Circadian Gene Expression in Cyanobacteria

Advisor: Professor Erin K. O'Shea **Author: Vikram Vijayan**

Circadian Gene Expression in Cyanobacteria

Abstract

Cyanobacteria are photosynthetic prokaryotes that live in aquatic environments. The cyanobacterium *Synechococcus elongatus* PCC 7942, (hereafter *S. elongatus*) coordinates its day and night behaviors via a circadian clock. The clock is entrained by light/dark cycles but continues to run in constant light conditions. The core circadian clock in *S. elongatus* is encoded by post-translational modifications of three Kai proteins, but the extent and mechanism of circadian gene expression are unknown.

We provide the first unbiased characterization of circadian gene expression in *S. elongatus*, demonstrating that ~65% of genes display oscillation in continuous light conditions, with some genes peaking in expression at subjective dawn and others at subjective dusk. We next sought to identify the mechanism by which such a large fraction of the genome could be rhythmically controlled. Through bioinformatic, correlative, and perturbation experiments, we find that circadian changes in chromosome topology/supercoiling are sufficient to drive rhythmic expression (Chapter 2).

To further investigate how chromosome topology can control gene expression we performed a high resolution characterization of transcripts and RNA polymerase across the *S. elongatus* genome (Chapter 3). Bioinformatic analysis of transcription start sites suggests that the AT/GC content a particular region of the promoter is informative in

defining the phase at which a transcript is maximally expressed. We find that these sequences are sufficient to drive circadian gene expression at a particular phase and that mutation of single nucleotides in this region can reverse the expression phase of a transcript (Chapter 4).

To understand the role of chromosome dynamics in circadian gene expression and cyanobacterial physiology, we tagged and followed chromosomes over multiple cell divisions. We find that *S. elongatus* cells harbor multiple ordered copies of a single chromosome, and the organization of chromosomes in the cytoplasm facilitates equal segregation of chromosomes to daughter cells (Chapter 5).

Table of contents

List of tables and figures

Statement of contributions

Chapter 1

Vikram Vijayan (VV) wrote text.

Chapter 2

VV and Erin K. O'Shea (EKO) designed research; VV and Rick Zuzow (RZ) performed research; VV analyzed data; and VV and EKO wrote the text.

Chapter 3

VV and EKO designed experiments; VV and Isha H. Jain (IHJ) performed experiments; VV and IHJ analyzed data; VV and EKO wrote text.

Chapter 4

VV and EKO designed experiments; VV performed experiments; VV analyzed data; VV wrote the text.

Chapter 5

VV, IHJ and EKO designed research; VV and IHJ performed research; VV and IHJ analyzed data; and VV, IHJ and EKO wrote the text.

Chapter 6

VV wrote text.

Funding

This work was funded by the Howard Hughes Medical Institute, National Defense Science and Engineering (VV) and National Science Foundation Graduate Research Fellowships (VV).

Acknowledgements

First and foremost, I would like to thank my advisor, Erin K. O'Shea, for providing me with support and independence throughout my Ph. D. Without Erin's guidance it would not have been possible for me to successfully try such a diverse array of experiments.

I would like to thank the members of my dissertation advisory committee for being helpful throughout my Ph. D.: Richard Losick, Vlad Denic, Aviv Regev, and Ethan Garner.

I would like to thank Rick Zuzow, a former technician in the lab, who established the lab's first circadian cyanobacteria cultures. Without Rick's initial experiments and help with establishing this model system in the lab, my Ph. D. may not have involved cyanobacteria. I would also like to thank Isha H. Jain, an undergraduate that I mentored for nearly four years. Without her help and encouragement I would not have been nearly as productive. I would like to thank all past and present lab members for making the lab a great intellectual and social environment.

Finally I would like to thank my parents and my sister for their support and for making my life easy for the past 27 (and 23) years.

Vikram Vijayan

September 2012

CHAPTER 1

Introduction to circadian gene expression rhythms in cyanobacteria

Circadian rhythms are endogenous oscillations with approximately 24 hour period. To be deemed circadian, a biological rhythm must satisfy four criteria: (1) period of approximately 24 hours; (2) rhythms must persist without external cues (ex. without light/dark oscillations); (3) external cues must be able to adjust the phase of the rhythm; and (4) the period of the rhythm must be temperature compensated.

The unicellular cyanobacterium, *Synechococcus elongatus* PCC 7942 (hereafter, *S. elongatus*) is the simplest model organism with a circadian clock. Cyanobacteria are responsible for a significant fraction of the earth's photosynthesis, and many species (albeit, not *S. elongatus)* can fix nitrogen. This raises the question of how an organism can fix nitrogen while generating oxygen from photosynthesis given that the enzyme nitrogenase is irreversibly inhibited by oxygen. One solution, discovered in 1968, and employed by filamentous cyanobacteria, is to spatially separate photosynthesis and nitrogen fixation (Fay et al., 1968). Dedicated cells termed heterocysts provide a microanaerobic environment for nitrogen fixation. However, unicellular cyanobacteria, like *S. elongatus,* cannot employ this strategy of spatial separation. In an effort to solve this paradox, Gallon and colleagues observed the temporal relationship between photosynthetic-oxygen evolution and nitrogenase activity in the cyanobacterium, *Gloeocapsa* and found that unicellular cyanobacteria may temporally separate photosynthesis and nitrogen fixation (Gallon et al., 1974). Eight years later, in 1981, Millineaux et al showed that when grown in 12 hour light/dark cycles *Gloeocapsa* fixed nitrogen only during the dark periods (Millineaux et al., 1981), and in 1985 Stal and Krumbein showed that the non-heterocyst forming (yet still filamentous) *Oscillatoria* species had oscillations in nitrogenase activity that persisted when cells were

transferred from light/dark to continuous light conditions (Stal et al., 1985) – providing the first strong evidence for an endogenous circadian clock in cyanobacteria. In 1987, Stal and Krumbein showed that photosynthesis activity oscillated out of phase with nitrogenase activity in continuous light conditions (Stal et al., 1987).

The circadian clock in cyanobacteria is thought to aide in coordinating behaviors – like photosynthesis and nitrogen fixation – with the diurnal changes in light and other parameters (ex. temperature) caused by the rotation of the earth. But what is the purpose of an endogenous clock? Could cyanobacteria simply respond to changes in light and dark instead of investing in an endogenous clock? An elegant experiment by Woelfle and colleagues provided evidence for a fitness advantage in *S. elongatus* with functioning clocks (Woelfle et al., 2004). Woelfle and colleagues found that when mutant strains with altered endogenous period are mixed together, the strain whose period resonated with the rhythm of the external environment outcompeted the others. That is a 22 hour period mutant strain would outperform at 30 hour period strain when light and dark are switched every 11 hours; while the 30 hour period strain would win if light and dark are switched every 15 hours.

What could be the source of this competitive advantage? The most probable explanation is that the circadian clock provides cyanobacteria with predictive information. That is, cyanobacteria may be able to “prepare” for a light to dark or dark to light transition; and mistimed or non-existent “preparation” may lead to a competitive disadvantage. Another intriguing, but unsubstantiated, advantage of a circadian clock is its ability to keep time in the presence of “noise”. In non-laboratory environments,

cyanobacteria may experience brief alterations in light levels. For example, a storm may provide a brief period of darkness during the day, and without a circadian clock, the cyanobacteria may mistake this storm for a transition into night. A transition from day to night physiology during a brief midday storm may result in a fitness disadvantage.

This brings us to the question, what aspects of physiology are controlled by the clock? The cyanobacterium *Oscillatoria* shows oscillations in nitrogenase activity and photosynthesis, but what about our non-diazotrophic model organism for circadian studies – *S. elongatus*? In the early 1990's Kondo and colleagues fused a bacterial luciferase reporter to the *psbAI* promoter in *S. elongatus* and observed robust 24 hour period oscillations in gene expression in continuous light conditions (Kondo et al., 1993). In 2005, a "promoter trap" analysis using bacterial luciferase integrated at approximately 30,000 random loci showed circadian oscillations in bioluminescence at all 800 locations where bioluminescence signal was detected, suggesting genome wide circadian control of gene expression (Liu et al., 2005). Liu et al found that genes typically oscillated with one of two physiologically relevant phases – either peaking at subjective dawn or subjective dusk.

