Supporting Information

Ito et al. 10.1073/pnas.0902587106

SI Methods

Microarray Analysis. Total RNA was purified with the modified acid-hot phenol method (1). From total RNA samples, 1.5 μg of cDNA was prepared with a SuperScript III reverse transcriptase (Invitrogen) with random hexamers as primers (Invitrogen), partially fragmented with DNaseI (Amersham Pharmacia Biotech) and biotinylated by using an ENZO BioArray Terminal Labeling kit (ENZO Life Sciences). The labeled cDNAs were hybridized in duplicate to the GeneChip arrays. The arrays were hybridized, washed, and stained by using standard Affymetrix prokaryotic GeneChip reagents and protocols. Affymetrix GeneChip software was used to determine the average difference between matched and mismatched oligonucleotide probes for each probe set. For better estimation of relative expression levels among different genes, we standardized each cDNA-derived signal with the corresponding genomic DNA-derived signal. We also performed hybridization of the *Synechococcus* oligonucleotide arrays with biotin-labeled DNaseI-fragmented genomic DNA. The genomic DNA-derived signals varied within a range of ≈10-fold, whereas the RNA-derived signals (wild-type RNA samples collected at LL 12 as example) varied within a range of ≈10^3-fold. The signals for genomic DNA were within a linear range. Genomic DNA signals were scaled so that their average level was 1. Each RNA signal was then divided by the corresponding scaled genomic DNA signal. Values of RNA signals given hereafter indicate these genomic-DNA-normalized values. Additionally, global normalization was applied to the RNA signal profiles under LL conditions so that the averages of the expression levels for all ORFs within a microarray were equal across the replicate arrays. The procedure for data processing is shown in Fig. S8.

Identification of Circadian Cycling Genes. We identified statistically significant circadian cycling expression profiles from 2515 *Synechococcus* ORFs under LL conditions using 2 criteria: "*correlation P value*" to extract genes with periodic expression patterns and "*amplitude*" to identify genes where the changes were above background level. The "*correlation P value*" (hereafter called "*P value*") is an index parameter that indicates the degree to which the gene expression profile has circadian periodicity (2). We defined *P value* as the significance of the Pearson's correlation between each expression profile and cosine wave for which the period is 24 h. We prepared 60 cosine waves with equally spaced phases. Pearson's correlation was calculated between each cosine wave and gene expression profile. We searched for the maximum value of the correlations to find the cosine wave with the best fit. Statistical significance of the maximum correlation of the gene profile was assessed by an empirical procedure. We generated 10^5 normally distributed random expression profiles. The maximum Pearson's correlation for each virtual random profile was calculated in the same way. Thus, the *P value* was obtained by calculating the ratio of the random temporal profiles assessed as a higher maximum correlation than that of the gene expression profile. *amplitude* is an index parameter that indicates to what extent the fluctuation in gene expression profile is above the background level. We defined *amplitude* as the coefficient of variation (standard deviation divided by the mean). Because the experiments under LL conditions were performed twice, 2 *amplitudes* and 2 *P values* were acquired for an ORF. The values were combined according to the following equations:

$$p\_value = \max(p\_value_{1st},\ p\_value_{2nd})$$

$$Amplitude = \min(Amplitude_{1st},\ Amplitude_{2nd})$$

The ORFs that had a higher *P value* and *amplitude* than the respective thresholds were identified as circadian cycling genes. In this research, both thresholds for the index parameters were assigned a value of 0.1. To estimate the false positive rate, we generated 10^5 random expression profiles that were normally distributed. Then we filtered these random expression profiles using the 2 thresholds. Random profiles produced 4203 genes classified as circadian cycling genes. We assume that this is the false positive rate (i.e., 4.2% of all identified genes would be false positive).

We also extracted circadian cyclic genes from *Synechocystis* expression profiles previously reported by Kucho et al. (3). They twice collected mRNAs every 4 h under LL conditions, and each sample was applied to DNA microarray and scanned by a higher- or lower-sensitivity detection method. The *P value* and *amplitude* were calculated for each time course and for *Synechococcus* and combined according to the following equations.

$$Pval = \min(\max(Pval_{highSensitivity,1st},\ Pval_{highSensitivity,2nd}),\ \max(Pval_{lowSensitivity,1st},\ Pval_{lowSensitivity,2nd}))$$

$$Amplitude = \max(\min(Amplitude_{highSensitivity,1st}, Amplitude_{highSensitivity,2nd}),\ \min(Amplitude_{lowSensitivity,1st}, Amplitude_{lowSensitivity,2nd})).$$

The same value was assigned to the thresholds for *Synechocystis* index parameters and the cycling genes were identified.

As mentioned above, we used global normalization for the RNA signals under LL conditions so that the averages of the expression levels for all ORFs within a microarray across the replicate arrays were equal. Note that some possible time-dependent bias may be involved in the normalization procedure, considering that a large portion of the genome is under the control of the clock. For example, as shown in Fig. 4*C* (white circle), the sum of raw hybridization signals for 2 independent experiments in LL show some fluctuations. However, it does not show circadian variation, and the *P value* and *amplitude* for the sum of raw signals in LL are 0.336 and 0.058 for the first experiment, and 0.78 and 0.083 for the second experiment, respectively. Note that we adopted more stringent values as the threshold to identify cycling genes (Fig. 1).

Extracting Peak Phase of Circadian Expression Profile. To determine the peak phase of cycling genes under LL conditions, we tested for Pearson's correlation between the temporal expression profiles of each gene and 24 h period cosine waves at 60 equally spaced phases. We estimated the phase of each cycling gene from the phase of the cosine wave with which it was most closely correlated. When the cosine function with the highest correlation coefficient is formulated as $\cos(2\pi\ (time\ -\ \varphi)/24)$, φ represents the estimated phase. Note that φ satisfies $0 \leq \varphi < 24$.

Estimation of Operons. The operons of *Synechococcus* ORFs were inferred by using 5 criteria including position on genome, mean expression and correlation of expression. We extracted any pairs of genes that satisfied all of the criteria for an operon.

Suppose that gene1 and gene2 are inferred to be an operon, these genes would satisfy the successive criteria as follows:

- Direction:
- gene1 and gene2 should be in the same transcriptional direction on the genome.
- Successive position:
- gene1 and gene2 should be next to each other on the genome.
- Positions on genome sequence.
- The intergenic distance between gene1 and gene2 is smaller than the threshold X.
- Expression average.
- The expression levels of gene1 and gene2 should not be significantly different. The time course average of the genes under constant light conditions ($Average_{1st}$, $Average_{2nd}$) should satisfy the following equation:

$$\frac{1}{Y} \leq \frac{Average_{1st}}{Average_{2nd}} \leq Y,$$

where threshold Y is >1.

- Expression correlations.
- The fluctuations of expression of gene1 and gene2 should correlate. The Pearson's correlation coefficient between gene1 and gene2 under constant light conditions should be higher than threshold Z.

