under diel conditions (MICHAEL et al. 2008; FILICHKIN et al. 2011) as well as a subset of circadian genes in *C. reinhardtii* (KUCHO et al. 2005).

**Phases of cycling gene expression are associated with a succession of biological functions**

Earlier studies have shown that multiple processes in *C. reinhardtii* have specific rhythms, including the expression of key photosystem components (HWANG and HERRIN 1994; JACOBSHAGEN and JOHNSON 1994) and the timing of gametogenesis (JONES 1970). Thus we first
asked which processes tend to be rhythmic by identifying GO terms with an over-represented number of cycling genes. We found that cycling genes were enriched in 44 GO terms, including those related to the chloroplast, photosynthesis, and ribosomal subunits (Table S1). Among these terms, the most striking pattern was that 207 of 252 flagella related genes showed cyclic expression. In particular, 80% of cyclically expressed flagella genes (167 of 207) had peak expression at ZT 21, suggesting that biological functions can be phase specific. To further explore the association between phase and function, cycling genes were assigned to eight “phase clusters” (ZT 0, 3, 6, 9, 12, 15, 18, and 21; Figure 2A) and enrichment of GO categories within each cluster was determined.

We found that 237 GO terms had over-represented numbers of genes in ≥1 phase cluster (Figure 2B). Enrichment values for each term in each phase group can be found in File S3. The greatest number of enriched terms was found in the ZT 21 cluster, just before the night-day transition, (40/237, 16.7%) and the ZT 9 cluster, just before the day-night transition (61/237, 25.7%). We also observed that over-represented GO terms tended to be phase-specific: of all 237 terms, only 19 were enriched in >1 phase and 12 of those were enriched only in two adjacent phases (Figure 2B, S3A). In contrast, the majority of under-represented categories (51%) spanned ≥ 4 phases (Figure S3B). Thus, genes involved in the same process not only tended to be enriched in a particular phase of expression, but were also depleted in other phases. This phase-specificity of functional categories was consistent with previous studies of light-response, metabolism, cell division, and flagellum biogenesis in C. reinhardtii demonstrating cyclic behavior at a specific time of the day (Jones 1970; Cavalier-Smith 1974; Teramoto et al. 2002). For example, DNA replication and mitotic events in C. reinhardtii are restricted to the early hours of the dark period (Jones 1970): not only is the transition into darkness required for
normal cell division (Voigt and Munzner 1987), but DNA replication and cell separation occur between 2-5 hours after the light-dark transition (Fang et al. 2006). This specific timing of DNA replication after the light to dark transition matches the phase of expression for cycling genes related to this process. Alternatively, the gradual increase in expression of replication associated genes towards a peak early in the dark period may track with increases in cell size, as it has been shown that the concentration of cell cycle regulatory proteins HA-MAT3, DP1, and E2F1 remain constant in spite of the increase in cell volume during G1 (Olson et al. 2010). We should note that many of the phase-specific functional categories uncovered here, such as amino acid biosynthesis, phosphorelay activity, and mRNA splicing were not previously known to show time-specific cycling behavior in C. reinhardtii. While correlation alone is insufficient to prove causation, the coordination between cyclic expression and function is highly suggestive that timing of transcription can regulate the timing of higher order biological processes.

Based on the apparent association between phase and function in this as well as in prior studies, GO terms were classified into broad “functional groups”: (1) ribosome and translation, (2) photosynthesis and light response, (3) mitochondria and metabolism, (4) cell cycle and mitosis and (5) microtubules and flagella (Figure 3, Table S2). We found that group 1, 2, and 3 were over-represented in the middle of the day (ZT 3 and 6), group 4 in the early and mid-night (ZT 12 and 15), and group 5 at the end of the dark-period (ZT 21) (Figure 3A). Consistent with the pattern of phase-specific enrichment of genes in different functional groups, the normalized expression profiles of cycling genes in each functional group clearly demonstrated phase specificity (Figure 3B-F). The diel expression data also highlighted the possibility of distinguishing different components of a biological process. For example, group 5 genes are involved in forming microtubules and subsequently flagella. Within this group, genes associated
with the microtubule cytoskeleton peaked earlier in the dark period while those associated with
flagellum assembly peaked toward the end (Figure 3L), representing a clear delineation between
spindle body formation and flagellar regeneration as described previously (CAVALIER-SMITH
1974). Taken together, our findings suggest that the timing of biological processes (translation,
cell-replication, and regeneration of the flagellum) may be determined by transcriptional
regulation.