Genetic studies in *S. elongatus* identified a three gene cluster – *kaiA, kaiB,* and *kaiC* – which when deleted abolished circadian rhythmicity of gene expression (Ishiura et al., 1998). It was later identified that oscillations in KaiC phosphorylation occurred with 24 hour period (Iwasaki et al., 2002) and these oscillations can be reproduced in-vitro in the presence of KaiA and KaiB (Nakajima et al., 2005). The oscillations in KaiC phosphorylation satisfy all four criteria for a circadian rhythm and are thought to

constitute the central pacemaker of the circadian clock in *S. elongatus.* Although much is known about the core post-translational oscillations in KaiC phosphorylation, very little is known concerning the extent and mechanism of gene expression control. A two-component regulatory system, SasA-RpaA, has been shown to be modulated by KaiC and deletion of the transcription factor RpaA abolishes circadian gene expression (Iwasaki et al., 2000; Takai et al., 2006). In addition, clock dependent circadian changes in supercoiling and chromosome topology have been implicated in circadian gene expression (Woelfle et al., 2007, Smith et al., 2006), but no causal evidence has been provided.

Here we provide the first unbiased characterization of circadian gene expression in *S. elongatus*, demonstrating that ~65% of genes display oscillation in continuous light conditions, with some genes peaking in expression at subjective dawn and others at subjective dusk. We next sought to identify the mechanism by which such a large fraction of the genome could be rhythmically controlled. Through bioinformatic, correlative, and perturbation experiments, we find that circadian changes in chromosome topology/supercoiling are sufficient to drive rhythmic expression (Chapter 2).

To further investigate how chromosome topology can control gene expression we performed a high resolution characterization of transcripts and RNA polymerase across the *S. elongatus* genome (Chapter 3). Bioinformatic analysis of transcription start sites suggests that the AT/GC content of the +1 to -35 region of the promoter is informative in defining the phase at which a transcript is maximally expressed. We find that these

sequences are sufficient to drive circadian gene expression at a particular phase and that mutation of just a single nucleotide in this region can reverse the expression phase of a transcript (Chapter 4).

To understand the role of chromosome dynamics in circadian gene expression and cyanobacterial physiology, we tagged and followed chromosomes over multiple cell divisions. We find that *S. elongatus* cells harbor multiple ordered copies of a single chromosome, and the organization of chromosomes in the cytoplasm facilitates equal segregation of chromosomes to daughter cells (Chapter 5).

References

Fay P, Stewart WDP, Walsby AE, Fogg GE (1968) Is the heterocyst the site of nitrogen fixation in blue-green algae? *Nature* 220:810-812.

Gallon JR, LaRue TA, Kurz WGW (1974) Photosynthesis and nitrogenase activity in the blue-green alga *Gloeocapsa. Can J Microbiol.* 20:1633–37.

Ishiura M, Kutsuna S, Aoki S, Iwasaki H, Andersson CR, Tanabe A, Golden SS, Johnson CH, Kondo T (1998) Expression of a gene cluster kaiABC as a circadian feedback process in cyabobactera. *Science* 281(1382):1519-1523.

Iwasaki H, Williams SB, Kitayama Y, Ishiura M, Golden SS, Kondo T (2000). A KaiC-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. *Cell* 101:223-233.

Iwasaki H, Nishiwaki T, Kitayama Y, Nakajima M, Kondo T (2002) KaiA-stimulated KaiC phosphorylation in circadian timing loops in cyanobacteria. *Proc Natl Acad Sci USA* 99*:*15788-15793.

Kondo T et al. (1993) Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria. *Proc Natl Acad Sci USA* 90:5672–76.

Liu Y, et al. (1995) Circadian orchestration of gene expression in cyanobacteria. *Genes Dev* 9:1469-1478.

Millineaux PM, Gallon JR, Chaplin AE (1981) Acetylene reduction (nitrogen fixation) by

cyanobacteria grown under alternating light-dark cycles. *FEMS Microbiol Lett* 10:245–47.

Nakajima M, Imai K, Ito H, Nishiwaki T, Murayama Y, Iwasaki H, Oyama T, Konso T (2005) Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. *Science* 308(5720):414-415.

Smith SM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium *Synechococcus elongatus*. *Proc Natl Acad Sci USA* 103:8564-8569.

Stal LJ, Krumbein WE (1985) Nitrogenase activity in the nonheterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. *Arch Microbiol* 143:67–71.

Stal LJ, Krumbein WE (1987) Temporal separation of nitrogen fixation and photosynthesis in the filamentous, nonheterocystous cyanobacterium *Oscillatoria* sp. *Arch Microbiol* 149:76–80.

Takai N, Nakajima M, Oyama T, Kito R, Sugita C, Sugita M, Kondo T, Iwasaki H (2006). A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria. *Proc Natl Acad Sci U S A* 103(32):12109-14.

Woelfle MA, Ouyang Y, Phanvijhirsiri K, Johnson CH (2004) The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. Curr Biol, 14(16), 1481-1486.

Woelfle MA, Xu Y, Qin X, Johnson CH (2007) Circadian rhythms of superhelical status of DNA in cyanobacteria. *Proc Natl Acad Sci USA* 104:18819-18824.

CHAPTER 2

Oscillations in supercoiling drive circadian gene expression in cyanobacteria

*This chapter contains text and figures from:

Vijayan V, Zuzow R, O'Shea EK (2009). Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proc Natl Acad Sci USA* 106(52):22564-22568.

Abstract

The cyanobacterium *Synechococcus elongatus* PCC 7942 exhibits oscillations in mRNA transcript abundance with 24-hour periodicity under continuous light conditions. The mechanism underlying these oscillations remains elusive – neither *cis* nor *trans*-factors controlling circadian gene expression phase have been identified. Here we show that the topological status of the chromosome is highly correlated with circadian gene expression state. We also demonstrate that DNA sequence characteristics of genes that appear monotonically activated and monotonically repressed by chromosomal relaxation during the circadian cycle are similar to those of supercoiling responsive genes in *E. coli.* Furthermore, perturbation of superhelical status within the physiological range elicits global changes in gene expression similar to those that occur during the normal circadian cycle.

Introduction

Circadian rhythms in gene expression have been identified in many organisms. In general, 5-15% of an organism's transcriptome oscillates with 24-hour periodicity in the absence of external cues such as light to dark or dark to light transitions (Dunlap et al., 2004). These transcriptional rhythms are controlled by an endogenous biological clock and allow organisms to schedule processes at appropriate times during the day and night cycle. The cyanobacterium *Synechococcus elongatus* PCC 7942 (hereafter, *S. elongatus*) is particularly striking because the majority of its gene expression is under circadian control in continuous light conditions. A "promoter trap" analysis using a bacterial luciferase reporter integrated at approximately 30,000 random loci showed

circadian oscillations in bioluminescence at all 800 locations where bioluminescence signal was detected (Liu et al., 2005). A recent measurement of mRNA levels by microarray analysis demonstrated that at least 30% of transcript levels oscillated in circadian fashion (Ito et al., 2009). The discrepancy between promoter trap and microarray analysis is not surprising and is at least partially attributable to a combination of: (1) the limited time-resolution of microarrays and; (2) the time-averaging of transcript levels observed in the bioluminescence output of promoter trap studies due to finite luciferase protein lifetime. The actual percent of circadian transcripts in *S. elongatus* is likely between 30% and 100%; here we observe that 64% of transcripts oscillate with circadian periodicity (Figure 2.1A).