The thresholds X, Y, and Z were set to 200 base pairs, 6.86 and 0.34, respectively. These threshold values were chosen to identify the already known operons of *Synechococcus* experimentally validated by Northern blotting, as follows:

- *idiB* operon (*syc1921_c/Synpcc_7942_2174*, *syc1922_c/Synpcc_7942_2173*, *syc1923/Synpcc_7942_2172*) (4)
- *kaiBC* operon (*syc0333_d/Synpcc_7942_1217*, *syc0334_d/Synpcc_7942_1216*) (5)
- *irpAB* operon (*syc0095_d/Synpcc_7942_1462*, *syc0096_d/Synpcc_7942_1461*) (6).
- *cmpABCD* operon (*syc2474_d/Synpcc7942_1488*, *syc2475_d/Synpcc7942_1489*, *syc2476_d/Synpcc7942_1490*, *syc2477_d/Synpcc7942_1491*) (7)
- *nirA-nrtABCD-narB* operon (*syc0310_d/Synpcc7942_1240*, *syc0311_d/Synpcc7942_1239*, *syc0312_d/Synpcc7942_1238*, *syc0313_d/Synpcc7942_1237*, *syc0314_d/Synpcc7942_1236*, *syc0315_d/Synpcc7942_1235*) (8)
- *moaCDEA* operon (*syc0268_d/Synpcc7942_1285*, *syc0269_d/Synpcc7942_1284*, *syc0270_d/Synpcc7942_1283*, *syc0271_d/Synpcc7942_1282*) (9).

If the pair gene1 and gene2 and the pair gene2 and gene3 were both identified as operons, we unified the 3 genes as an operon including gene1, gene2 and gene3. We defined the unified operons and the genes that were not included in any operons as transcriptional units.

1. Iwasaki H, et al. (2000) A KaiC-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. *Cell* 101:223–233.
2. Yamada R, Ueda HR (2007) Microarrays: Statistical methods for circadian rhythms. *Circadian Rhythms*: *Methods and Protocols*, Methods in Molecular Biology, ed. Rosato E (Humana Press, Clifton, NJ), Vol 362, pp 245–264.
3. Kucho K, et al. (2005) Global analysis of circadian expression in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *J Bacteriol* 187:2190–2199.
4. Yousef N, Pistorius EK, Michel KP (2003) Comparative analysis of *idiA* and *isiA* transcription under iron starvation and oxidative stress in *Synechococcus elongatus* PCC 7942 wild-type and selected mutants. *Arch Microbiol* 180:471–483.
5. Ishiura M, et al. (1998) Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. *Science* 281:1519–1523.
6. Nodop A, et al. (2008) Transcript profiling reveals new insights into the acclimation of the mesophilic freshwater cyanobacterium *Synechococcus elongatus* PCC 7942 to iron starvation. *Plant Physiol* 147:747–763.
7. Omata T, et al. (1999) Identification of an ATP-binding cassette transporter involved in bicarbonate uptake in the cyanobacterium *Synechococcus* sp. strain PCC 7942. *Proc Natl Acad Sci USA* 96:13571–13576.
8. Suzuki I, Sugiyama T, Omata T (1993) Primary structure and transcriptional regulation of the gene for nitrite reductase from the cyanobacterium *Synechococcus* PCC 7942. *Plant Cell Physiol* 34:1311–1320.
9. Rubio LM, Flores E, Herrero A (1998) The *narA* locus of *Synechococcus* sp. strain PCC 7942 consists of a cluster of molybdopterin biosynthesis genes. *J Bacteriol* 180:1200–1206.

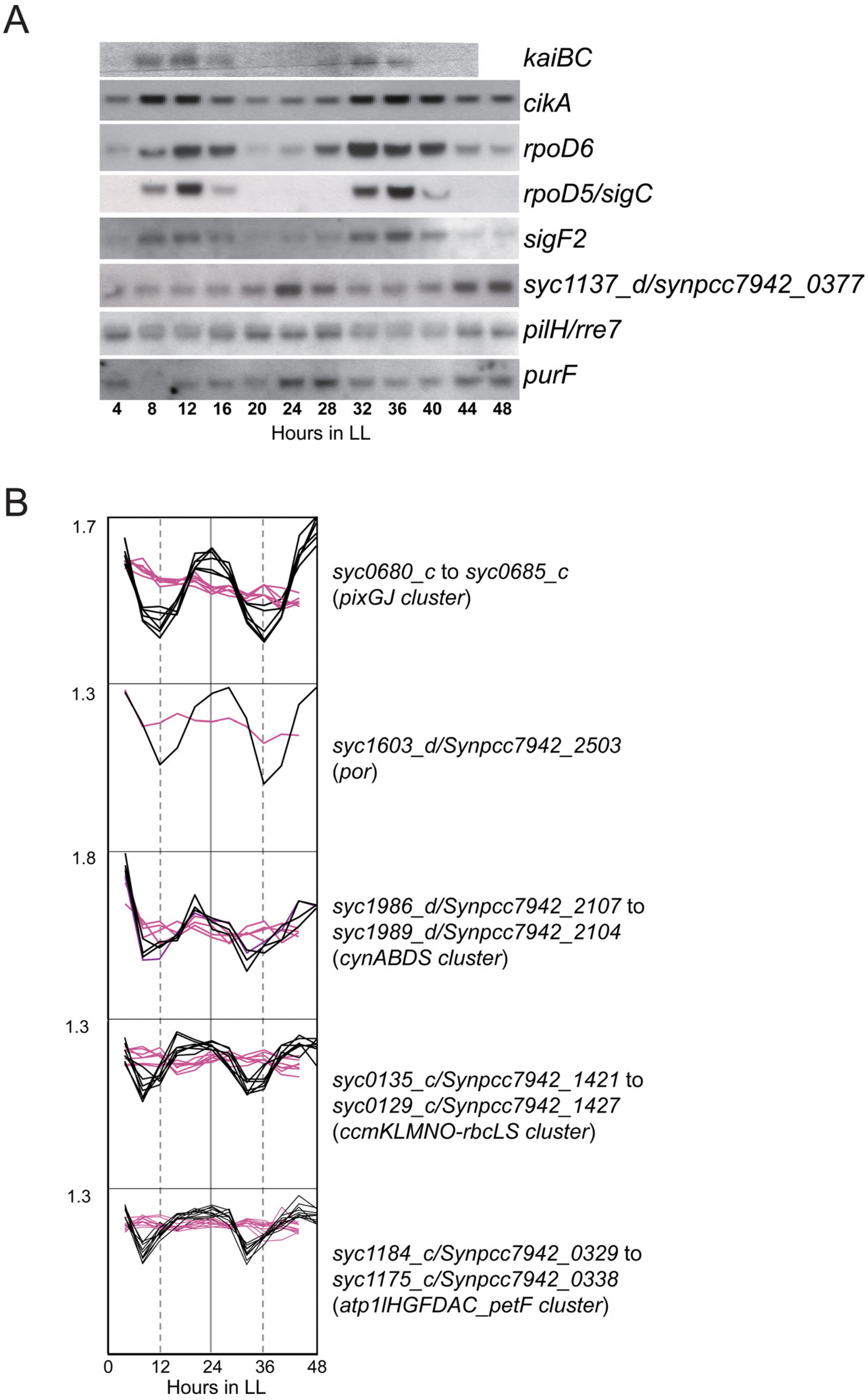


Fig. S1. Clock-controlled gene expression. (*A*) Validation of circadian expression profiles by Northern hybrid analysis. Wild-type *Synechococcus* cells were cultivated under LL after 2 LD cycles and subjected to RNA blot analysis (2 μg of total RNA per lane). Results for 5 subjective dusk genes (*kaiBC*, *cikA*, *rpoD6*, *rpoD5/sigC*, *sigF2*) and 3 subjective dawn genes (*syc1137_d/Synpcc7942_0377*, *syc0684_c/pilH/rre7*, *purF*) are shown. (*B*) DNA microarray data showing identified genes peaking at subjective dawn. Expression of genes in wild-type (black) and *kaiABC*-null (red) strains are shown. Average values from 2 independent experiments are plotted. The number on the ordinate indicates relative expression level (the mean value of expression levels in each strain was normalized to 1). Note that the expression profile on DNA microarray of additional dawn genes, *purF*, *syc1137_d/ Synpcc7942_0377* and *syc0684_c/pilH/rre7*during the 24 h in LL are also shown in Fig. 4.

