Supporting Information

Vijayan et al. 10.1073/pnas.0912673106

ST Text

Breakdown of Worksheets contained in Dataset S1 is listed below.

Worksheet 1: Analysis of Circadian Gene Exp. Column A: Joint Genome Institute (JGI) ORF ID. Column B: 7942 ID. Columns C through F: Location of ORF on chromosome, pANL (large endogenous plasmid), or pANS (small endogenous plasmid) from the respective GenBank files CP000100, CP000101, and S89470. Columns G through I: Average GC content along coding region and promoter regions. Columns J through N: Log_2(amplitude/2), phase, period, cluster, and classification (subjective dawn, subjective dusk, and noncircadian) of all predicted ORFs. Column O: Temporal gene expression correlation between ORF and its clockwise neighbor. Column P: Temporal gene expression correlation between ORF and supercoiling waveform.

Worksheet 2: KEGG Designations. Columns A: Joint Genome Institute (JGI) ORF ID. Columns B: KEGG pathway designation. Columns C: ORF classification (subjective dawn, subjective dusk, and noncircadian).

Worksheet 3: KEGG Pathway *p* values. Columns A and J: KEGG pathway designation. Columns B through I: KEGG pathway statistics.

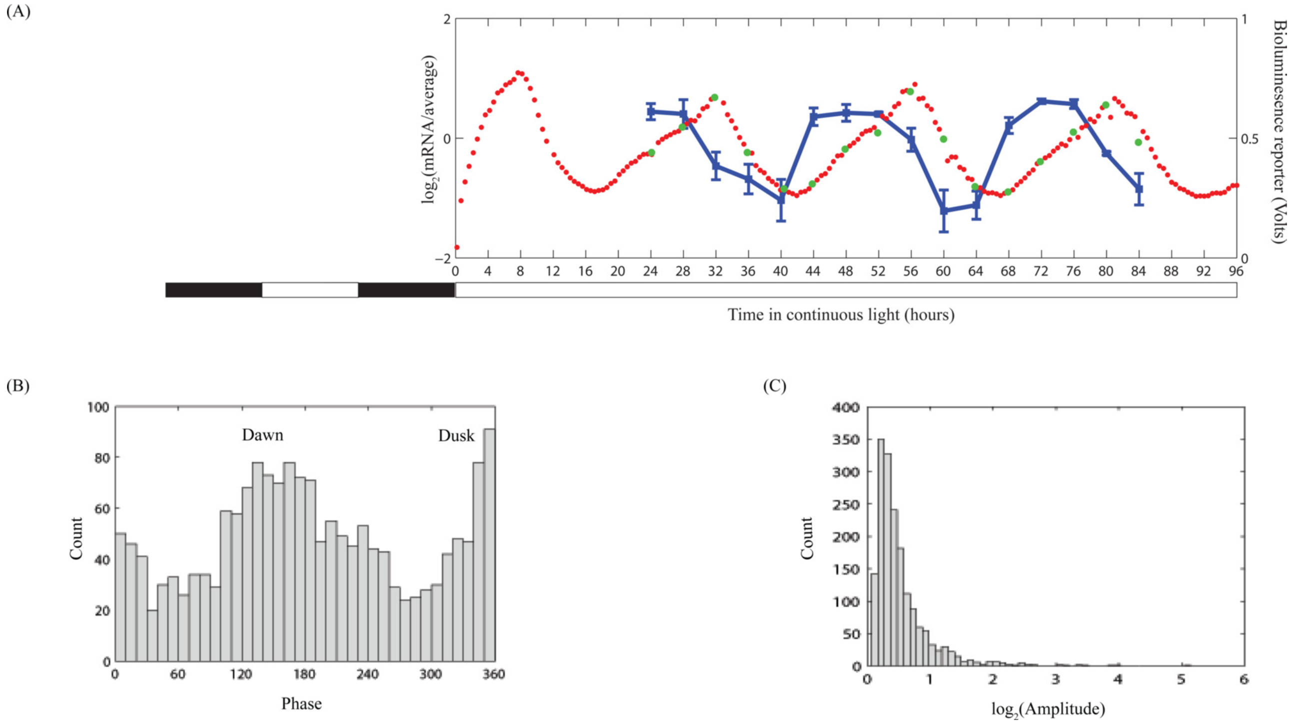


Fig. S1. Circadian gene expression in *Synechococcus elongatus*. (*A*) *S. elongatus* strain AMC 408 was subjected to two consecutive alternating 12-h light–dark cycles for entrainment and subsequently released into continuous light at T = 0 h. A bioluminescence reporter (bacterial luciferase) under the control of the *purF* promoter was monitored continuously as an indicator of oscillation phase (red dots). Cells were sampled at 4-h intervals between T = 24 and T = 84 h (green dots), and gene expression was measured by microarray. The mRNA levels for the *purF* gene are shown in blue with the standard deviation on each side of the mean from four separate microarray probes corresponding to the gene. (*B*) Phase distribution of circadian genes. Gene expression primarily consists of two phases—subjective dawn and subjective dusk. (*C*) Amplitude distribution of circadian genes. Most circadian genes oscillate with low amplitude. Only 186 of 1,748 circadian genes have amplitude of 2-fold or greater.