In *S. elongatus*, circadian oscillations in transcriptional activity require three genes, *kaiA*, *kaiB*, and *kaiC*, whose products comprise the core circadian oscillator. The proteins encoded by the *kai* genes interact with one another to generate circadian rhythms in KaiC phosphorylation at serine and threonine residues (Iwasaki et al., 2002). Amazingly, an *in vitro* mixture of the Kai proteins and ATP reproduces *in vivo* oscillations in the phosphorylation state of KaiC (Nakajima et al., 2005). Inactivation of any *kai* gene abolishes circadian oscillations in both transcription and phosphorylation, and when *kaiC* is mutated such that phosphorylation oscillations occur with periods other than 24 hours, the period of transcriptional oscillation is similarly affected (Ito et al., 2009; Iwasaki et al., 2002; Ishiura et al., 1998). Though the link between KaiC phosphorylation and transcription is strong, it is still conditional – when *S. elongatus* is exposed to constant darkness, oscillations in KaiC phosphorylation persist for several

cycles, but transcriptional activity is greatly reduced and no longer appears circadian (Ito et al., 2009; Tomita et al., 2005).

Although a role for the Kai proteins in circadian gene expression is evident, the mechanism of promoter control is unclear. Circadian gene expression in *S. elongatus* is primarily divided into two phases – one subset of transcripts peaks at subjective dawn and the other subset at subjective dusk (Liu et al., 1995; Ito et al., 2009). Systematic analysis of single promoters has yet to identify specific *cis*-elements controlling expression phase (Min et al., 2004; Liu et al., 1996), and neither random mutagenesis nor transposon insertion screens have identified *trans*-factors responsible for expression phase (Min et al., 2000). Interestingly, several heterologous non-circadian promoters from *E. coli* can drive circadian expression when integrated in *S. elongatus*, suggesting that a specific *cis*-sequence is not necessary for circadian promoter activity (Ditty et al., 2003). The elusiveness of *cis* and *trans* factors has prompted the "oscilloid model" for circadian control of gene expression (Mori et al., 2001). In this model, circadian rhythms in the topology of the *S. elongatus* chromosome are thought to impart circadian gene expression patterns by modulating the affinity of the transcription machinery for promoters during the circadian cycle. After this model was proposed, two studies demonstrated circadian oscillations in the compaction and superhelical status of the *S. elongatus* chromosome and an endogenous plasmid, respectively (Smith et al., 2006; Woelfle et al., 2007). Although circadian oscillations in the topology of the chromosome are evident, it is unclear whether these oscillations cause circadian changes in gene expression. To test whether changes in chromosome topology are responsible for generating circadian gene expression, we concurrently measured gene

expression and endogenous plasmid topology. We demonstrate that each topological state corresponds to a distinct state in gene expression. We also observe that DNA sequence features of promoters and coding regions of *S. elongatus* genes that appear monotonically relaxation repressed or monotonically relaxation activated during the circadian cycle are similar to features found in relaxation sensitive genes in *E. coli.* Furthermore, we show that perturbation of topological status results in rapid and predictable changes in global transcript levels.

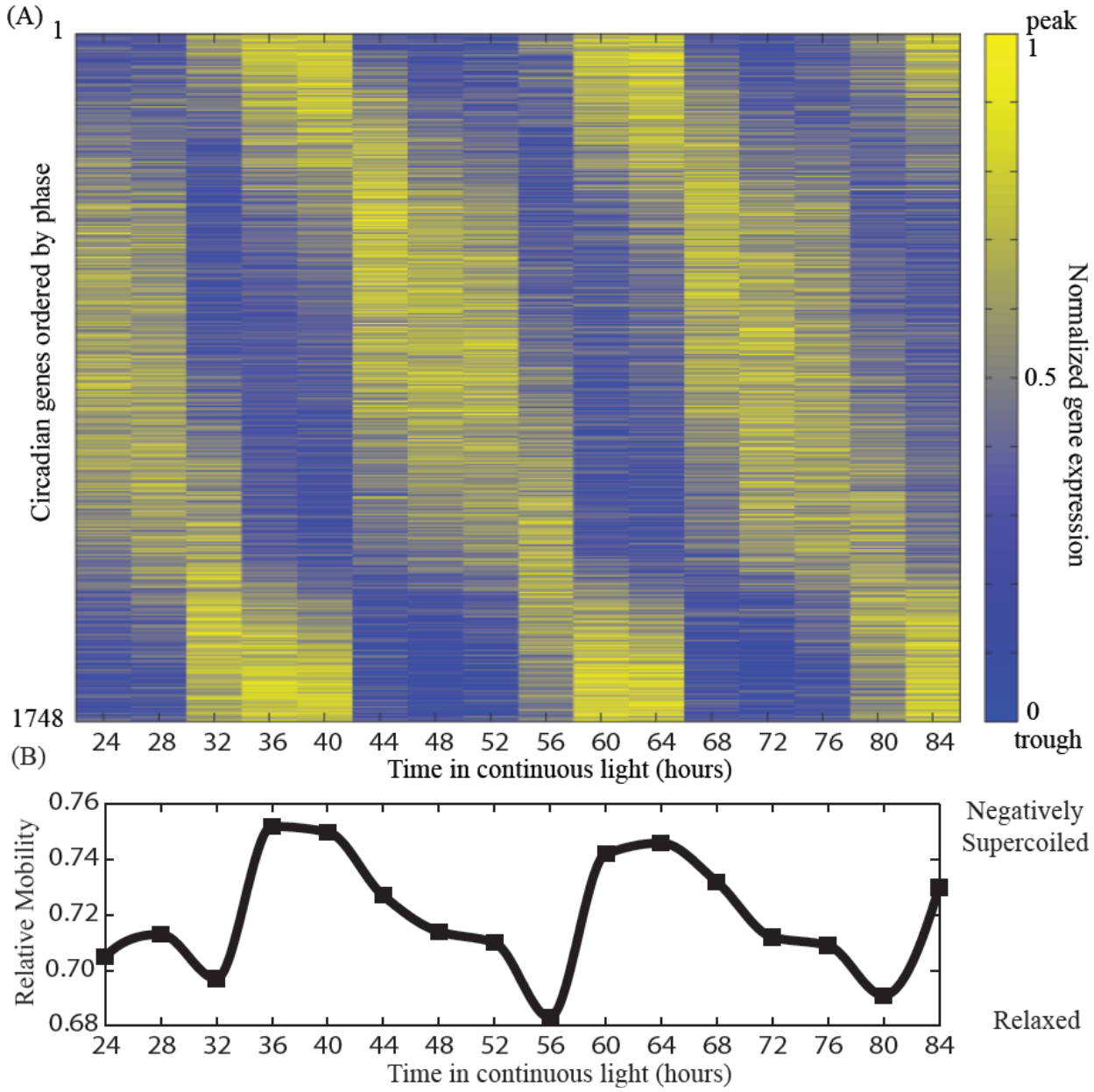


Figure 2.1: Circadian gene expression and topology in *S. elongatus*.

(A) Heat map of circadian gene expression (1748/2724 = 64% of total predicted ORFs) ordered by phase. Each gene is normalized between 0 and 1 such that yellow (1) and blue (0) indicate a peak and trough of expression, respectively. Circadian genes were identified on the chromosome and both endogenous plasmids.

(B) Oscillations in the superhelicity of an endogenous plasmid (pANS) from same time-course as expression analysis. Each superhelical state corresponds to a unique state of gene expression suggesting that oscillations in supercoiling can drive oscillations in gene expression.

Results

Measurement and analysis of circadian gene expression in *S. elongatus*

We analyzed changes in transcript abundance in *S. elongatus* using whole-genome microarrays. Cells were entrained by two consecutive alternating 12 hour light-dark cycles and subsequently released into continuous light (T = 0 hours). Samples were collected every 4 hours for 60 hours from T = 24 to T = 84 hours (Figure S2.1A). Our analysis revealed that 1,748 of 2,724 predicted ORFs, equivalent to 64%, oscillated with 22 to 26-hour period (Figure 2.1A, GSE18902). Hereafter, we refer to these ORFs as circadian genes.