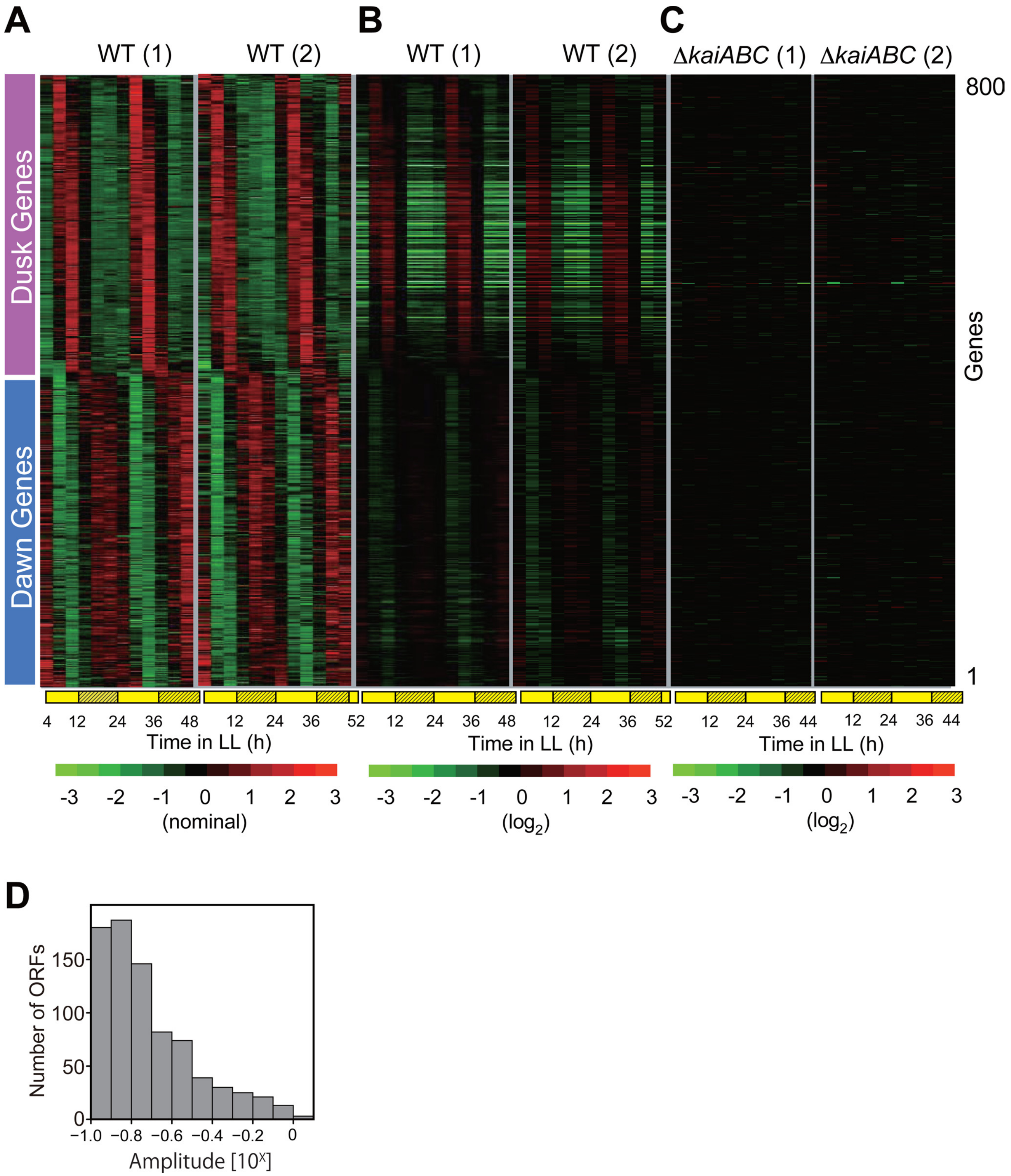


Fig. S2. Temporal expression profiles under continuous LL conditions (duplication). (*A* and *B*) Expression profiles of 800 cycling genes in wild-type strains sorted by peak time. (*C*) Expression profiles of the clock-controlled genes in *kaiABC*-null mutant strains. Results from 2 independent experiments are shown. Representation of data are the same as for Fig. 1, where one of the 2 (Panel 1 here) is shown. (*D*) A histogram showing a continuous distribution of the *amplitude* index for 800 clock-controlled genes.