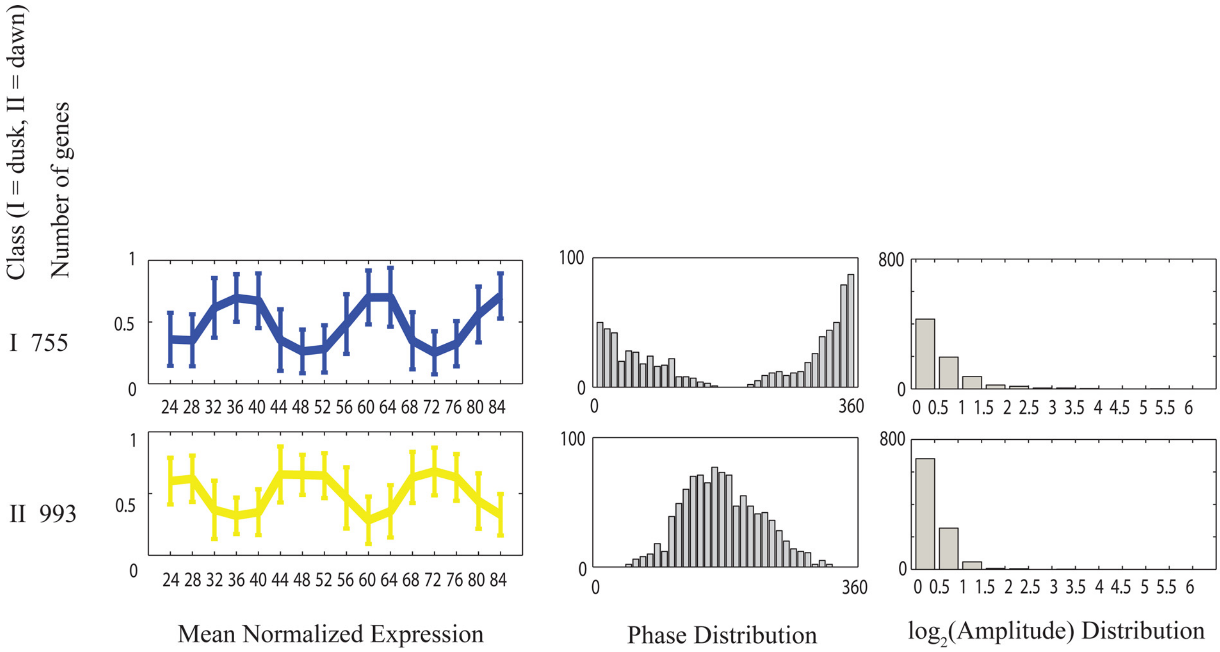


Fig. S2. K-means clustering for analysis of circadian gene expression. Subjective dawn and subjective dusk genes were identified by K-means clustering. Temporal expression profiles were clustered using K = 6 by Euclidean distance. Each generated cluster peaked at a unique time during the circadian cycle. Subjective dawn genes were defined as genes whose mRNA levels peaked at 20, 24, or 4 h in the circadian cycle, and subjective dusk genes were defined as genes whose mRNA levels peaked at 8, 12, or 16 h in the circadian cycle. Designations for each individual gene are specified in Dataset S1.