We observed that the majority of circadian genes either peaked in the subjective dawn or subjective dusk, with approximately 30% more genes peaking in the subjective dawn (Figure S2.1B, Figure S2.2) (Ito et al., 2009). Circadian genes existed with amplitudes as high as a 32-fold change in expression, but approximately 90% of oscillating genes had amplitudes less than 2-fold (Figure S2.1C). Surprisingly, a few genes exhibited oscillations with periods less than 24-hours. For example, *rpaA* – encoding a two component transcriptional regulator – has transcript levels that oscillate with a 12-hour period. Genes with a 12-hour period may be inherently 24-hour period genes whose expression is controlled (activated or repressed) by another 24-hour period gene.

To investigate whether chromosomal location influences gene expression we explored the location dependence of gene expression profiles. We found that genes located on the same operon had almost identical gene expression profiles. For example, the Pearson correlation of the temporal profiles of *kaiB* and *kaiC*, two genes

known to exist on the same operon, is 0.99 (Ishiura et al., 1998). In the extreme case of the ribosomal protein cluster, a contiguous 8.5 kb segment featuring 18 genes showed remarkable co-regulation with mean Pearson correlation between neighbors of 0.89 (Ito et al., 2009). Despite the highly correlated expression profiles of neighboring genes, the overall transcription architecture of the chromosome appears random (Figure S2.3A) (Ito et al., 2009). In fact, when the spatial organization of expression profiles is conditioned on putative operon structure, the correlation between neighboring operons is close to random (Figure S2.3B, Figure S2.3C). That is, genes on the same mRNA transcript are highly correlated, but neighboring transcripts show close to no correlation. Similarly, the amplitude distribution along the chromosome appears random after conditioning on putative operon structure (Figure S2.4). In addition to a ~2.7 Mb chromosome, *S. elongatus* contains two endogenous plasmids, pANL and pANS. We see several circadian genes on both endogenous plasmids. The existence of circadian genes on both the chromosome and plasmids suggests that circadian gene expression is regulated by a global mechanism.

To explore the function of circadian gene expression, we investigated the distribution of subjective dawn and subjective dusk genes in various cellular and metabolic pathways (Ito et al., 2009). Interestingly, we found that 41 of 46 circadian genes involved in photosynthesis were most highly expressed in the subjective dawn. Since protein levels have been shown to lag transcript abundance by 4 to 6 hours (Kondo et al., 1993; Liu et al., 1995), these photosynthesis proteins are likely up-regulated during the subjective day and down-regulated during the subjective night. Several other metabolic and cellular pathways show significant enrichment for

subjective dawn or subjective dusk genes (Figure S2.2, Figure S2.5). In total, 70% of genes annotated in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa et al., 2002) exhibit circadian periodicity in their expression. Although it is clear that circadian expression has major metabolic and cellular consequences in continuous light, it is unclear whether and how this control is utilized in more physiologically relevant conditions where light and dark periods alternate.

Correlation between topological state and gene expression

The “oscilloid model” suggests that circadian change in the topology of the chromosome drives circadian gene expression. If this model is true, we expect each topological state to map to a single gene expression state. As a proxy for the topological status of the *S. elongatus* chromosome, we quantified the superhelical status of the endogenous pANS plasmid (Figure 2.1B) sampled from the same time-course as the expression analysis (Figure S2.6) (Woelfle et al., 2007). A comparison of the topological analysis with the microarray data suggests that each topological state corresponds to a unique state of gene expression (Figure 2.1A, Figure 2.1B). That is, circadian times with similar superhelical states generally have similar gene expression states. Interestingly, a dramatic increase in superhelicity occurs eight hours into the circadian cycle (CT 8), allowing each superhelical state to be sampled only once during the circadian cycle. If each gene requires a particular level of supercoiling for maximal transcription, then this increase in superhelicity could explain the predominance of sinusoidal 24-hour period oscillations over oscillations with significant components at smaller periods.

Studies of global expression changes due to chromosomal relaxation in *E. coli* have identified general sequence characteristics of relaxation activated and relaxation repressed genes (Peter et al., 2004; Jeong et al., 2006). To investigate if supercoiling plays a role in circadian gene expression we analyzed the sequence content of genes that appear activated and repressed by relaxation in *S. elongatus.* To identify genes whose expression was highly correlated (relaxation repressed) or anti-correlated (relaxation activated) with relaxation, we computed the Pearson correlation coefficient between each temporal expression profile and the supercoiling waveform. The 250 most strongly correlated genes were designated monotonically relaxation repressed and the 250 most strongly anti-correlated genes were designated monotonically relaxation activated. By selecting the most correlated and most anti-correlated genes, we eliminate the genes whose relationship with supercoiling is non-linear, including the genes for which maximal expression occurs at an intermediate physiological supercoiling state. Analysis of global expression changes due to relaxation in *E. coli* demonstrated that genes monotonically repressed in response to relaxation have a lower than expected AT content in both the coding region and promoter, whereas genes monotonically activated by relaxation have a higher than expected AT content (Peter et al., 2004; Jeong et al., 2006). Similarly, we see a significant difference in the AT content of genes that appear monotonically repressed versus monotonically activated by relaxation (Figure 2.2A). This difference in AT content is highly significant in both the promoter region and coding region (all $p < 0.05$) (Figure S2.7).

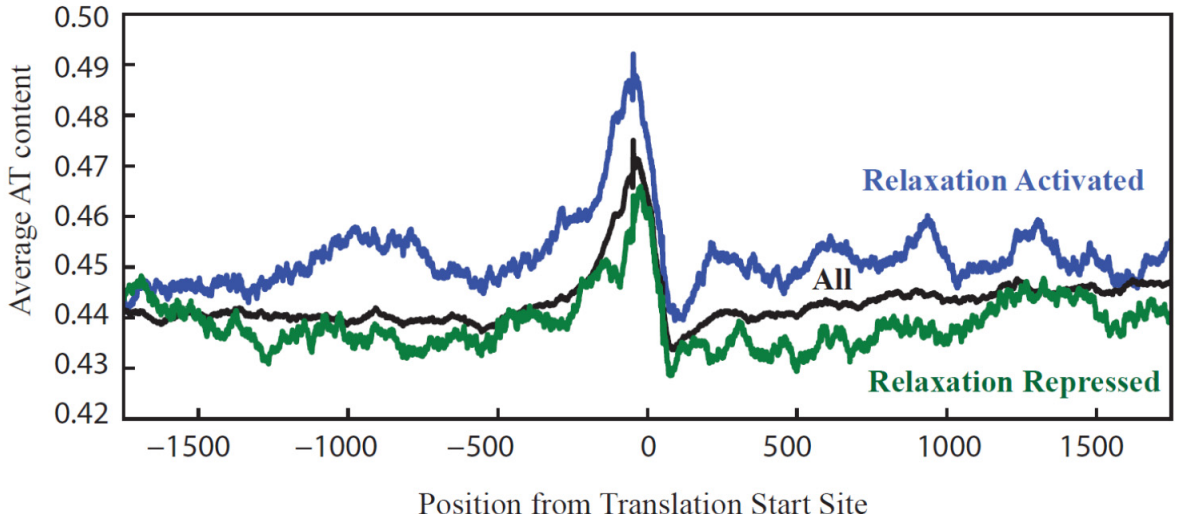


Figure 2.2: Sequence characteristics of genes suggest supercoiling mediated control.

Genes that are monotonically relaxation activated (blue) and monotonically relaxation repressed (green) have increased and decreased AT content (in promoter and coding region), respectively, compared to all the genes in the genome. Genes were aligned by translational start site and the mean AT content along -2000 to 2000 was calculated with a smoothing window of 100 nucleotides.

Perturbation of topology results in predictable expression changes

If superhelical state does in fact regulate circadian gene expression, then artificial manipulation of supercoiling should result in a predictable change in gene expression. In particular, if we are able to rapidly change the superhelical status of the chromosome from its most supercoiled to most relaxed state, then the instantaneous change in gene expression should be very similar to the change in gene expression caused by normal circadian relaxation. That is, the expression of genes that are expressed most strongly during supercoiled circadian times (Figure 2.3A, purple) should immediately decrease, whereas the expression of genes that are expressed most strongly during relaxed circadian times (Figure 2.3A, red) should immediately increase.