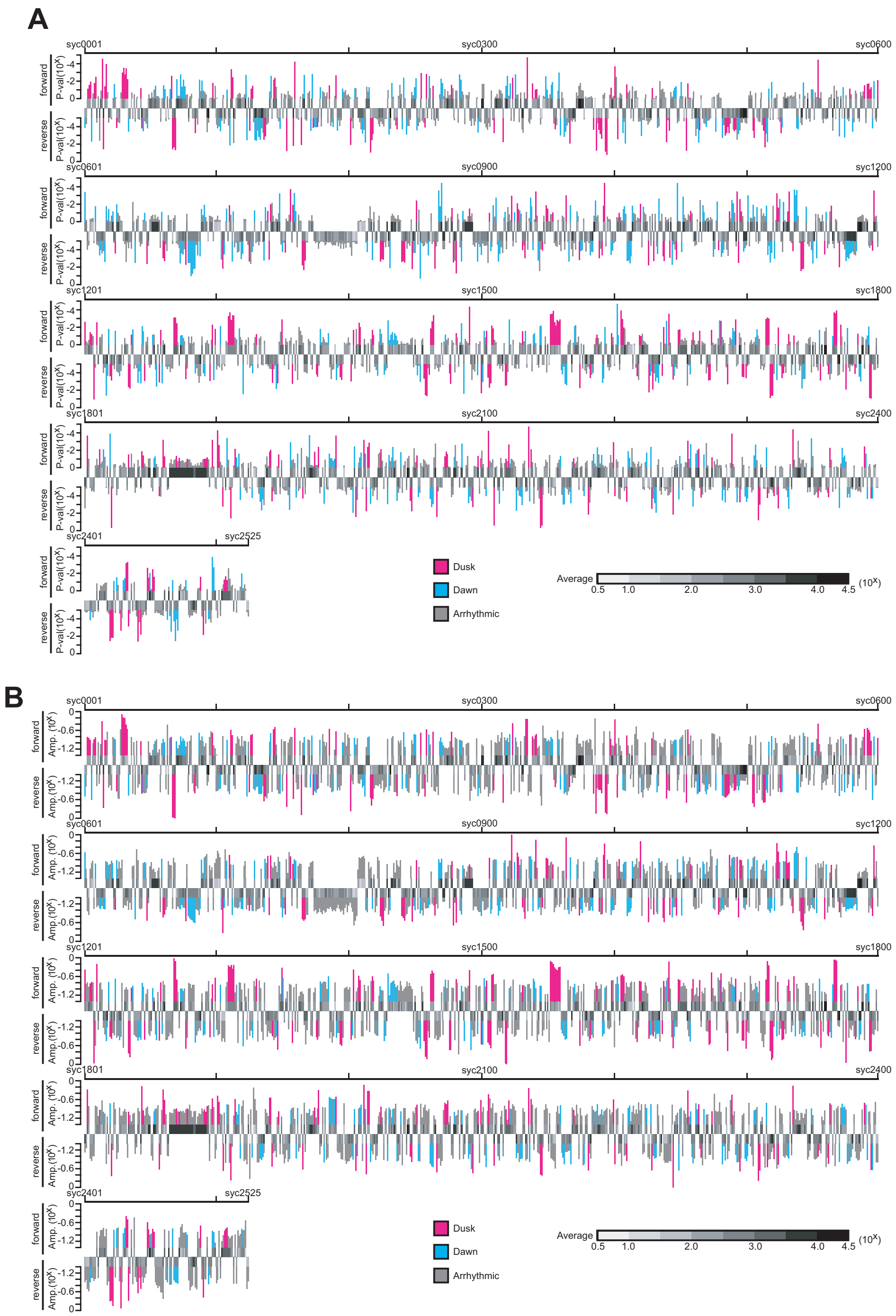


Fig. S3. Distribution of clock-controlled genes and possible operons in the *Synechococcus* genome. (*A* and *B*) Genome-wide profile of oscillatory index parameters. The expression averages and oscillatory index parameters (*P value* in panel A or *amplitude* in panel B) are arranged in ascending order of ORF number. The average signal during LL conditions is represented by the gray scale image. The oscillatory indexes are shown as a bar plot on the upper side (forward direction) or lower side (reverse direction). The color of the bar plot represents the peak phase of expression in wild-type strains under LL; red for subjective dusk genes, blue for subjective dawn genes and gray for arrhythmic genes. Derivation of *P value*, peak phase and *amplitude* is described in the *SI Methods*. (*C* and *D*) Profiles of oscillatory index parameters of transcriptional units. We identified as an operon ORF pairs that satisfied 5 criteria: expression levels were almost identical; the time course correlation was sufficiently high, and they are close together on the genome sequence (details are described in *SI Methods*). The oscillatory parameters of single ORFs and predicted operons are displayed in the same way as in panels A and B. (*E*) A correlation map of all ORFs. Pearson's correlation coefficients in all combinations of 2 ORFs (for the total 2515 ORFs) were calculated from expression profiles in the wild-type strain under LL conditions. Two giant gene clusters (presumed operons) are indicated by *a* and *b* and include genes for 29 ribosomal proteins (*syc1865_d* to *syc1893_d*) and 33 hypothetical proteins (*syc0775_c to syc0807_c*), respectively. Some small clusters with high correlation values derived from the operonic structure are located diagonally. (*F*) The correlation map of the 1681 transcriptional units (Table S1) eliminating possible operonic redundancy. Correlations between transcriptional units were recalculated and displayed in the same way as in *E*.

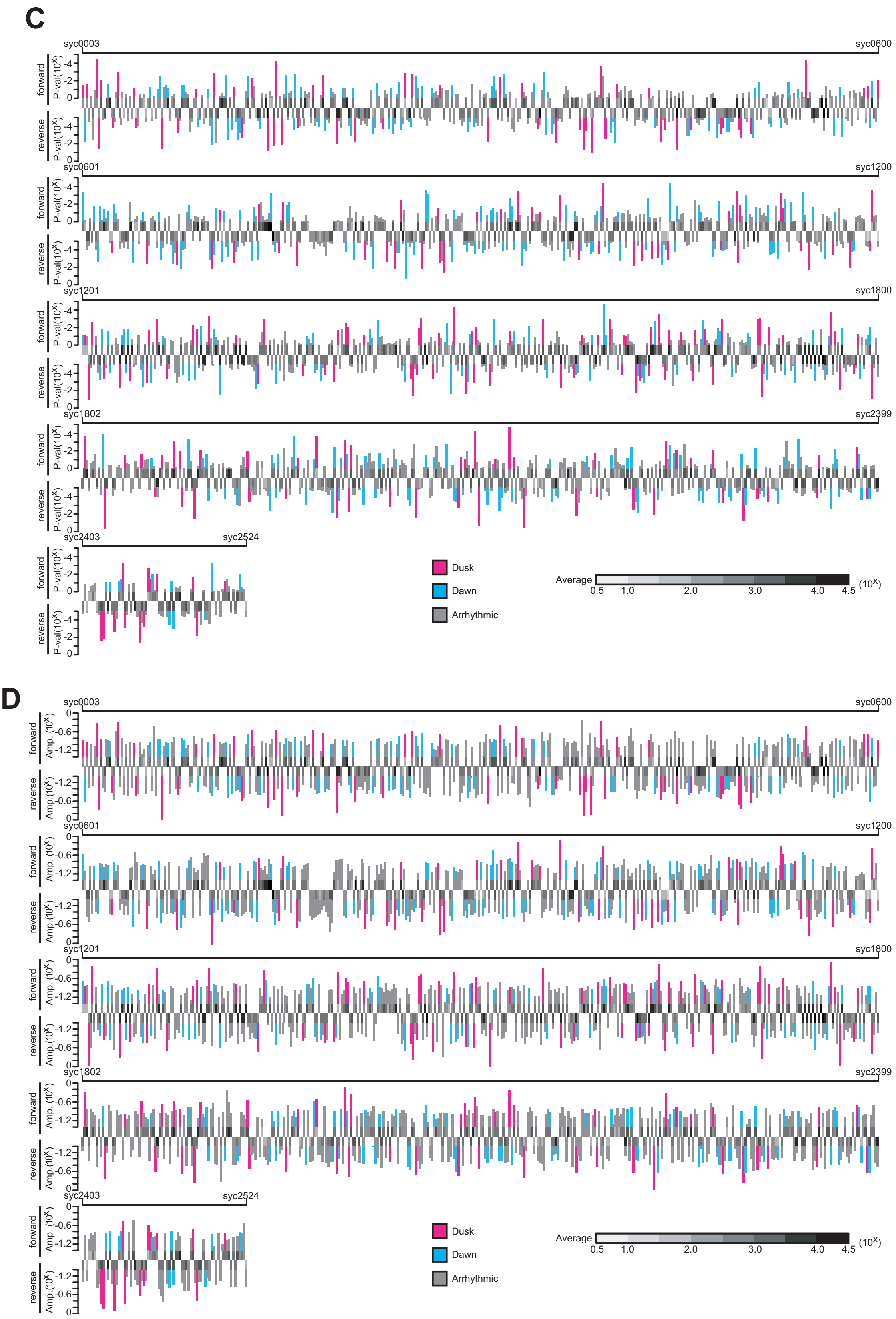


Fig. S3. Continued.