**C. reinhardtii** and **A. thaliana** orthologs have limited conservation in cycling gene
eexpression patterns

To test if the functional coordination and phase specificity of cyclic expression observed
in **C. reinhardtii** can be found in related multicellular species, cycling genes were identified in **A.
thaliana** using the same methods and cutoff values applied to **C. reinhardtii** on an existing diel
expression data (BLASING et al. 2005). A total of 4945 genes in **A. thaliana** were identified as
cycling (21.7% of the annotated genes), less than half of what was seen in **C. reinhardtii.** This
difference is in part due to a lower sampling density of the **A. thaliana** data (once every 4 hours),
though the overall time span covered was longer (3 days). It is also possible that the mixture of
different cell types in **A. thaliana** samples could mask some rhythmic expression patterns. We
also observed that 992 GO terms in **A. thaliana** were over-represented in ≥1 phases compared to
237 in **C. reinhardtii**, which is likely a function of significantly better annotation (Figure 2C).
Enrichment values for each term in each phase group can be found in File S4.

In contrast to the strict phase-specificity in **C. reinhardtii**, **A. thaliana** group 2 GO terms
(photosynthesis and light response) were enriched amongst cycling genes in all six time points,
but were predominant at ZT 4. The other three groups (group 1, 3, and 4) were restricted to at
most two adjacent phases (Figure 3G). Compared to **C. reinhardtii**, there is a greater variance in
the phase of expression amongst the *A. thaliana* cycling genes within each group, potentially due to the fact that the *A. thaliana* expression data was derived from samples of mixed tissues and cell types. Nonetheless, the peak expression of photosynthetic, mitochondrial, and ribosomal genes occurred at a similar time, as was observed in *C. reinhardtii* (Figure 3H-K). These results suggest that cyclic expression is conserved between a subset of functionally related genes, in both unicellular and multi-cellular plant systems.

Due to the concern that the phase-specificity differences between *C. reinhardtii* and *A. thaliana* might be due to annotation quality difference, we next examined the degree to which cycling gene expression was conserved between orthologous genes in these two species. Among 11,845 putative orthologs, 1,464 (12.4%) showed cyclic expression in both species (referred to as “co-cycling” orthologs), which is significantly higher than the random expectation (Chi-Squared Test, \( p < 0.001 \)). The conserved co-cycling genes encode components of the ribosome (particularly the small subunit), plasma and thylakoid membrane components, or are involved in stress response (Fisher Exact Tests, \( p < 0.05 \)). Nonetheless, we should emphasize that the difference between the observed and expected proportion of co-cycling orthologs was only 2.4%. Thus most cycling genes in *C. reinhardtii* are not cyclic in *A. thaliana* and vice versa. In addition, while the amplitude of cyclic expression is significantly correlated among co-cycling orthologs \( (r^2 = 0.30, p < 10^{-100}) \), there are only weak relationships between their phases \( (r^2 < 0.01, p < 0.006) \). The *A. thaliana* and *C. reinhardtii* lineages diverged 650-800 million years ago (SANDERSON *et al.* 2004) and have extensive differences in life histories, distribution, complexity, and physiology. Thus the conserved components of cyclic expression are likely core processes strongly selected to be maintained, including photosynthetic, mitochondrial, and ribosomal genes (Figure 3H-K). However, most orthologs between green algae and flowering
plants have divergent patterns of cyclic expression, and the extent of cyclic expression divergence highlights the fact that cycling gene expression can be plastic.

**Conservation of cyclic expression is more prevalent amongst older duplicate genes**

To further assess how quickly cyclic expression divergence occurred, we asked how the pattern of cycling gene expression evolved between duplicated genes in *C. reinhardtii*. Gene trees were inferred based on similarity of known protein domains, and we retained only the closest pairs of paralogs (i.e. those separated by only a single ancestral node) for subsequent study (see Methods). The frequency with which the pattern of gene expression (cycling or non-cycling) was identical or divergent was compared to the timing of the inferred duplication event, estimated using the synonymous substitution rate (*Ks*) (**Figure 4A**). The overall frequency of diverged duplicates (one paralog cycling, the other non-cycling) increased with *Ks*, approaching an asymptote of ~0.45 for *Ks* > 0.9. While the frequency of non-cycling duplicates decreased with *Ks*, the frequency of cycling duplicates was greater on average for *Ks* > 0.9 indicating a net gain of cycling expression as duplicates age. We hypothesized that this gain of cycling expression results from a bias in the rate at which duplicate genes diverge that favors the cycling state.