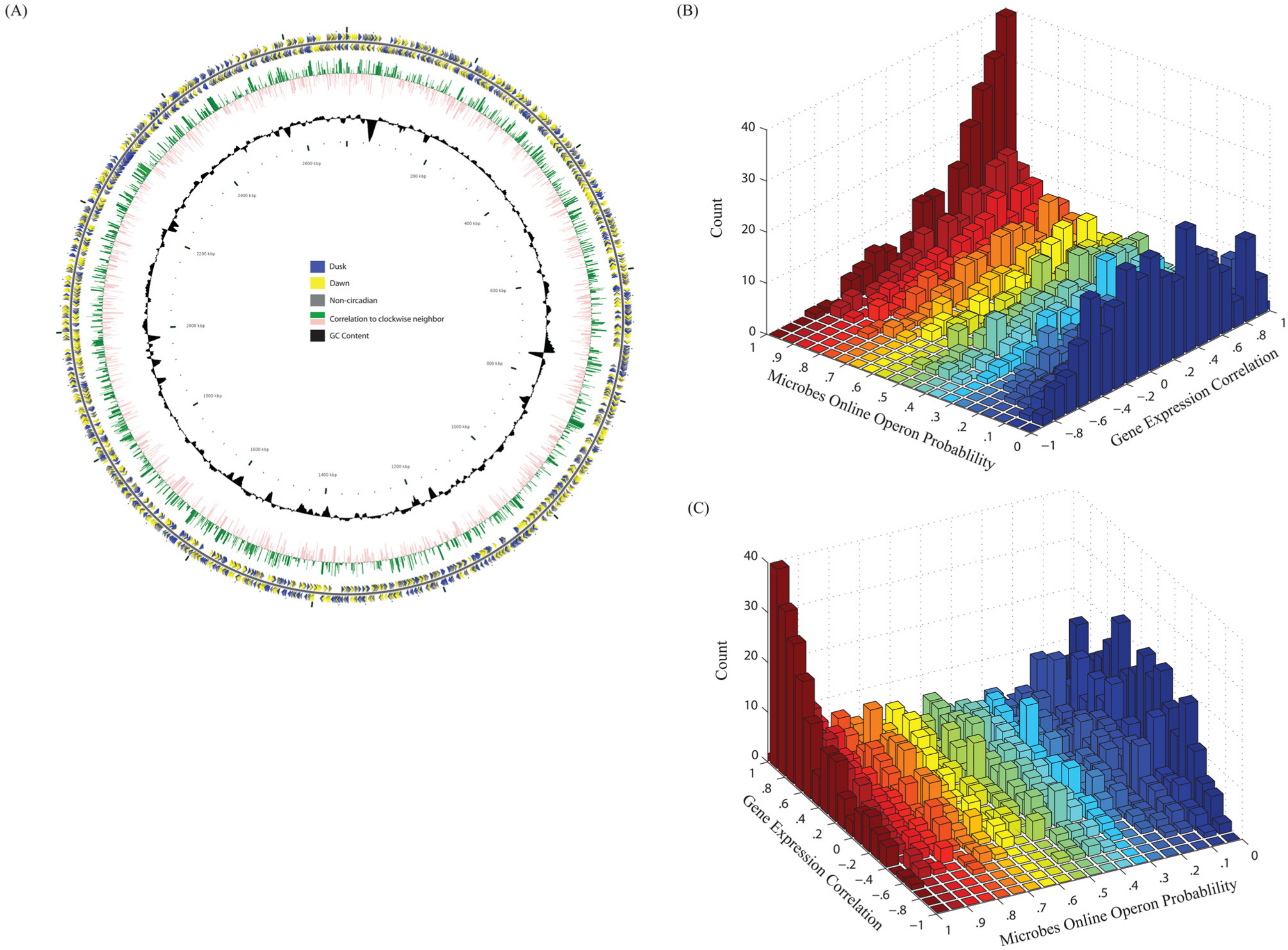


Fig. S3. Spatial organization of gene expression phase. (*A*) Classification (subjective dawn, subjective dusk, or noncircadian), Pearson correlation of temporal gene expression profile to clockwise neighbor, and GC content along the *S. elongatus* chromosome. Negative and positive correlations are colored in green and pink, respectively. Locations are based on GenBank file CP000100 and chromosome diagram was made with CGView (1). The overall phase distribution along the chromosome appears random although adjacent genes are often highly correlated in expression. (*B* and *C*) are two separate views of the same 2-D histogram comparing temporal gene expression correlation (Pearson) between neighboring genes and the probability of existence on the same operon. Operon probabilities were calculated by MicrobesOnline and are determined by four metrics: (*i*) the distance between genes; (*ii*) whether the genes are near each other in other genomes; (*iii*) whether genes both belong to a narrow GO category; and (*iv*) whether genes share a COG functional category (2). Neighboring genes with low probability of being in the same operon tend to have an equal chance of being negatively or positively correlated with each other. Similarly, most neighboring genes that have a high positive correlation in gene expression tend to have a high probability of existing on the same operon. Together, this suggests that the organization of phase along the *S. elongatus* chromosome is relatively random after operon structure is taken into account.

1. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. *Bioinformatics* 21:537–539.
2. Alm EJ, et al. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res* 17:1015–1022.