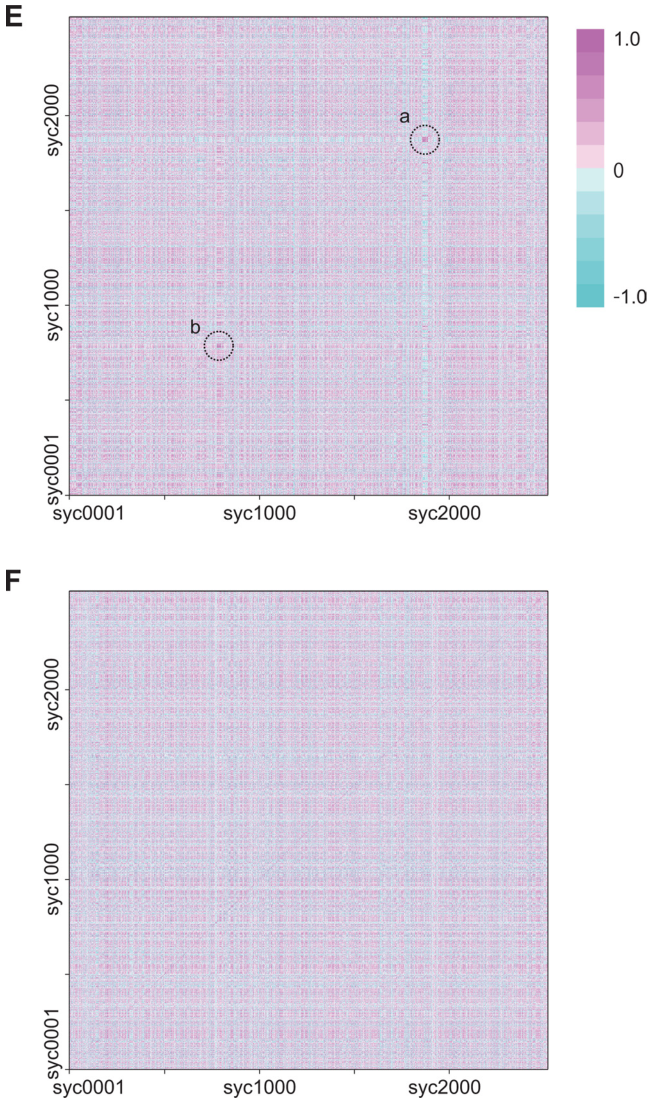


Fig. S3. Continued.

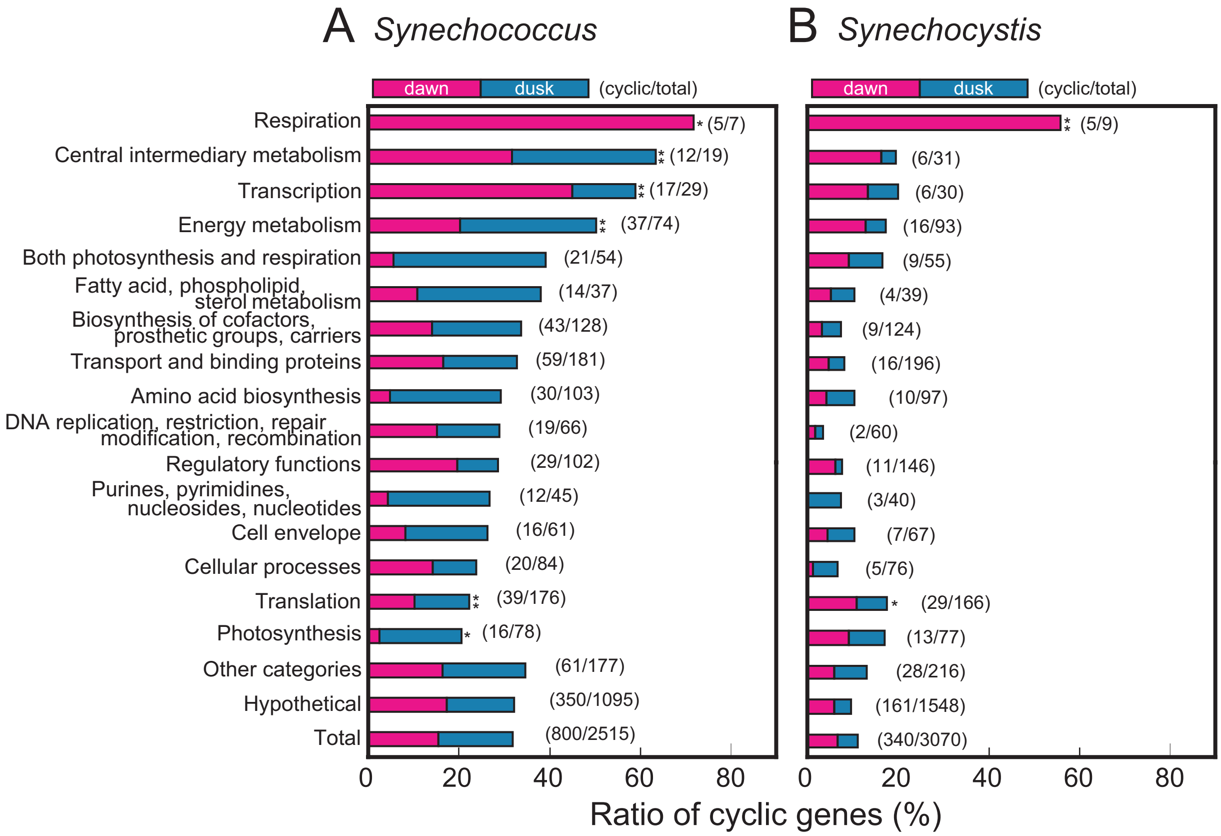
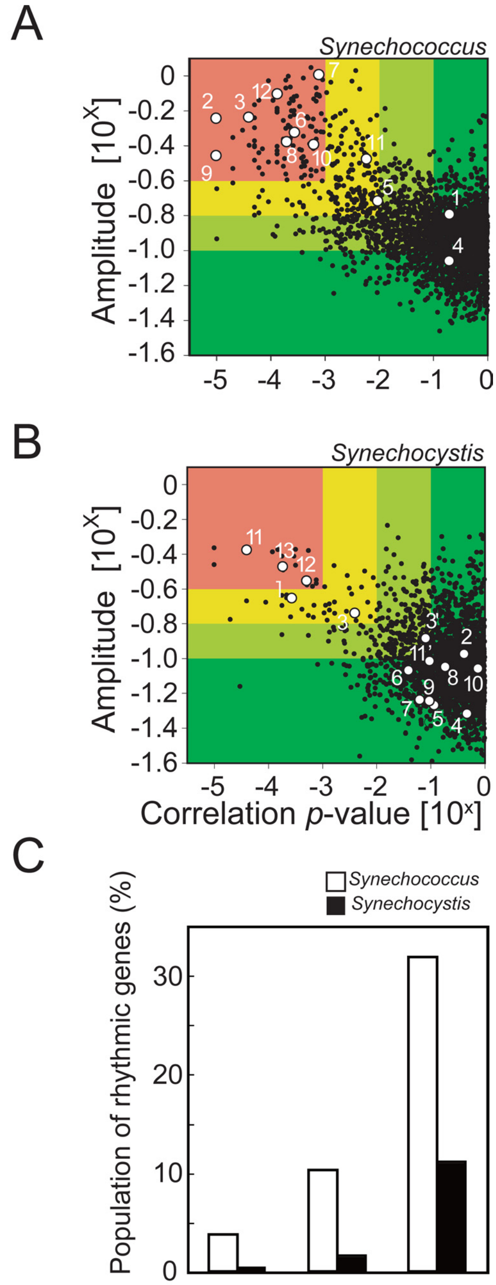


Fig. S4. Functional categories of clock-controlled genes. Clock-controlled genes in *Synechococcus* (*A*) and *Synechocystis* (*B*) were classified by their functional category. The proportion of clock-controlled genes to total genes in each category is represented as stacked bar graphs. Red and blue bars indicate the ratio of dawn and dusk genes, respectively, to total cycling genes in each category. The numbers of cyclic genes and total genes in each category are shown in parentheses. Asterisks indicate the categories containing significantly higher or lower ratios of cyclic genes compared with the ratio of identified cyclic genes to total genes (**, $P < 0.01$; *, $P < 0.05$; Fisher's exact test).