To test this hypothesis, we examined if the observed changes in the frequency of retention can be explained without assuming different rates of divergence. Therefore, a null model of duplicate gene divergence was created using a system of difference equations (see Methods). We fit the transition probabilities using the difference in frequencies between *Ks* 0.6 and 0.9, and the predicted frequencies of identical and divergent duplicates closely matched our observed results at all time points (root mean squared error = 0.03), showing the same pattern of increases and decreases (**Figure S4**). Hence, we have no evidence of a differential rate in
divergence between cycling and non-cycling duplicates, however the predicted probability of transition from identical to divergent (0.42) is less than the probability of transition from divergent to identical (0.53), suggesting that there is a preference for maintaining duplicates in an identical state. This is consistent with our finding that the observed frequency of paralogs with identical states tends to be significantly higher than expected under random association (Z-test, p < 10^{-17}; Figure 4B). In contrast, the frequency of paralogs with divergent state is significantly lower than expected (Figure 4B).

Next, we examined the frequency with which phase is identical amongst pairs of the cycling duplicates. Overall, the number of co-cycling paralogs for which the phase of cyclic expression was identical is more than twice the number randomly expected (Z-test, p < 10^{-85}) with 33.7% of co-cycling duplicates sharing the same phase. The identical phase state was more common amongst cycling duplicates with lower $K_s$ and there was a sharp decrease in the frequency of duplicates with identical phases going from a $K_s$ of 0.9 to 1.2 (Figure 4C). Next we explored if there was a bias in the magnitude of phase change between co-cycling duplicates (Figure 4D). We found that small phase divergences of +/- 3 hr (covering 28.3% of all duplicates) tended to be enriched relative to random expectation, in particular at ZT0/ZT3 and ZT12/ZT15, although the identical phase state is still the most highly enriched scenario. Additionally, there was an inverse, linear relationship between the magnitude of the difference in phase between cycling duplicates and the enrichment of phase-shift events relative to random expectation (all cycling duplicates, $r^2 = 0.91$; duplicates with $K_s > 0.9$, $r^2 = 0.93$), indicating that large differences in phase between duplicates occur less frequently than expected by random chance. Furthermore, we found 33 GO terms enriched (adjusted p-value < 0.05) amongst cycling
duplicates with the same phase, the majority of which (88%) were previously found to be enriched in a specific phase of cyclic expression.

**Cycling genes are enriched for specific putative Cis-regulatory elements**

The coordinated expression of functionally related genes suggests the existence of one or more regulatory mechanisms that drive phase specific expression. While mRNA levels may be affected at multiple levels of regulation, we chose to focus on transcriptional regulation driven by cis-regulatory sequences as circadian rhythm related cis-elements have previously been identified in plant and animal models (Michael and McClung 2002; Ueda et al. 2005; Michael et al. 2008). Using a motif finding pipeline (Zou et al. 2011), we found 687 putative cis-regulatory elements (pCREs) in the 1kb regions upstream of the transcriptional start sites of cycling genes for each of the eight C. reinhardtii phase clusters (Fisher Exact Test, adjusted $p < 0.05$). The top enriched motifs for each phase can be found in Figure 5, and the entire list of enriched motifs can be found in File S5. Each phase had 60-84 associated pCREs, except for ZT 15 with 169; however, more than 20% of pCREs (141/687) were enriched in >1 phase and 43.8% of ZT 15 pCREs (74/169) were enriched in ≥1 other phases (mostly ZT 12; Figure 6A).

Therefore, each pCRE was assigned to the phase cluster in which it was most significantly enriched.