To test this hypothesis, we treated cells with the gyrase inhibitor novobiocin (0.1 μg/ml novobiocin sodium salt) during the most supercoiled state in the circadian cycle (CT 16, T = 64). At this concentration of novobiocin, the growth rate is not significantly affected and supercoiling resumes the circadian course after a brief excursion. Within 5 minutes of novobiocin treatment, the pANS plasmid reached a supercoiling state similar to the most relaxed state observed during the normal circadian cycle (Figure 2.3B, Figure S2.8AB, GSE18902). As expected, expression from the well-characterized supercoiling sensitive promoters in *E. coli* – *gyrA*, *gyrB*, and *topI* – exhibited a homeostasis-generating response upon relaxation (Menzel et al., 1983; Figure S2.8C). That is, since gyrase induces negative supercoils and topoisomerase I induces positive supercoils (relaxation), the expression of gyrase increases and the expression of topoisomerase I decreases in response to relaxation to regain the original superhelical state. Amazingly, genes whose expression is decreased or increased with circadian

relaxation took opposite and predictable trajectories after novobiocin-induced relaxation (Figure 2.3C). In fact, there is high correlation (Pearson correlation coefficient 0.85) between the change in gene expression after only 5 minutes of drug-induced relaxation and the change in gene expression due to relaxation in the circadian cycle (Figure 2.4). Of the 1748 circadian genes, 1380 (79%) moved in the expected direction 5 minutes after novobiocin treatment. Of the genes whose expression differed by more than 2-fold between relaxed and supercoiled circadian time-points, none moved in the direction contrary to that expected. These results demonstrate that direct modulation of supercoiling within the physiological regime elicits the expected changes in expression, suggesting that oscillations in superhelicity may be sufficient to impart circadian gene expression.

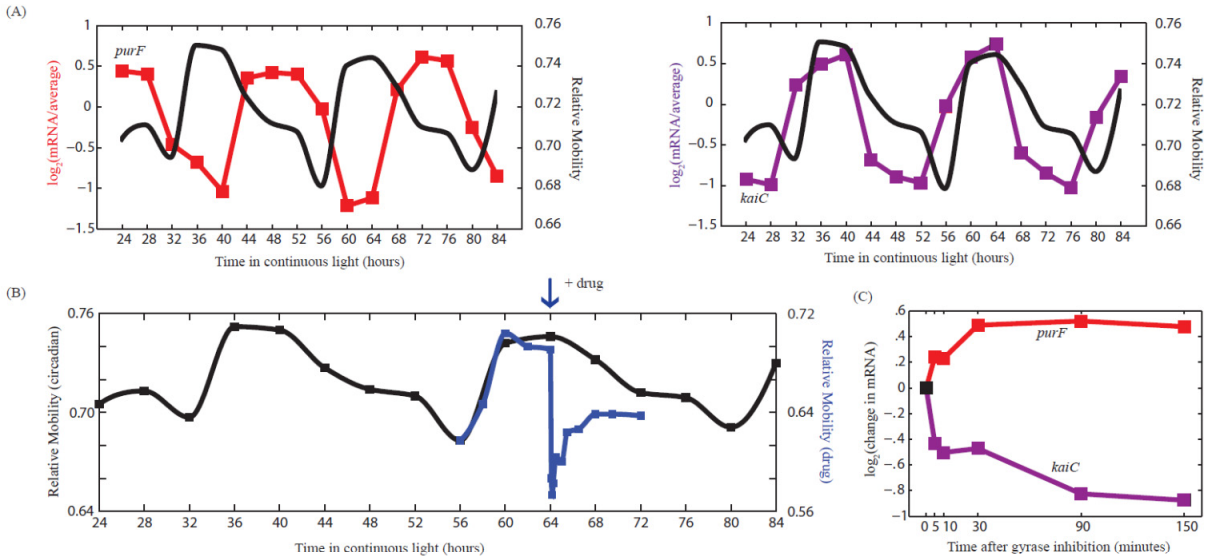


Figure 2.3: Manipulation of supercoiling via gyrase inhibition results in expected changes in gene expression.

(A) Canonical subjective dawn (*purF*, red) and subjective dusk (*kaiC*, purple) genes oscillate with opposite phase in gene expression. Superimposed is the circadian oscillation in superhelicity of the endogenous pANS plasmid (black). Relaxed supercoiling states have lower relative mobility than negatively supercoiled states. *KaiC* expression is decreased during circadian relaxation while *purF* expression is increased.

(B) Quantification of pANS plasmid supercoiling prior to and after addition of novobiocin at T = 64 hours (blue) superimposed on supercoiling changes during the circadian cycle (black). Supercoiling was measured 5, 10, 15, 30, 60, 90, 150, 240, 330, and 480 minutes after novobiocin addition. Novobiocin addition immediately relaxes the pANS plasmid to a level similar to the most relaxed state during the circadian cycle.

(C) Gyrase inhibition at T = 64 hours in the circadian cycle results in an immediate change in gene expression due to drug-induced relaxation. Genes that have higher expression during relaxed circadian times immediately increase in gene expression (*purF*, red) whereas genes that have lower expression during relaxed circadian times immediately decrease in gene expression (*kaiC*, purple).

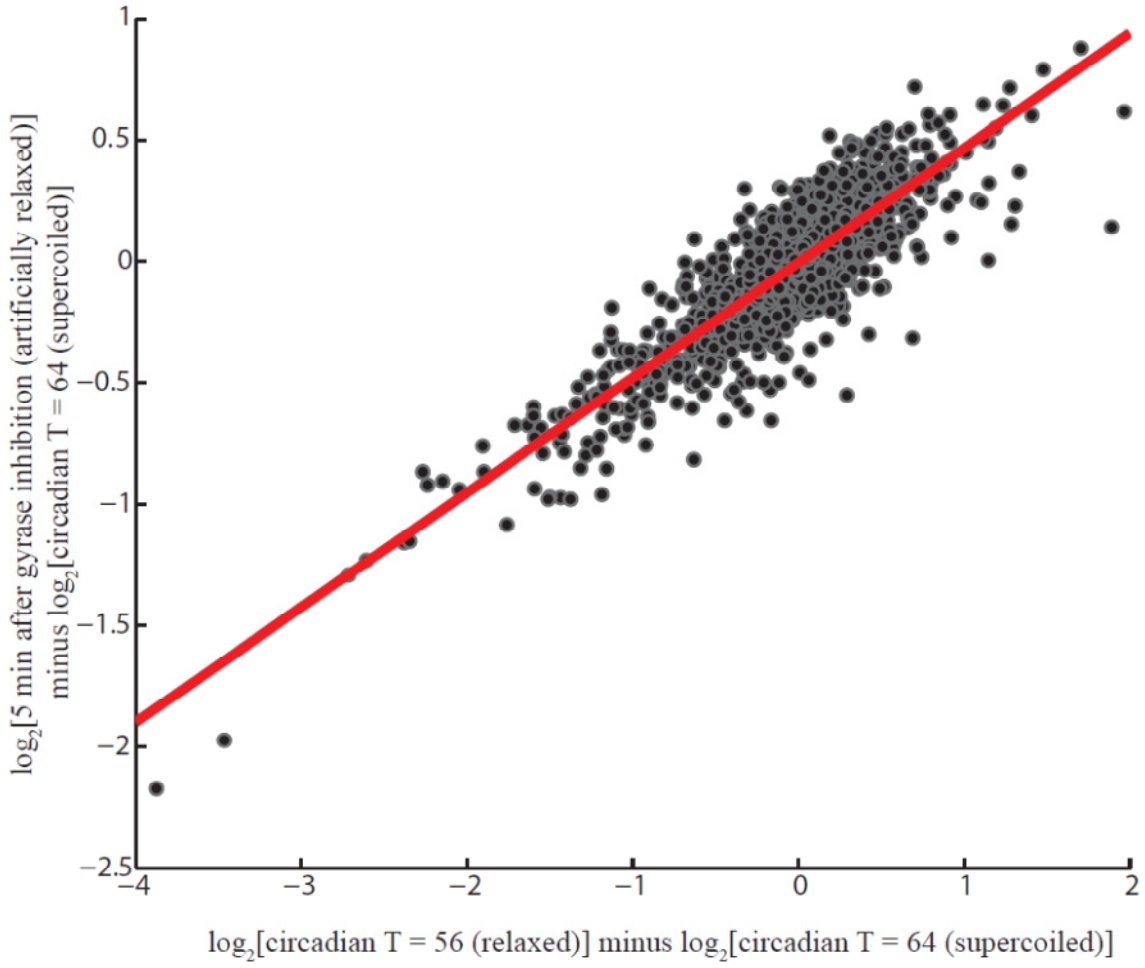


Figure 2.4: Genome-wide changes in expression due to gyrase inhibition induced relaxation.