Correlation *p*-value	< 0.001	< 0.01	< 0.1
Amplitude	> $10^{-0.6}$	> $10^{-0.8}$	> $10^{-1.0}$
Area	🟥	🟨	🟩

Fig. S5. Cycling genes in the *Synechococcus* and *Synechocystis* genomes. (*A* and *B*) Variations in *amplitude* (abscissa) and a cosine-fitting correlation score (*correlation P value*; ordinate) from each transcript under LL conditions in wild-type *Synechococcus* (*A*) and *Synechocystis* strains. Data processing and representation are the same as in Fig. 1. 1. *kaiA*, 2. *kaiB* (*A*) or *kaiB1* (*B*), 3. *kaiC* (*A*) or *kaiC1*, 3′. *Synechocystis kaiC3*, 4. *sasA*/*hik8*, 5. *rpaA*/*rre31*, 6. *cikA*, 7. *rpoD5*/*sigC*, 8. *rpoD6*/*sigB*, 9. *purF*, 10. *pilH*/*rre7*, 11. *ctaC*/*coxB*, 12. *opcA*, and 13. *Synechocystis sigE* (B only). (*C*) Population of rhythmic genes in total ORFs dependent on different filtering parameters. Note that when a more stringent filtering set is used, a greater difference in the robustness of circadian control appears between the 2 species.

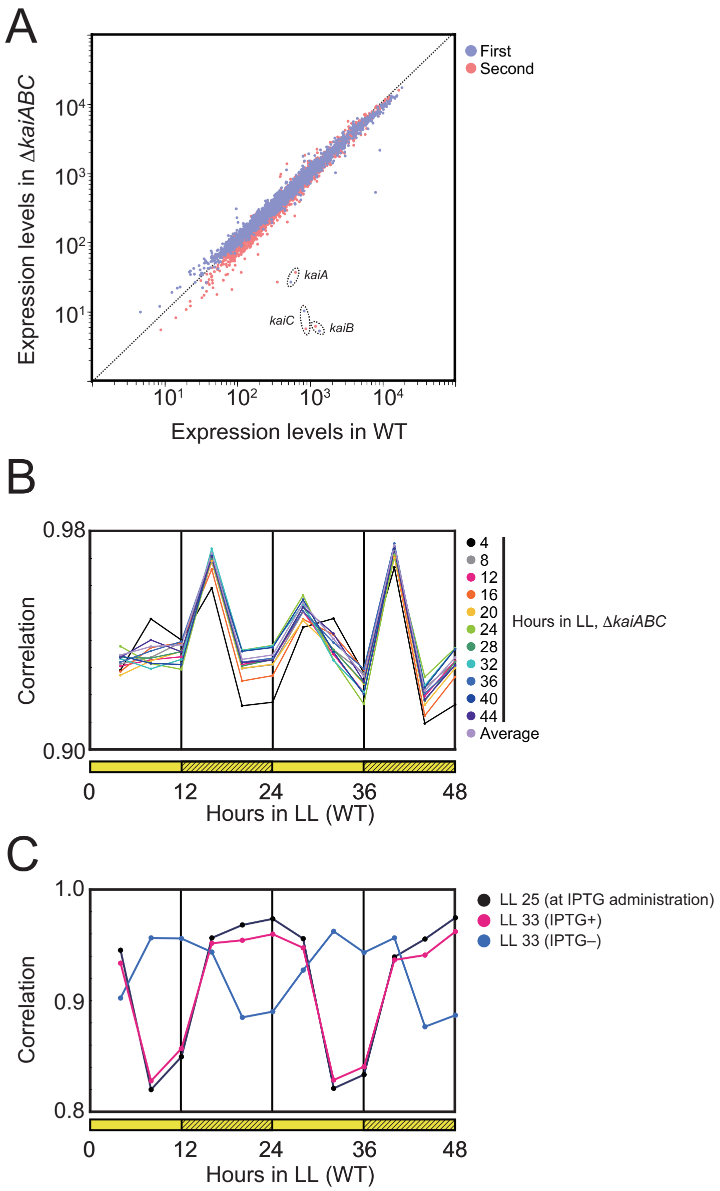


Fig. S6. Correlation analysis. (*A*) Average expression level is not much affected by *kaiABC*-nullification. The average expression level of each gene over 48 h in LL was compared in wild-type and *kaiABC*-deficient strains. Two independent experiments for both strains were compared. (*B*) Correlation coefficients were calculated on a dataset covering 2515 ORFs to compare transcription profiles in *kaiABC*-null and wild-type strains grown under LL with a combination of the indicated sampling times. Note that the correlation coefficient is lower at subjective dawn and dusk in the wild-type strains compared with each of the samples collected from *kaiABC*-null mutant strains. (*C*) Correlation coefficients comparing *kaiC*-overexpressing and wild-type strains under LL with a combination of indicated sampling times.