To further assess whether the pCREs are meaningful, they were used to establish classifiers to predict cyclic expression in different phases. First, the pCREs assigned to each phase were used to predict which genes are cyclic in a naïve manner. That is, for pCREs enriched in a particular phase, we simply predicted that all genes with ≥1 pCREs mapped to their promoters would cycle at that phase. The performance of these predictions was assessed using the area under the receiver operating curve (AUC-ROC), a metric which quantifies the ability of
a method to predict positive examples which, in our case, is phase specific expression. Perfect predictors have an AUC-ROC of 1 while random guessing has a value of 0.5; our naïve classification of phase had AUC-ROCs that ranged from 0.58 (ZT 9) to 0.62 (ZT 12) indicating that this simple classification procedure performed marginally better than randomly assigning phase (Figure 6B). The same conclusion can be reached based on the F-measure, another prediction performance metric (Figure 6C). Next, to further improve the prediction of the phase of cyclic expression, we used the support vector machine (SVM) algorithm to classify cycling genes according to the presence or absence of all pCREs (see Methods). The SVM classifier shows improved performance compared to naïve classification (Figure 6B-C, Table S3) but AUC-ROC values are still relatively low, ranging from 0.58 (ZT 9) to 0.65 (ZT 18) (Figure S5). We also identified two pCRE association rules enriched in specific phases of cyclic expression using CBA (Liu et al. 1998); however, adding these rules to the SVM prediction models did not significantly improve the overall predictive power of our pCREs as the AUC-ROC increased by at most 0.01.

Given that the C. reinhardtii pCREs are computationally derived, we next asked how well a known, experimentally verified, phase-specific cis-regulatory element may predict cyclic expression. For this purpose we examined the Evening Element that is necessary and sufficient to drive circadian expression in A. thaliana (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008). Using motifs related to the Evening Element identified in Michael et al. (2008), we generated a cycling gene classifier to predict the phase of the 4,945 A. thaliana cycling genes introduced in the earlier phase-specificity comparison section. The optimal AUC-ROC of the Evening Element classifier was 0.57 and 0.56 at ZT 0 and 12 hours, respectively (compared to 0.58-0.65 in C. reinhardtii pCRE predictions). Therefore, although the Evening
Element is known to function as a circadian regulator, similar to *C. reinhardtii* pCRES, it has only limited predictive power on a genome wide scale. To obtain accurate predictions the presence or absence of pCREs needs to be supplemented with additional information regarding the regulation of cycling expression.

**Phase of cyclic expression can be predicted for groups of genes with common expression patterns or common function**

The weak predictive power of pCREs likely results from an underlying complexity in the regulation of the phase of cyclic expression, either in the form of additional control mechanisms or the existence of more discrete regulatory groups. Timing of cyclic expression may be modified by interactions amongst regulatory motifs or post-transcriptional mechanisms. It is also possible that our phase clusters might consist of multiple regulatory subgroups. To address the latter possibility, we further classified genes in each phase group into sub-clusters containing genes with highly similar expression profiles (phase-expression clusters). Using SVM, 28 of 190 phase-expression clusters covering 584 genes (7.23% of cycling genes) could be classified with an AUC-ROC > 0.7 (these clusters are described in **File S6**), which is better than any individual phase alone. The best predicted phase-expression clusters do not necessarily have stronger cyclic signals (**Figure 7A**) compared to the worst predicted (**Figure 7B**). Additionally, we eliminated size ($r^2 = 0.15$) and the correlation of expression profiles within each phase-expression cluster ($r^2 < 0.01$) as possible variables explaining the observed variance in AUC-ROC (**Figure S6**). These results suggest that phase specific regulation does occur at the *cis*-regulatory level for particular groups of cycling genes and that presence or absence of pCREs alone is sufficient to accurately predict the pattern of phase specific expression for these clusters. Those pCREs which were
informative (i.e. had the non-zero weights) when predicting the 28 best phase-expression clusters are listed in File S7.

In addition to using highly similar expression patterns as a way of subdividing phase clusters, we looked for evidence of phase specific regulation amongst groups of genes in the same phase cluster that had related annotated function (phase-function clusters). Among 71 phase-function clusters, genes belonging to 12 of these clusters could be classified with an AUC-ROC > 0.7. These clusters covered 12.2% (175/1434) of genes present in all phase-function clusters, a higher percentage than the phase-expression clusters, although they constitute a smaller portion of all cycling genes due to limited GO annotation in C. reinhardtii (these clusters are described in File S8). Genes in most of these functional groups displayed a clear cyclic signal (Figure 7C), except for the groups related to the nucleolus and cell wall, which were predominantly non-cyclic genes but had a statistically significant subset of phase-specific genes. Amongst the best classified sub-clusters contained genes relating to the large cytosolic ribosomal subunits (AUC-ROC = 0.73), cilium (0.72), small cytosolic ribosomal subunit (0.72), translation (0.71), and the chloroplast (0.68). This supports our earlier observation that the cyclic patterning of large scale processes such as photosynthesis, translation, and motility may be regulated at the transcriptional level. The pCREs which had non-zero weights when predicting the 12 best phase-function clusters are listed in File S9.