Scatter plot of the change in gene expression due to circadian relaxation versus drug-induced relaxation for circadian genes. The normal relaxation induced change in gene expression (x-axis) is highly correlated with the rate of gene expression due to drug-induced relaxation (y-axis) suggesting that supercoiling plays a primary role in dictating circadian gene expression. The linear best fit line is shown in red with Pearson correlation coefficient 0.85.

Discussion

By measuring genome-wide gene expression and superhelicity of an endogenous plasmid, we demonstrate that each superhelical state corresponds to a unique state of expression. In addition, we show that genes monotonically repressed and monotonically activated by relaxation are depleted and enriched for AT content, respectively, similar to observations made for relaxation sensitive genes in *E. coli* (Peter et al., 2004; Jeong et al., 2006). Further, we show that the instantaneous change in gene expression following drug-induced relaxation closely correlates with the change resulting from relaxation during the normal circadian cycle.

But what is controlling the rhythmicity in supercoiling? Since we know that circadian oscillations in gene expression and supercoiling are dependent on KaiC (Ito et al., 2009; Ishiura et al., 1998; Smith et al., 2006; Woelfle et al., 2006), the primary hypothesis is that KaiC either directly or indirectly controls supercoiling. KaiC sequence analysis suggests that it belongs in the RecA/DnaB superfamily (Leipe et al., 2000) and KaiC has been shown to have DNA binding activity *in vitro* (Mori et al., 2002), allowing for the possibility that KaiC may be able to directly change the superhelicity of DNA (Woelfle et al., 2007). An alternative model is that KaiC controls the ratio of negative versus positive supercoiling generating activity (Woelfle et al., 2007). For example, KaiC or an interacting protein may transcriptionally or post-transcriptionally control an enzyme that modulates supercoiling or affects the [ATP] / [ADP] ratio. In *E. coli*, changes in supercoiling are often accompanied by changes in the [ATP] / [ADP] ratio (Hseih et al., 1991; Hseih et al., 1991; Camacho-Carranza et al., 1995), and *in vitro,* the [ATP] / [ADP] ratio has been shown to modulate the activity of gyrase (Westerhoff et al., 1988).

Supercoiling has been shown to play an important role in transcription regulation in *E. coli* (Perez-Martin et al., 1994). Superhelicity imposes torsional stress on DNA, which affects transcription by modulating the stability of interactions between RNA polymerase, nucleoid-like proteins, and transcription factors with the promoter or coding region of genes (Perez-Martin et al., 1994). Changes in the level of supercoiling have been observed during a variety of stress responses – temperature, peroxide, and osmotic – as well as growth conditions (Hseih et al., 1991; Hseih et al., 1991; Camacho-Carranza et al., 1995; Rui et al., 2003), and changes in supercoiling have been shown to elicit a global transcriptional response (Cheung et al., 2003). A role for supercoiling in mediating changes in expression is not limited to *E. coli* and has been observed in a variety of bacteria including *Salmonella typhimurium, Vibrio cholera,Bacillus subtilis* and *Synechocystis* PCC 6803 (Higgins et al., 1998; Parsot et al., 1992; Alice et al., 1997; Prakash et al., 2009).

Supercoiling mediated control of circadian gene expression poses several advantages over other modes of regulation: (1) supercoiling can globally affect all promoters, enabling genome-wide oscillations in transcription; (2) promoters can have differential sensitivity to supercoiling, allowing different phases of circadian oscillation; and (3) supercoiling is already under feedback control in bacteria, so evolution of an oscillator requires only a circadian perturbation of supercoiling levels since the return to homeostasis is already pre-programmed.

In *S. elongatus,* the extent of supercoiling mediated circadian gene expression, natural variation in superhelical status, and robust outputs – amplitude and phase – of supercoiling sensitivity make it an ideal system for understanding how supercoiling

affects transcription. Although several mechanisms have been proposed for how supercoiling affects the transcription of individual promoters in *E. coli*, genome-wide analysis in *S. elongatus* may reveal the common mechanistic and sequence motifs of supercoiling mediated transcriptional control. By studying how sequence determines phase and how the association of transcription machinery with promoters and genes is modulated during the circadian cycle, we may be able to gain a better understanding of how supercoiling can affect transcription.

Materials and Methods

Continuous Culture of Cyanobacteria

A continuous culture apparatus was developed to keep cells in constant conditions and provide real-time bioluminescence readings. *S. elongatus* (strain AMC 408 (Min et al., 2004): *psbAI::luxCDE* fusion in NS1 (Andersson et al., 2000) (spectinomycin) and *purF::luxAB* fusion in NSII (Andersson et al., 2000) (chloramphenicol)) was grown in a 6 liter cylindrical spinner flask (Corning) at a volume of 4.5 liters. Cells were grown in modified BG-11 medium (Bustos et al., 1991) with the following modifications: 0.0010 g/liter of $FeNH_4$ citrate was used instead of 0.0012 g/liter of $FeNH_4$ citrate and citric acid was supplemented at 0.00066 g/liter. Cells were initially inoculated in the presence of antibiotics (5 µg/ml spectinomycin and 5 µg/ml chloramphenicol), and subsequently diluted with modified BG-11 only. Cells were exposed to surface flux of ~25 µmol photons m^{-2} s^{-1} white light, bubbled with 500 ml min^{-1} 1% CO_2 in air, maintained at 30º C, and stirred at 1 rotation per second. Constant

optical density (OD_{750} 0.15) and volume are achieved via a two state controller. OD does not fluctuate more than 8% during an experiment.

Cells are exposed to two light/dark cycles for entrainment prior to release into continuous light, collected at designated times by vacuum filtration, snap frozen in liquid nitrogen, and stored in -80º C. Duplicate 120 ml cultures were harvested at each time-point – one for the supercoiling assay and one for the expression analysis.

Expression Microarrays

RNA was extracted from frozen cells in two steps. First, cells were lysed in 65° C phenol/SDS (*sodium dodecyl* sulfate) by vortexing and the total RNA was purified by phenol/chloroform extraction. Second, total RNA was subjected to DNase I (Promega) treatment followed by a second phenol/chloroform extraction. Total RNA was analyzed on agarose gel and Agilent Bioanalyzer for integrity.