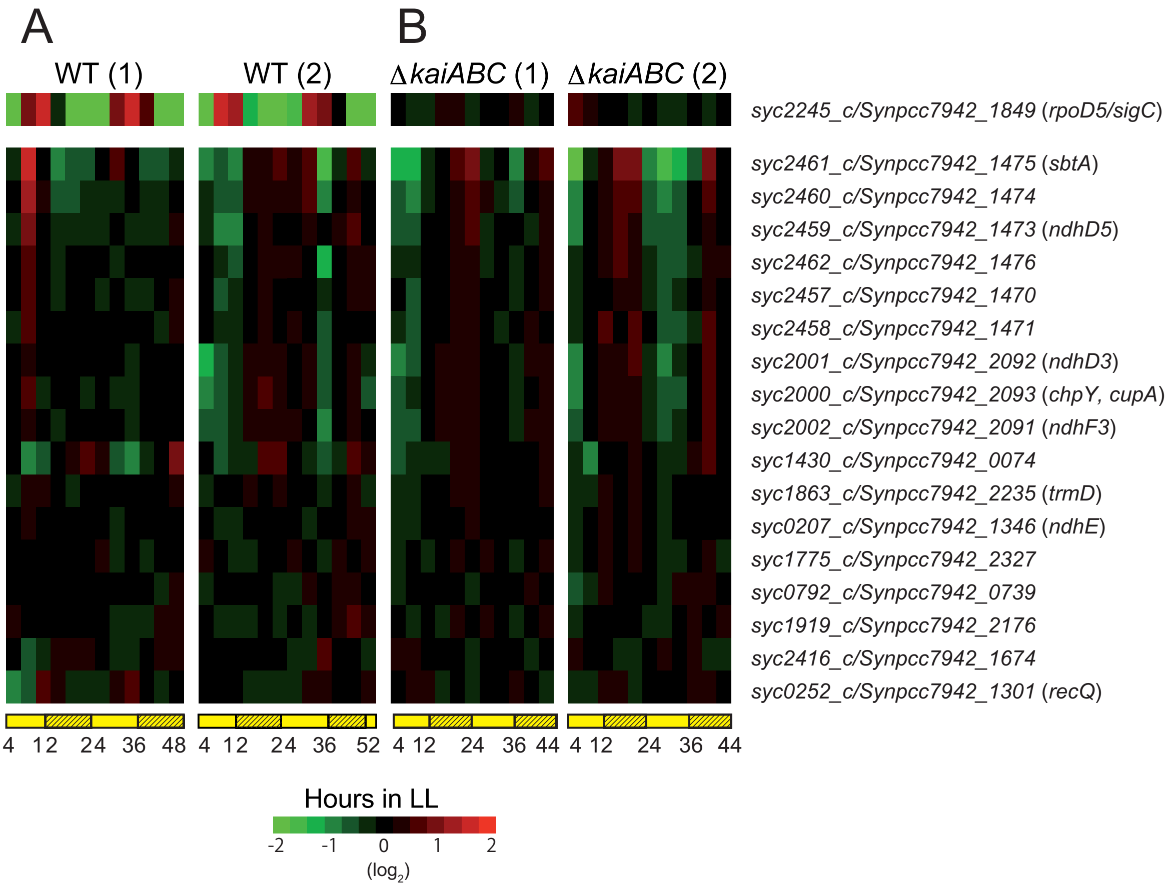


Fig. S7. Expression of *kai*-independent cycling transcripts. Expression profiles of 17 *kai*-independent cycling genes in the wild-type and *kaiABC*-null mutant strains. Results from 2 independent experiments are shown. Representation of data are the same as for Fig. S2. As a reference, expression profiles in both strains of a representative high-amplitude *kai*-controlled gene, *rpoD5/sigC*, are also shown at the top.

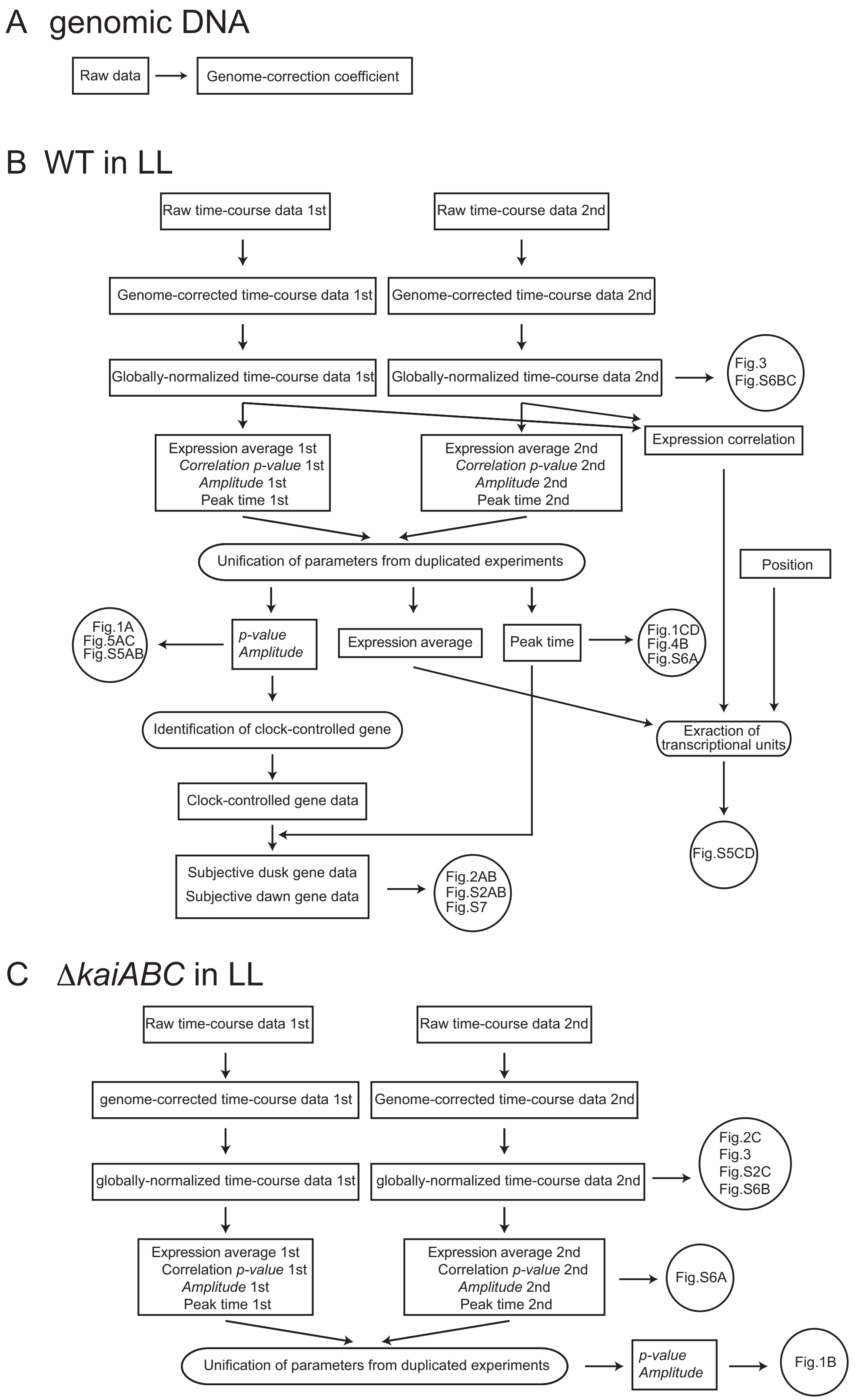


Fig. S8. Schematic diagram for data processing. To extract oscillatory index parameters or normalized expression profiles, we processed the raw expression data from genomic DNA signals (*A*), mRNA signals from *Synechococcus* wild-type strains under LL (*B*), *Synechococcus* Δ*kaiABC* strains under LL (*C*), *Synechococcus* wild-type strains under DD (*D*), *Synechococcus kaiC*-overexpressor strains under LL (*E*) and *Synechocystis* strains under LL (*F*). Details are described in *SI Methods*.