DISCUSSION

We have determined that cyclic expression is prevalent in the C. reinhardtii genome, and nearly half of all annotated genes cycle under diel conditions. There is a strong link between rhythmic patterns at the molecular and physiological levels. Diel cycling expression is influenced
both by environmental factors, such as the availability of light, and endogenous factors, including metabolism and the circadian clock (Farre 2012; Kinmonth-Schultz et al. 2013; Song et al. 2013; Fonken and Nelson 2014). While the importance of photoperiod can be inferred for light-dependent (i.e. photo-synthesis) and light-sensitive (i.e. DNA replication) processes, for most cycling related functions it remains an open question as to what extent each factor influences cycling expression. This is particularly true of functions which were not previously known to exhibit cycling expression in green algae, for example, the regulation of RNA processing and amino-acid synthesis.

In addition to the relationship between cyclic expression and gene function, we found that cyclic expression was significantly conserved between paralogous genes. The proportion of divergent duplicates reaches an asymptote at Ks > 0.9, which is similar to what was previously observed for stress responsive duplicate genes (Zou et al. 2009). However, while there appears to be a clear preference for the partitioning of ancestral expression states in stress responsive genes (Zou et al. 2009; Dong and Adams 2011), we found that duplicates genes tend to share the same expression state with respect to cycling and that cycling duplicates preferentially retain the same or similar phase of expression. We hypothesize this pattern of cyclicity/phase conservation among duplicates points to a fundamentally distinct regulatory logic from that of stress response. In stress response, a duplicate which has lost response to one condition may still be responsive to other conditions and thus retained. However, either loss or gain of cyclicity in a duplicate gene would mean it is no longer temporally in sync with other genes in the processes which it was originally involved in. For example, if a replication initiation factor duplicate was not in sync with the expression of other components of the replication machinery, the duplicated factor would not be functional and eventually eliminated from the genome. This argument may
also apply to the conservation of phase among duplicate cycling genes. Indeed, we found that most GO terms enriched amongst co-cycling duplicates with the same phase were highly phase specific, including those associated with DNA replication and flagellar components.

Based on prior studies of stress response genes (Zou et al. 2009), we expected that the conservation of cycling expression state, particularly the phase of expression, would be correlated with the presence of shared cis-regulatory elements. However, contrary to this expectation, the set of putative cis-regulatory elements enriched in cycling genes does not accurately distinguish phase expression. While our results suggest that cis-regulation plays a significant role in controlling cyclic expression in C. reinhardtii, the presence or absence of promoter elements alone was insufficient to fully explain the observed patterns of cyclic variation across the entire C. reinhardtii genome. This suggests that additional regulatory components are involved in controlling cyclic expression. In other organisms the combinatorial interactions amongst regulatory factors play an important role in controlling the phase of cyclic gene expression (Harmers and Kay 2005; Ueda et al. 2005), but in C. reinhardtii there is evidence that response to changing light levels is mediated by multiple copies of the same or similar promoter elements (Von Gromoff et al. 2006). While we did not see significant improvement when rules considering combinatorial relationships between pCREs were included in our model, this may be due to the fact that we were able to explore only a subset of all possible combinatorial interactions in our pCRE set. Additionally, post-transcriptional regulation has been implicated in regulating circadian processes in Neurospora crassa, A. thaliana, and D. melanogaster (Kojima et al. 2011). In C. reinhardtii, the over or under expression of the RNA-binding protein CHLAMY1 is known to result in the disruption or loss of circadian rhythms.
Further studies incorporating post-transcriptional regulatory features will be necessary to improve the prediction of phase specific cyclic expression.

The inability of pCREs to classify phase specific cycling expression on a genome wide scale does not contradict prior observations that certain cis-elements are necessary for cycling expression (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008). Rather it suggests that cis-elements alone are insufficient to explain the variation in cycling expression on a genome wide scale and that additional regulatory components remain to be discovered. Post-transcriptional regulatory mechanisms and chromatin state are two promising avenues of investigation which, in conjunction with the cis elements we have identified, could be used to better predict the state of cycling expression. Although there remains substantial room for further improvement, our findings contribute to a better understanding of both the function and evolutionary origins of cyclic expression in a green alga, laying the foundation for further molecular dissection of the relationships between the rhythmic gene expression and physiological functions for potential biotechnological applications.