Expression was measured using custom designed two-color 8x15k microarrays (Agilent, Array ID 020846). Microarrays were designed using genome and plasmid sequences from Genbank CP000100, CP000101, and S89470. Four separate melting temperature matched probes (80° C) for each predicted ORF were designed using eArray (Agilent). All probes less than 60 nucleotides were appended with the Agilent standard linker to reach 60 nucleotides. Probes against *luxA* through *luxE* and *Arabidopsis* spike-in controls (Ambion) were also included on the microarray.

cDNA was prepared for each individual time-point (foreground channel) as well as for a pool of all time-points (background channel). Spike-in RNA was introduced at different concentrations and ratios to the foreground and background channels prior to

reverse transcription in order to ensure proper ratio detection in a wide dynamic range. 5 μg total RNA (plus spike-ins) was reverse-transcribed with random 15-mer primers (Operon) and a 2:3 ratio of amino allyl-UTP:dTTP (Sigma) using SuperScript III reverse-transcriptase (Invitrogen) without amplification. RNA was hydrolyzed and cDNA was purified using Microcon 30 spin column (Millipore).

cDNA was labeled with N-Hydroxysuccinimide-ester Cyanine 3 (Cy3, foreground) or Cyanine 5 (Cy5, background) (GE Biosciences) in 0.1 M sodium bicarbonate pH 9.0 for 6 hours. Labeled cDNA was purified (Microcon 30, Millipore) in preparation for hybridization. Each array was hybridized with 150 to 300 ng Cy3 and 150 to 300 ng Cy5 labeled cDNA and rotated (5 rotations min^{-1}) at 60º C for 17 hours in SureHyb chambers (Agilent). Arrays were subsequently washed in 6.7X SSPE and 0.005% *N-lauryl* sarcosine buffer for at least 1 minute, 0.67X SSPE and 0.005% *N-lauryl* sarcosine buffer for 1 minute, and then Agilent drying and ozone protection wash for 30 seconds at room temperature (1X SSPE = 0.15 M NaCl, 10 μM sodium phosphate, 1 mM EDTA, pH 7.4). The arrays were immediately scanned using an Axon 4000B scanner at 5 μm resolution. The average intensity of the Cy3 and Cy5 florescence at each spot was extracted using the GenePix software (Molecular Devices) and loaded into MATLAB (Mathworks). Loess and quantile normalization were performed using the MATLAB bioinformatics toolbox. All subsequent analysis was performed in MATLAB. Spike-in ratios were compared from array to array to ensure proper normalization and hybridization. Constant spike-in signal also ensures that rRNA to mRNA ratio is constant during all experimental conditions.

Identification and Classification of Circadian Genes

The 4 melting temperature matched probes per predicted ORF were normalized and averaged. A modified Cosiner method was used to identify cycling genes and their period, phase, and amplitude (Kucho et al., 2005). Briefly, the expression data for each predicted ORF was linearly de-trended and the first Fourier component was calculated assuming periods between 12 and 36 with increment of 0.1. The period that minimized the Euclidian distance between the first Fourier component and the experimental signal was designated the period of the predicted ORF. If the period was between 22 and 26, the ORF was considered circadian and its phase, period, and amplitude were given by the first Fourier component. To separate subjective dawn and subjective dusk genes as well as genes that peaked at particular times during the circadian cycle, the circadian genes were subjected to K-means clustering by Euclidean distance. Clustering with K = 6, generated clusters that peaked at unique times during the circadian cycle (CT 0, 4, 8, 12, 16, and 20). Genes peaking at CT 20, 0, and 4 were designated subjective dawn genes and genes peaking at CT 8, 12, and 16 were designated subjective dusk genes (Figure S2.2).

Figure S2.9 compares the data from this study with Ito et al. (Ito et al., 2009). Here, we identify as circadian 86% of the genes identified as circadian in Ito et al. (Figure S2.9A). In addition, for the genes which both studies identified as circadian, the phase of expression was almost identical (Figure S2.9B).

Isolation of Plasmid DNA and Chloroquine Agarose Gel Electrophoresis (CAGE)

Plasmid isolation and CAGE technique were performed similarly to (Woelfle et al., 2007). Endogenous *S. elongatus* plasmid (pANS) was isolated using QIAprep

miniprep kit (Qiagen). Approximately 150 ng of isolated plasmid was run on a gel containing 0.8% agarose, 0.5X TBE, and 10 μg/ml chloroquine diphosphate (1X TBE = 90 mM Tris, 64.5 mM boric acid, 2.5 mM EDTA, pH 8.3). Plasmid topoisomers were separated by electrophoresis at 50 volts / 14 cm for 24 hours in the dark at room temperature with buffer exchange. Gels were incubated for one hour in 0.5X TBE and 1X SybrGold stain (Invitrogen). Gels were imaged on Typhoon Imager (GE). Chloroquine titration and gyrase inhibition (novobiocin treatment) were utilized to determine the position of relaxed and supercoiled forms on gel. Southern blotting was utilized to prove that topoisomers visualized by SybrGold staining were from the pANS plasmid. Blotting was performed by upward capillary transfer (Sambrook et al., 2001) onto Biodyne B nylon membrane (Pierce) and visualized by horseradish peroxidase-streptavidin chemiluminesence assay (Pierce). Probe template was synthesized by amplification with PCR primers pANSorfAfor (Woelfle et al., 2007), 5′-GGAAGTAGAAGGCTT-3′, and pANSorfArev (14), 5′-GGCAATGGCCAGCATCG-3′. The 468 nucleotide PCR product was amplified with random heptanucleotide primers in the presence of biotin-11-dUTP using random priming kit (Pierce). Gels were analyzed using ImageJ and MATLAB software. From densitometry traces, the relative mobility was calculated by the position of the mean of the Boltzmann distribution of topoisomers relative to the position of the open circular (oc) and relaxed (rel) forms. RM = 1 - (mean – oc) / (oc - rel).

Abbreviations

T, experimental time where T = 0 is defined as time of release into continuous light; CT, circadian time; CAGE, chloroquine agarose gel electrophoresis.

Data Deposition

The microarray data has been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) Database, www.ncbi.nlm.nih.gov/geo (accession no. GSE18902).

Acknowledgments

We thank members of the O'Shea lab for discussion and commentary. We thank Dr. Susan Golden for the *S. elongatus* strain AMC 408. This work was supported by the Howard Hughes Medical Institute. VV was supported by National Defense Science and Engineering and National Science Foundation Graduate Research Fellowships.

References

Alice AF, Sanchez-Rivas C (1997) DNA Supercoiling and Osmoresistance in *Bacillus subtilis* 168. *Curr Microbiol* 35:309-315.

Andersson CR et al. (2000) Application of bioluminescence to the study of circadian rhythms in cyanobacteria. *Methods Enzymol* 305:527-542.

Bustos SA, Golden SS (1991) Expression of the psbDII Gene in Synechococcus sp. Strain PCC 7942 Requires Sequences Downstream of the Transcription Start Site. *J Bacteriol* 173:7525-7533.

Camacho-Carranza R, et. al (1995) Topoisomerase Activity during the Heat Shock Response in *Escherichia coli* K-12. *J Bacteriol* 177:3619-3622.

Cheung KJ, Badarinarayana V, Selinger DW, Janse D, Church GM (2003) A Microarray-Based Antibiotic Screen Identifies a Regulatory Role for Supercoiling in the Osmotic Stress Response of *Escherichia coli. Genome Res* 13:206-215.

Ditty JL, Williams SB, Golden SS (2003) A cyanobacterial circadian timing mechanism. *Annu Rev Genet* 37:513–543.

Dunlap JC, Loros JJ, DeCoursey PJ (2004) Chronobiology – Biological Timekeeping. (Sinauer Associates, Sunderland, MA).

Higgins CF et al. (1988) A physiological role for DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli*. *Cell* 52:569-584.

Hseih LS, Rouviere-Yaniv J, Drlica K (1991) Bacterial DNA supercoiling and [ATP]/[ADP] ratio: changes associated with salt shock. *J Bacteriol* 12:3914-1917.

Hseih LS, Burger RM, Drlica K (1991) Bacterial DNA supercoiling and [ATP]/[ADP]. Changes associated with a transition to anaerobic growth. *J Mol Bio* 219:443-450.

Ishiura M, et al. (1998) Expression of a Gene Cluster KaiABC as a Circadian Feedback Process in Cyanobacteria. *Science* 281:1519-1523.

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.

Iwasaki H, Nishiwaki T, Kitayama Y, Nakajima M, Kondo T (2002) KaiA-stimulated KaiC phosphorylation in circadian timing loops in cyanobacteria. *Proc Natl Acad Sci USA* 99:15788-15793.

Jeong KS, Xie Y, Hiasa H, Khodursky AB (2006) Analysis of Pleiotropic Transcriptional Profiles: A Case Study of DNA Gyrase Inhibition. *PLoS Genet* 2:1464-1476.

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30:42–4.

Kondo T et al. (1993) Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria. *Proc Natl Acad Sci USA* 90:5672–76.