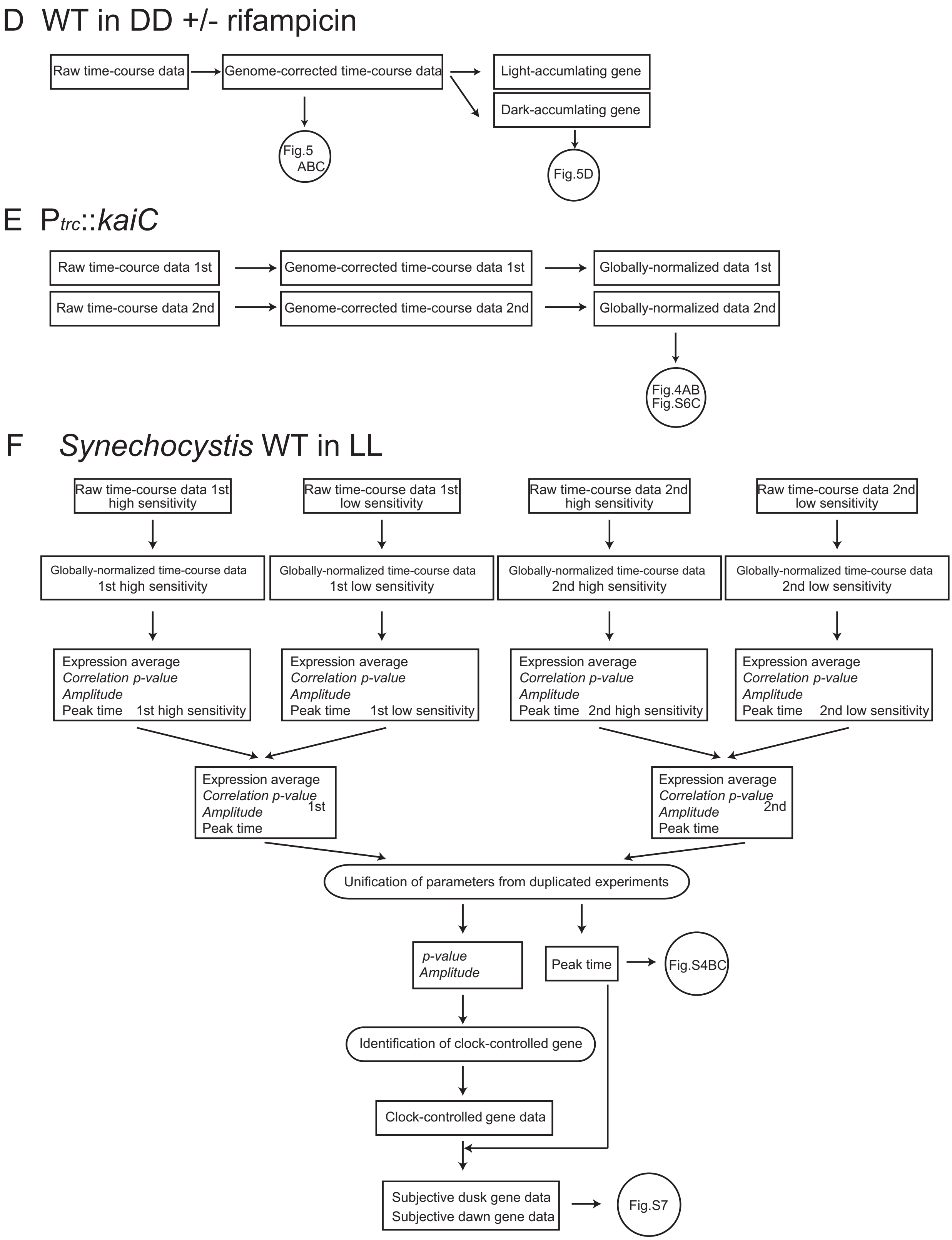

D WT in DD +/- rifampicin
Raw time-course data
Genome-corrected time-course data
Light-accumlating gene
Dark-accumlating gene
Fig.5 ABC
Fig.5D
E P trc::kaiC
Raw time-cource data 1st
Raw time-course data 2nd
Genome-corrected time-course data 1st
Genome-corrected time-course data 2nd
Globally-normalized data 1st
Globally-normalized data 2nd
Fig.4AB Fig.S6C
F Synechocystis WT in LL
Raw time-course data 1st high sensitivity
Raw time-course data 1st low sensitivity
Raw time-course data 2nd high sensitivity
Raw time-course data 2nd low sensitivity
Globally-normalized time-course data 1st high sensitivity
Globally-normalized time-course data 1st low sensitivity
Globally-normalized time-course data 2nd high sensitivity
Globally-normalized time-course data 2nd low sensitivity
Expression average Correlation p-value Amplitude Peak time 1st high sensitivity
Expression average Correlation p-value Amplitude Peak time 1st low sensitivity
Expression average Correlation p-value Amplitude Peak time 2nd high sensitivity
Expression average Correlation p-value Amplitude Peak time 2nd low sensitivity
Expression average Correlation p-value Amplitude Peak time 1st
Expression average Correlation p-value Amplitude Peak time 2nd
Unification of parameters from duplicated experiments
p-value Amplitude
Peak time
Fig.S4BC
Identification of clock-controlled gene
Clock-controlled gene data
Subjective dusk gene data
Subjective dawn gene data
Fig.S7

Fig. S8. Continued.

Table S1. Populations of rhythmic genes based on operonic organization in the genome

	Subjective dawn	Subjective dusk	Arrhythmic	Total
Possible operons	98 (20.1%)	78 (16.0%)	312 (63.9%)	488 (100%)
Singly regulated genes	208 (17.4%)	155 (13.0%)	830 (69.6%)	1193 (100%)
Total transcription units	306 (18.2%)	233 (13.9%)	1143 (68.0%)	1681 (100%)
Total genes	413 (16.4%)	387 (15.4%)	1715 (68.2%)	2515 (100%)

Table S2. *kai*-independent rhythmic genes

6301 ID	7942 ID	Gene	Annotation	WT*	Δ*kaiABC**
syc0207_c	*Synpcc7942_1346*	*ndhE*	NADH dehydrogenase subunit 4L	AR	*a*
syc0252_c	*Synpcc7942_1301*	*recQ*	ATP-dependent DNA helicase RecQ	AR	*a*
syc0792_c	*Synpcc7942_0739*		Hypothetical protein	AR	*a*
syc1430_d	*Synpcc7942_0074*		Hypothetical protein	*b*	*a*
syc1775_d	*Synpcc7942_2327*		Hypothetical protein	AR	*a*
syc1863_c	*Synpcc7942_2235*	*trmD*	tRNA (guanine-N1)-methyltransferase	AR	*a*
syc1919_d	*Synpcc7942_2176*		Hypothetical protein	AR	*a*
syc2000_c	*Synpcc7942_2093*	*chpY, cupA*	Protein involved in low CO_2-inducible high affinity CO_2 uptake	AR	*a*
syc2001_c	*Synpcc7942_2092*	*ndhD3*	NADH dehydrogenase subunit 4	AR	*a*
syc2002_c	*Synpcc7942_2091*	*ndhF3*	NADH dehydrogenase subunit 5	AR	*a*
syc2416_d	*Synpcc7942_1674*		Hypothetical protein	AR	*a*
syc2457_c	*Synpcc7942_1470*		Hypothetical protein	AR	*a*
syc2458_c	*Synpcc7942_1471*		Hypothetical protein	AR	*a*
syc2459_c	*Synpcc7942_1473*	*ndhD5*	NADH dehydrogenase subunit 4	AR	*a*
syc2460_c	*Synpcc7942_1474*		Hypothetical protein	AR	*b*
syc2461_c	*Synpcc7942_1475*	*sbtA*	Sodium-dependent bicarbonate transporter	AR	*b*
syc2462_d	*Synpcc7942_1476*		Hypothetical protein	AR	*a*

*AR indicates arrhythmic using the standard filtering condition (*P value* of >0.1, and/or *amplitude* of $<10^{-1}$). In this table, *a* and *b* indicate low-amplitude and medium-amplitude cycling genes: *a* (*P value* of < 0.1, and/or *amplitude* of $> 10^{-1}$), *b* (*P value* of < 0.01, and/or *amplitude* of $< 10^{-0.8}$). Note that there were no high-amplitude genes (*P value* of < 0.001, and/or *amplitude* of $> 10^{-0.6}$) in the *kaiABC*-null mutant (see also Fig. 1*B*), whereas 97 genes were found in the wild-type strain.

Other Supporting Information Files

Dataset S1