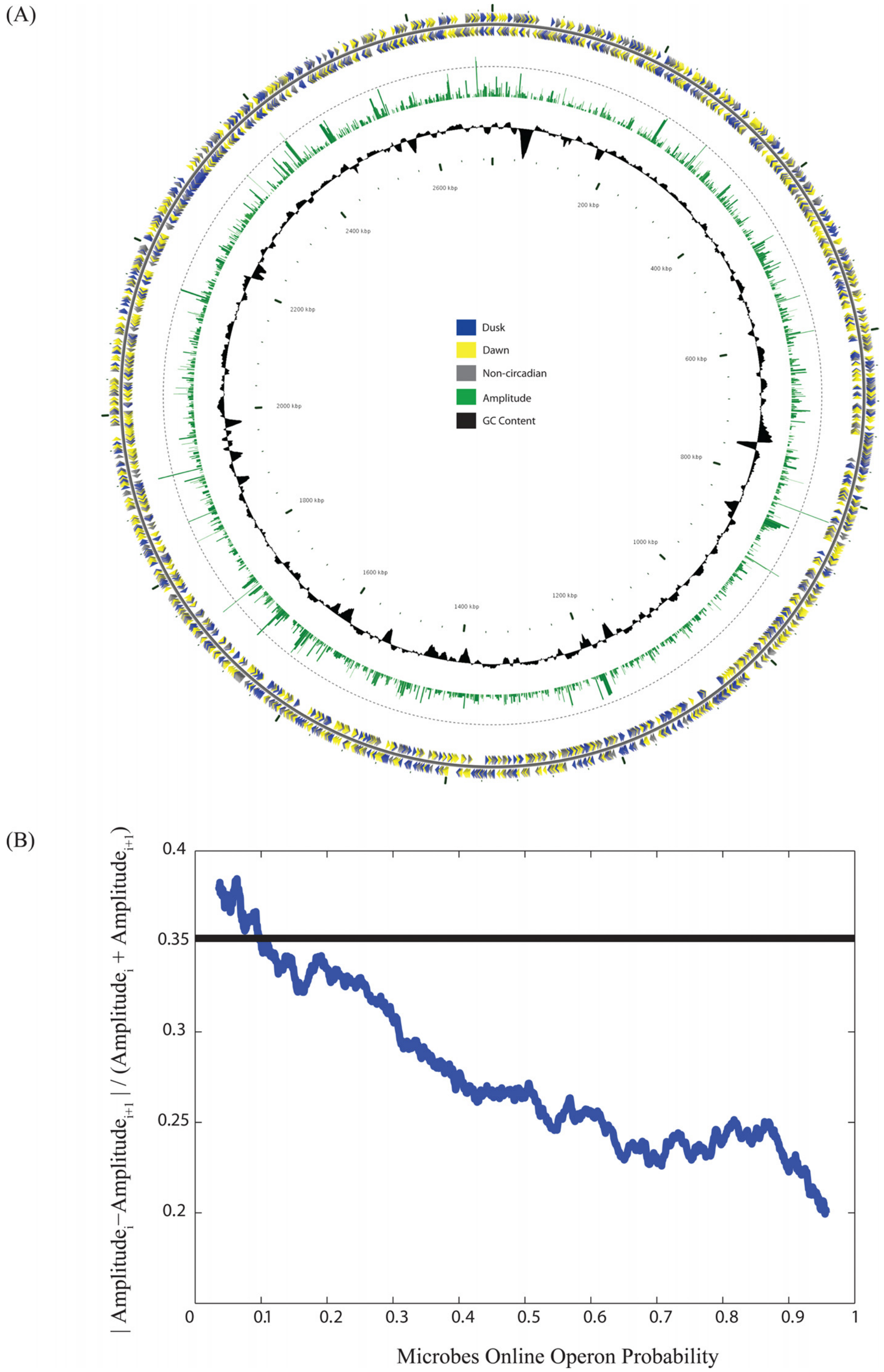


Fig. S4. Spatial organization of gene expression amplitude. (*A*) Classification (subjective dawn, subjective dusk, or noncircadian), amplitude, and GC content along the *S. elongatus* chromosome. Locations are based on GenBank file CP000100 and chromosome diagram was made with CGView (1). The overall amplitude distribution along the chromosome appears random. (*B*) Normalized amplitude difference between neighbors versus MicrobesOnline operon probability (smoothing = 200, blue). In black is the average normalized amplitude difference between two randomly selected genes (mean of 100,000 repetitions). As the probability of sharing the same operon decreases, the amplitude difference between neighbors becomes more random. That is, neighboring genes that are on different transcripts tend to not be correlated in amplitude.

1. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. *Bioinformatics* 21:537–539.