ACKNOWLEDGEMENTS
We thank Alexander Seddon and John Lloyd for help in identifying pCREs. We also thank all the members of the algae group at Michigan State University for their attention and critique. This work was supported by a Michigan State University Strategic Partnership Grant to C.B, E. F., and S.-H. S. and a National Science Foundation grant (MCB-1119778) to S.-H. S.

REFERENCES


LIU, B., W. HSU and Y. MA, 1998 Integrating Classification and Association Rule Mining Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.


**FIGURE LEGENDS**

**Figure 1:** Period, amplitude, and phase of cyclic expression. (A) Three properties of cyclic variation: period, amplitude, and phase. (B) The distribution of period of cycling genes identified in *C. reinhardtii*. (C) The relationship between amplitude and mean expression level in FPKM (Fragments per Kilobase of transcript per Million mapped reads). (D) The distribution of the phase of cycling genes.
**Figure 2:** Phase of gene expression and cyclically expressed GO terms. (A) The normalized (relative) expression of each cycling gene in *C. reinhardtii* (each row) across the 48-hour period (columns). Genes were assigned to phase clusters based on the predicted time of peak expression. Genes in each phase cluster were ordered using hierarchical clustering. The white and black bars below indicate samples from the light and the dark periods, respectively. (B) The test statistics of GO term (rows) enrichment in each phase (columns) in *C. reinhardtii*. The -log(*p*-value) of the Fisher exact test is plotted. GO terms are ordered along the y-axis according to the most enriched phases and hierarchically clustered within each phase. (C) The test statistics of GO term enrichment in each phase in *A. thaliana*. Methods for assigning GO terms to phase, clustering, and the color legend are the same as in (B).

**Figure 3:** Phase specific expression of broad functional categories. (A) Enrichment test statistics in each functional group (row) and in each phase cluster (column) among *C. reinhardtii* cycling genes. The color indicates the averaged –log(*p*-value) of GO terms in a functional group (Table S4). (B-F) Normalized expression profiles of genes in each functional group in *C. reinhardtii*. The black line indicates average expression values. The grey area represents plus/minus one standard deviation. (B) Ribosomes/Translation (C) Photosynthesis/Light-response (D) Mitochondria/Metabolism (E) Cell-cycle/Mitosis (F) Microtubules/Flagella (G) Enrichment test statistics for functional groups in *A. thaliana*. The functional group designation and color legends are the same as (A). Gray: not applicable. (H-K) Normalized expression profiles of genes in each functional group in *A. thaliana*. (H) Ribosomes/Translation (I) Photosynthesis/Light-response (J) Mitochondria/Metabolism (K) Cell-cycle/Mitosis (L) Expression profiles of genes in the
microtubule cytoskeleton (red), flagellum assembly (blue) and cell projection organization (black) categories.

**Figure 4:** Conservation of cyclic expression and phase of cyclic expression. (A) The frequency at which duplicate pairs of genes in *C. reinhardtii* maintain cycling expression, maintain non-cycling expression, or diverge as a function of the synonymous substitution rate (Ks). (B) Distribution of cycling retention, non-cycling retention and divergence between duplicate pairs in random simulations. The black bars cover the inter-quartile range of each distribution, and error bars represent the 95% confidence interval. Observed values are indicated by asterisks. (C) The frequency at which the phase is retained in pairs of cycling duplicates as a function of Ks. (D) Enrichment values for phase retention (diagonal values) and phase change (off diagonal values) between actual duplicates and duplicate pairs in random simulations.

**Figure 5:** Top three pCREs enriched in each phase cluster of cyclic genes. Sequence logos representing the top three putative *cis* regulatory elements (pCREs) enriched in each phase cluster of cycling genes in *C. reinhardtii*.

**Figure 6:** Enrichment and performance of phase-specific pCREs. (A) The enrichment test statistics of 687 pCREs (rows) in genes of each phase cluster and non-cyclic (NC) genes (columns). (B) The area under the curve of the receiver operating characteristic (AUC-ROC) for phase expression prediction with naïve (green) and Support Vector Machine (SVM, white) classifiers. (C) The F-measures for phase expression prediction based on random guess (black), naïve (green) and SVM (white) classifiers.
**Figure 7:** Expression of best predicted co-expression cluster and GO terms. (A) Averaged, normalized expression profile of genes in the top 28 co-expression clusters whose phase of expression can be predicted with AUC-ROC > 0.7. (B) Averaged, normalized expression profiles of genes in the bottom 28 co-expression clusters whose phase of expression can be predicted with AUC-ROC < 0.56. (C) Averaged, normalized expression profiles of genes in the 12 GO terms whose phase of expression can be predicted with AUC-ROC > 0.7. Both cycling and non-cycling genes annotated in each GO term are included.

**Figure S1:** Expression profiles of cell cycle genes (MAT3, E2F, CDKA1, and CDKB1) in *C. reinhardtii* grown in TAP (Tris-Acetate-Phosphate) culture. As observed in previous studies of *C. reinhardtii* grown on autotrophic conditions (BISOVA et al. 2005) MAT3, CDKA1, and CDKB1 are most highly expressed between 12 and 18 hours after dawn, while E2F expression increases slightly earlier (between 6 and 9 hours).

**Figure S2:** Period, amplitude, and phase of cyclic expression amongst predictions made by COSPOT, DFT, and both methods combined. (A) The distribution of the period of expression in cycling genes predicted by COSPOT (red), DFT (yellow) and both methods combined (blue). (B) Mean expression (x-axis) vs. the amplitude of cyclic expression (y-axis) of cycling genes. Color labels follow (A). (C) Phase of expression of cycling genes. Color labels follow (A).

**Figure S3:** Most over- and under-enriched GO terms amongst phase clusters of cycling genes. (A) Heatmap showing the –log10 transformed Fisher’s exact test *p*-values (pval) of the top five
GO terms with over-represented numbers of genes in each phase cluster (ZT 0, 3, 6, 9, 12, 15, 18, and 21). (B) Heatmap showing transformed \( p \)-values of GO terms with under-represented numbers of genes in at least one phase cluster (same as in (A)). \( p \)-values were calculated and transformed as in part (A) except that the left-tail \( p \)-value was used.

**Figure S4:** Divergence of duplicate gene expression state modeled as a system of difference equations. The frequency at which duplicate pairs in *C. reinhardtii* are both cycling (blue), both non-cycling (red), or divergent expression (orange) as a function of the synonymous substitution rate (\( Ks \)). The difference equations used to generate these data are described in the Supporting Information.

**Figure S5:** Precision-recall and AUC-ROC curves of SVM predictions for *C. reinhardtii*. (A) Precision-recall curves for the prediction of the each of the eight phase clusters in cycling genes in *C. reinhardtii* as classified using SVM. Each phase-cluster is represented as a different colored line: 0 (red), 3 (orange), 6 (lime), 9 (green), 12 (teal), 15 (blue), 18 (purple), 21 (pink). Error bars represent the variance in 10 separate runs of the SVM classifier at optimal parameters. (B) ROC curves for the prediction of each of the eight phase clusters in cycling genes in *C. reinhardtii* as classified using SVM. Line color and error bars are assigned as in (A).

**Figure S6:** Regression of the AUC-ROC of phase-expression clusters against cluster size, and Pearson Correlation Coefficient (PCC) of genes in the cluster. (A) Plot of phase-expression cluster size against AUC-ROC. The black line indicates the best linear regression of AUC-ROC against cluster size. The equation is reported above the figure. (B) Plot of the mean PCC amongst
genes in each phase-expression cluster against AUC-ROC. The black line indicates the best linear regression of AUC-ROC against mean PCC. The equation is reported above the figure. (C) Plot of the standard deviation of PCC amongst genes in each expression cluster against AUC-ROC. The black line indicates the best linear regression of AUC-ROC against standard deviation of PCC. The equation is reported above the figure.
A. Diagram showing the concept of expression level, phase, period, and amplitude over time.

B. Histogram showing the frequency distribution of period (hours).

C. Scatter plot of log(amplitude FPKM) vs. log(mean expression FPKM).

D. Histogram showing the frequency distribution of phase (hours).
A. Co-expression clusters

B. Co-expression clusters

C. GO terms

- phosphorus-oxygen lyase activity
- cyclic nucleotide biosynthetic process
- translation
- cytosolic small ribosomal subunit
- rRNA binding
- chloroplast thylakoid lumen
- cytosolic large ribosomal subunit
- cell wall
- nucleolus
- photosystem 1
- cilium
- cell projection organization

Relative expression

0 0.5 1