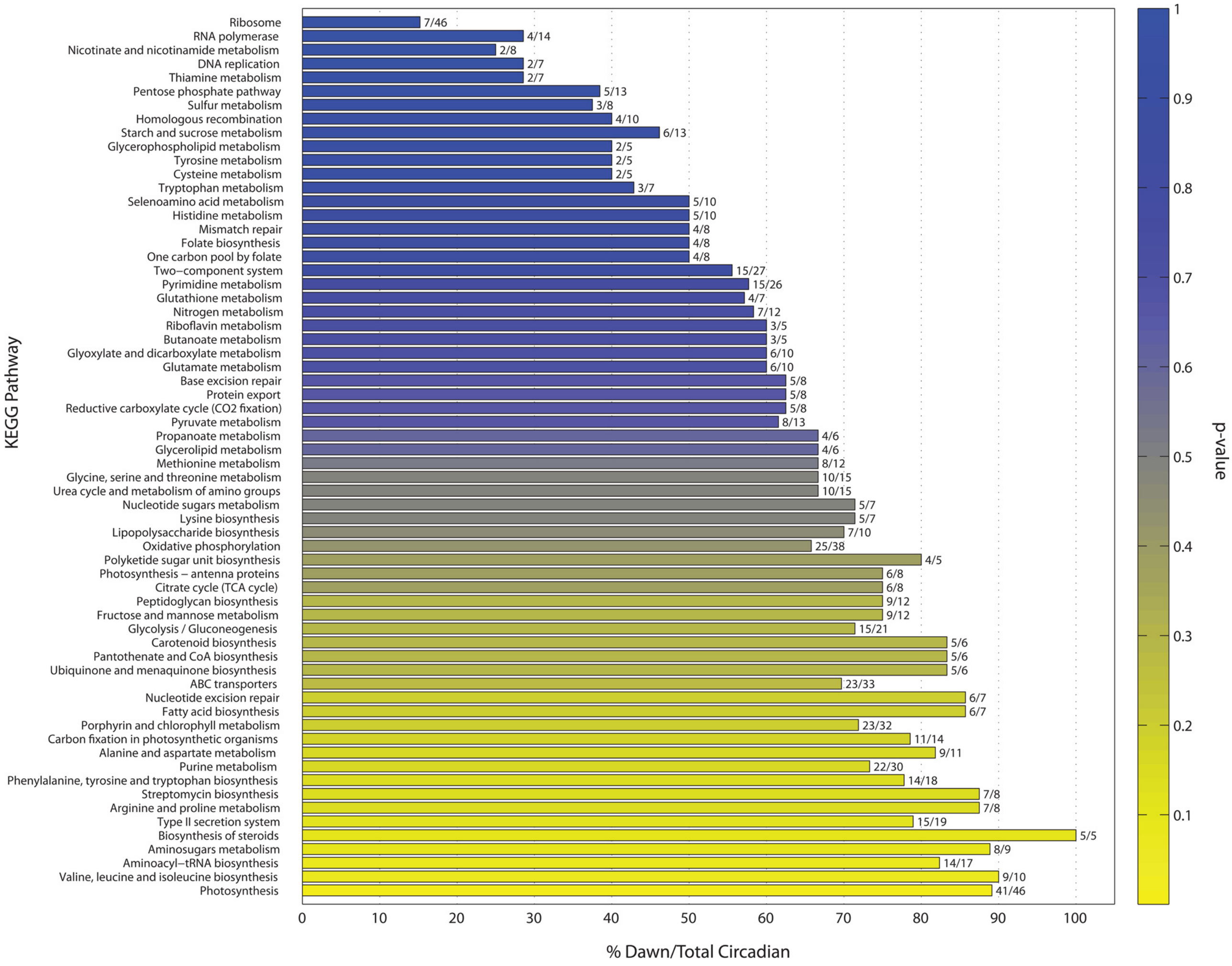


Fig. S5. Functional significance of circadian gene expression. Several KEGG categories are highly enriched for subjective dawn or subjective dusk genes suggesting a physiological role for circadian gene expression. KEGG groups are colored by significance of enrichment (yellow = dawn, blue = dusk). The enrichment for photosynthesis genes ($P = 6.9e\text{-}5$) and ribosomal protein genes ($P = 3.4e\text{-}11$) in the subjective dawn and subjective dusk are particularly striking. Numbers at the end of each horizontal bar indicate fraction of subjective dawn genes to the total number of circadian genes within a KEGG category. Details including mapping of genes to KEGG categories are shown in Dataset S1. *P* values were calculated using the cumulative hypergeometric distribution.

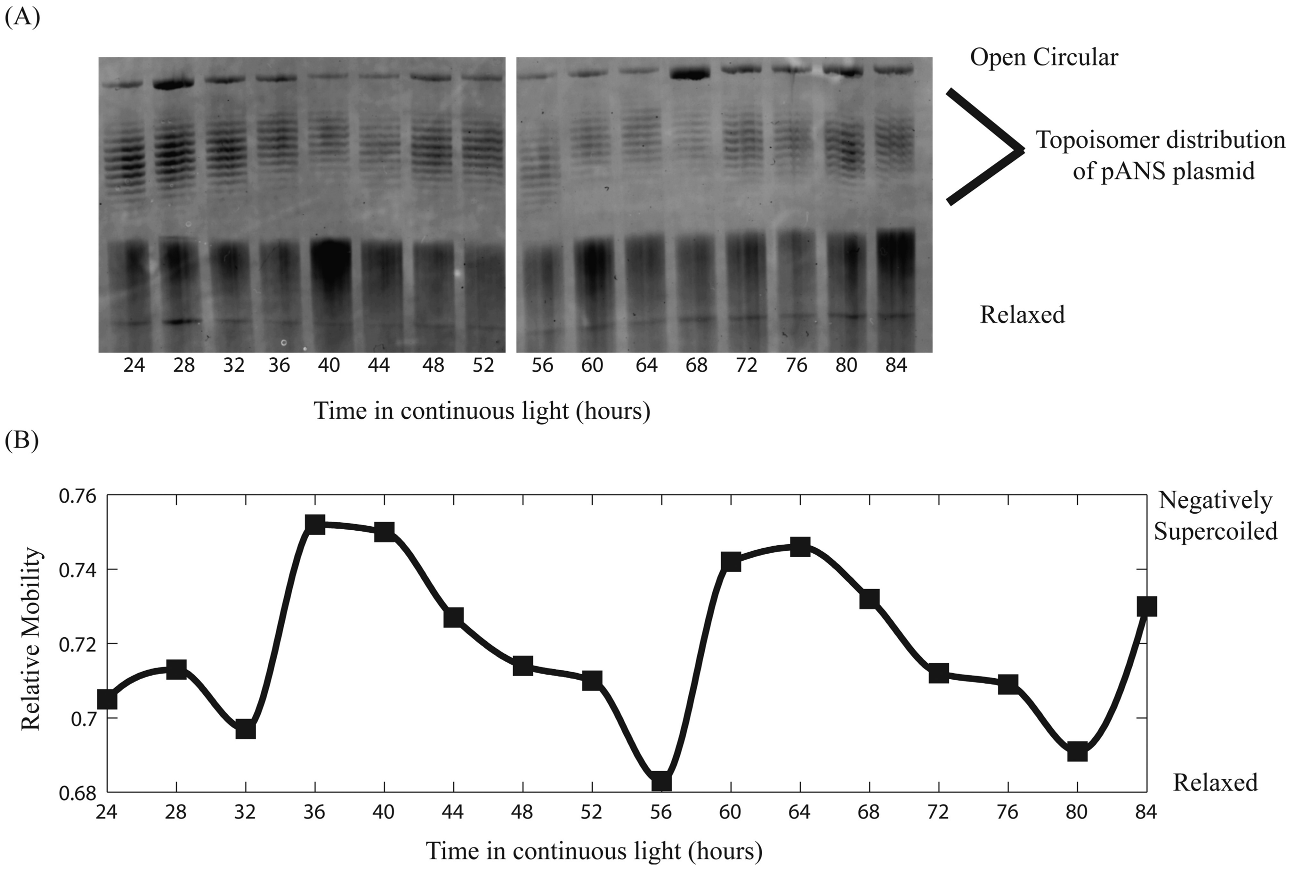
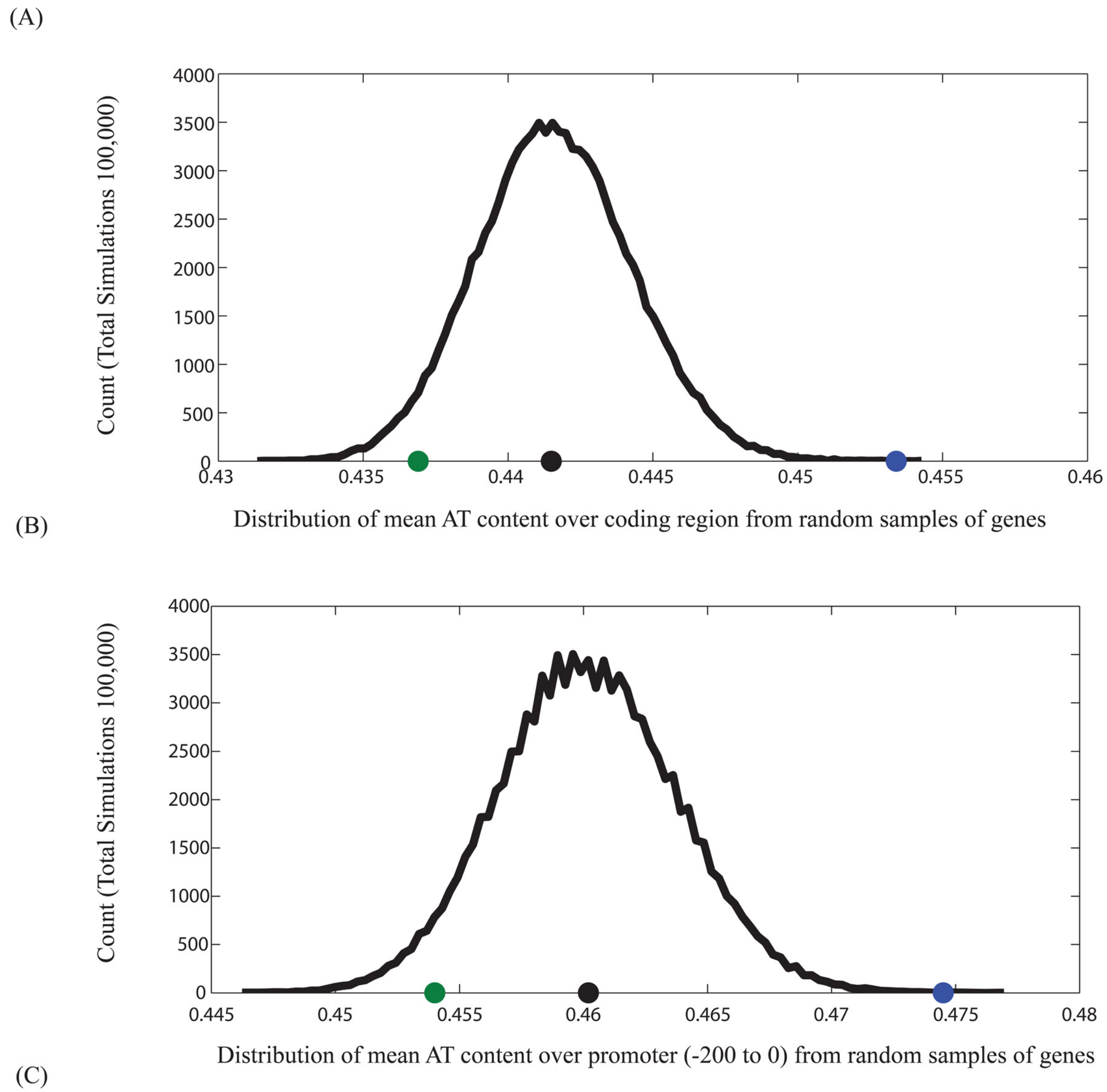


Fig. S6. Chloroquine gel electrophoresis (CAGE) for determination of endogenous plasmid superhelicity during the circadian cycle. (*A*) SybrGold (Invitrogen) stained CAGE gel identifying distribution of topoisomers during the circadian cycle. (*B*) Quantification of supercoiling from (*A*) as described in *Materials and Methods*.



(C)

	Mean AT Coding Region	p-value Coding Region	Mean AT -200 to 0	p-value -200 to 0
Relaxation Repressed	0.4369	3,597 / 100,000	0.4540	4,073 / 100,000
Relaxation Activated	0.4534	4 / 100,000	0.4745	16 / 100,000
All Genes	0.4415		0.4602	

Fig. S7. Statistical significance of increased and decreased AT content in monotonically relaxation activated and monotonically relaxation repressed genes, respectively. (*A*) Two hundred fifty genes were selected at random from the full genome 100,000 times and the mean AT content of their coding regions plotted (black histogram). The average AT content of all of the coding regions, the monotonically relaxation activated set, and the monotonically relaxation repressed set are shown with black, blue, and green dots, respectively. The monotonically relaxation activated and monotonically relaxation repressed sets are defined as the 250 genes with lowest (closest to -1) and highest (closest to $+1$) Pearson correlation to the supercoiling waveform represented in Fig. 1*B*. (*B*) Same as (*A*) with promoter (-200 to 0) AT content plotted instead of coding region. (*C*) The table shows the number of times the random simulation chose a set of genes with a more extreme AT content (*P* value). All *P* values are <0.05.

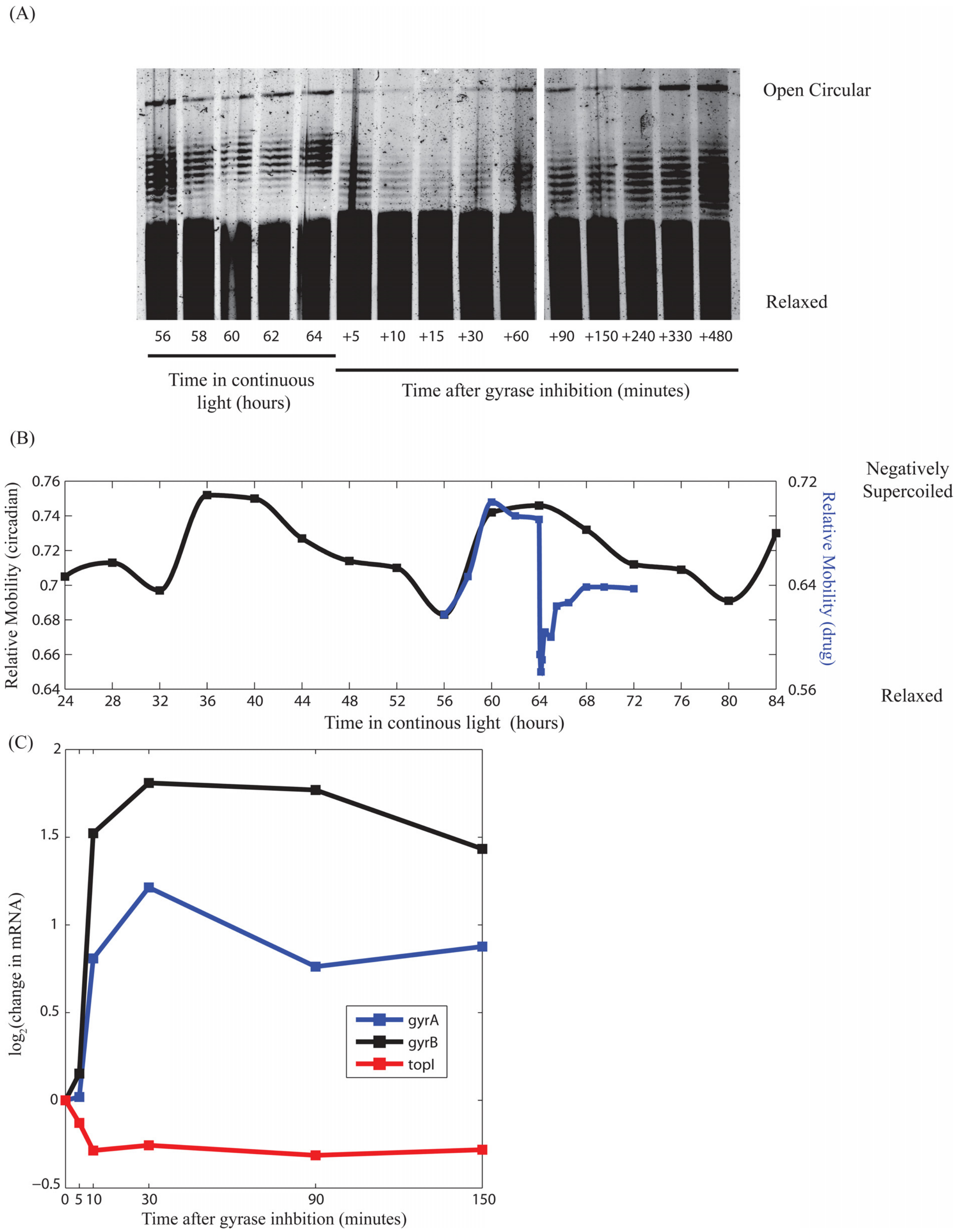


Fig. S8. Chloroquine gel electrophoresis (CAGE) for determination of endogenous plasmid superhelicity after novobiocin induced relaxation. (*A*) SybrGold (Invitrogen) stained CAGE gel identifying distribution of topoisomers both before and after novobiocin (0.1 μg/mL novobiocin sodium salt) addition. (*B*) Quantification of supercoiling from (*A*) (blue) superimposed on supercoiling changes during the circadian cycle. Novobiocin addition immediately relaxes the pANS plasmid to a level similar to the most relaxed state during the circadian cycle. (*C*) Known supercoiling responsive genes in *E. coli* show a similar homeostatic response in *S. elongatus*.

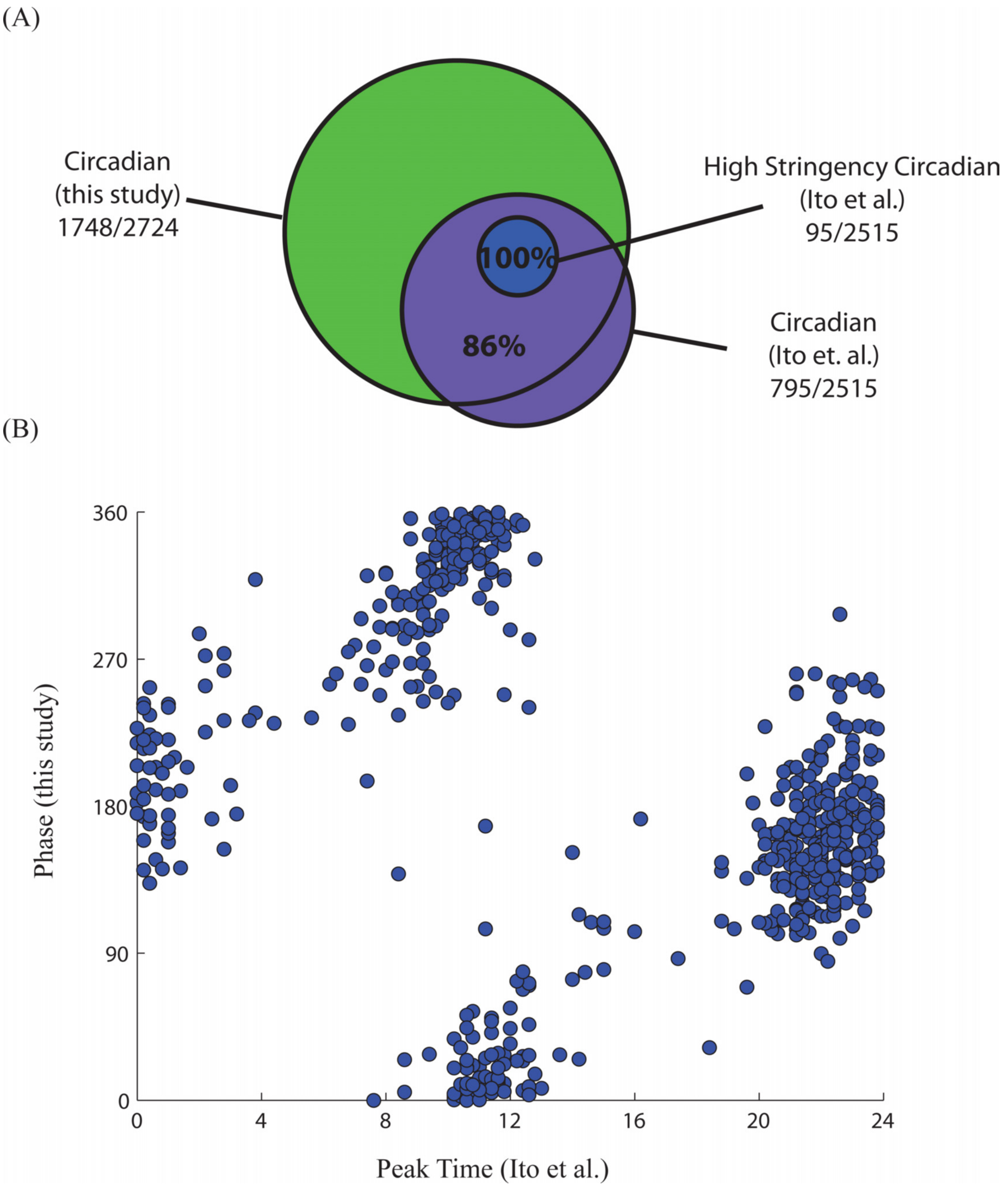


Fig. S9. Comparison of circadian gene expression to ref. 1. (*A*) Comparison of circadian genes found in this study vs. ref. 1. Eighty-six percent and 100% of the genes indentified by ref. 1 as low and high stringency circadian genes, respectively, were classified as circadian in this study. Although most circadian genes identified in ref. 1 are identified in this study, we additionally identify 955 circadian genes. Some of these genes are predicted ORFs not present in the ref. 1 microarray experiments. Part of the remaining circadian genes identified in this study can be attributed to our identification criteria. Here, we do not filter genes with low amplitude and do not require expression profiles to necessarily be cosine-like. (*B*) Comparison of phase (this study) vs. peak time (ref. 1) in all genes identified as circadian in both studies. Genes oscillate with similar phase in both studies.

1. Ito H, et al. (2009) Cyanobacterial daily life with Kai-based circadian and diurnal genome-wide transcriptional control in *Synechococcus elongatus*. *Proc Natl Acad Sci USA* 106:14168–14173.

Other Supporting Information Files

Dataset S1 (XLS)