Kucho K et al. (2005) Global Analysis of Circadian Gene Expression in the Cyanobacterium *Synechocystis sp.* Strain PCC 6803. *J Bacteriol* 187:2190-2199.

Leipe DD, Aravind L, Grishin NV, Koonin EV (2000) The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome Res* 10:5-16.

Liu Y, Golden SS, Kondo T, Ishiura M, Johnson CH (1995) Bacterial luciferase as a reporter of circadian gene expression in cyanobacteria. *J Bacteriol* 177:2080-2086.

Liu Y, et al. (1995) Circadian orchestration of gene expression in cyanobacteria. *Genes Dev* 9:1469-1478.

Liu Y, Tsinoremas NF, Golden SS, Kondo T, Johnson CH (1996) Circadian expression of genes involved in the purine biosynthetic pathway of the cyanobacterium *Synechococcus* sp. Strain PCC 7942. *Mol Microbiol* 20:1071-1081.

Menzel R, Gellert M (1983) Regulation of the genes for E. coli DNA gyrase: Homeostatic control of DNA supercoiling. *Cell* 35:105-113.

Min H, Golden SS (2000) A New Circadian Class 2 Gene *opcA,* Whose Product is Important for Reductant Production at Night in *Synechococcus elongatus* PCC 7942. *J Bacteriol* 182:6214-6221.

Min H, Liu Y, Johnson CH, Golden SS (2004). Phase Determination of Circadian Gene Expression in Synechococcus Elongatus PCC 7942. *J Biol Rhythms* 19:103-112.

Mori T, Johnson CH (2001) Circadian programming in cyanobacteria. *Semin Cell Dev Biol* 12:271-278.

Mori T et al. (2002) Circadian clock protein KaiC forms ATP-dependent hexameric rings and binds DNA. *Proc Natl Acad Sci USA* 99:17203-17208.

Nakajima M, et al. (2005) Reconstitution of Circadian Oscillation of Cyanobacterial KaiC Phosphorylation in Vitro. *Science* 308:414-415.

Parsot C, Mekalanos JJ (1992) Structural Analysis of the *acfA* and *acfD* Genes of *Vibrio Cholerae*: Effects of DNA Topology and Transcriptional Activators on Expression. *J Bacteriol* 174:5211-5218.

Perez-Martin J, Rojo F, Lorenzo V (1994) Promoters Responsive to DNA Bending: a Common Theme in Prokaryotic Gene Expression. *Microbiol Rev* 58:268-290.

Peter BJ et al. (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol* 5:R87.

Prakash JS et al. (2009) DNA supercoiling regulates the stress-inducible expression of genes in the cyanobacterium *Synechocystis. Mol Biosyst* 5(12):1904-1912.

Rui S, Tse-Dinh YC (2003) Topoisomerase function during bacterial responses to environmental challenge. *Front Biosci* 8:256-263.

Sambrook J, Russell DW (2001) Molecular Cloning – A Laboratory Manual. (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), 3rd Ed.

Smith SM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium *Synechococcus elongatus*. *Proc Natl Acad Sci USA* 103:8564-8569.

Tomita J, Nakajima M, Kondo T, Ishiura M (2005). No Transcription-Translation Feedback in Circadian Rhythm of KaiC Phosphorylation. *Science* 307:251-254.

Westerhoff HV, O'Dea MH, Maxwell A, Gellert M (1988) DNA supercoiling by DNA gyrase. A static head analysis. *Cell Biophys* 12:157-181.

Woelfle MA, Xu Y, Qin X, Johnson CH (2007) Circadian rhythms of superhelical status of DNA in cyanobacteria. *Proc Natl Acad Sci USA* 104:18819-18824.

CHAPTER 3

A high resolution map of a cyanobacterial transcriptome

*This chapter contains text and figures from:

Vijayan V, Jain IH, O'Shea EK (2011). A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47.

Abstract

Previous molecular and mechanistic studies have identified several principles of prokaryotic transcription, but less is known about the global transcriptional architecture of bacterial genomes. Here we perform a comprehensive study of a cyanobacterial transcriptome, that of *Synechococcus elongatus* PCC 7942, generated by combining three high-resolution data sets: RNA sequencing, tiling expression microarrays, and RNA polymerase chromatin immunoprecipitation sequencing. We report absolute transcript levels, operon identification, and high-resolution mapping of 5′ and 3′ ends of transcripts. We identify several interesting features at promoters, within transcripts and in terminators relating to transcription initiation, elongation, and termination. Furthermore, we identify many putative non-coding transcripts. We provide a global analysis of a cyanobacterial transcriptome. Our results uncover insights that reinforce and extend the current views of bacterial transcription.

Introduction

Over the past few decades considerable progress has been made in understanding the mechanisms and regulation of bacterial transcription. However, relatively few studies have attempted to identify the prevalent features of bacterial transcription *de novo* using an unbiased genome-wide approach. This approach to analyzing the bacterial transcriptome may not only help reinforce the progress made from traditional molecular and mechanistic studies, but may also identify new global features in transcription that have previously been underappreciated.

The advent of next-generation sequencing allows for a complete characterization of bacterial genomes that was previously not possible. RNA sequencing gives unprecedented insights into transcription unit architecture, while RNA polymerase chromatin immunoprecipitation (ChIP) sequencing reveals the flow of information into the transcriptome. We provide a comprehensive analysis of a cyanobacterial transcriptome - that of *Synechococcus elongatus* PCC 7942 - integrating data from RNA sequencing, tiling expression microarrays, and RNA polymerase (RNA pol) ChIP sequencing.

The unicellular cyanobacterium *S. elongatus* PCC 7942 is a genetically tractable model organism for prokaryotic photosynthesis (Frenkel et al., 1950), bioenergy production, and circadian rhythms (Kondo et al., 1993). The circadian clock of *S. elongatus* is built on a three-protein central oscillator that controls the global rhythmic expression of the majority of the genome (Vijayan et al., 2009; Ito et al., 2009). Our transcriptome characterization will facilitate the further use of *S. elongatus* as a model organism.

Results and Discussion.

The transcriptome

We used RNA sequencing, tiling expression microarrays, and RNA pol ChIP sequencing to interrogate transcription in the cyanobacterium *S. elongatus*. RNA was isolated at 4-hour intervals from circadian free-running cells grown in constant light conditions and RNA from a pool of circadian timepoints was sequenced (Materials and methods). Strand-specific RNA sequencing was performed on the Illumina platform

yielding over 22 million uniquely mappable non-rRNA reads and over 620 million nucleotides of coverage, strand-specifically covering each nucleotide of the approximately 2.7 Mb genome an average of approximately 115 times (Ingolia et al., 2009) (Materials and methods). Agilent two-color microarrays with a total of approximately 488,000 strand-specific 60-nucleotide probes spaced every 12 nucleotides were hybridized with cDNA from individual circadian timepoints to supplement RNA sequencing analysis (Materials and methods). RNA pol ChIP sequencing of subjective dawn and subjective dusk circadian timepoints was performed on the Illumina platform, yielding a total of over 19 million uniquely mappable reads, covering each nucleotide over approximately 1,055 times after extension of reads by 150 bp to cover the average length of sequenced DNA fragments (Materials and methods). All analysis of RNA pol ChIP was performed on the combination of the two circadian timepoints unless otherwise specified.

The RNA sequencing and RNA pol ChIP sequencing profiles demonstrate that the transcription landscape in *S. elongatus* is rather dense with very small inter-transcript regions (Figure 3.1A). Assuming a relatively strict cutoff of at least two reads per nucleotide for transcription, approximately 88% of the genome is transcribed on either the plus or minus strand, and approximately 55% of each strand is transcribed (Materials and methods). Approximately 82% of all non-coding sequence is transcribed on either the plus or minus strand, highlighting the density of transcription in *S. elongatus*. Fewer than 10% of the 2,612 chromosomally encoded Joint Genome Institute (JGI) predicted ORFs have negligible transcription (less than a mean of two reads per nucleotide across the ORF), and the remaining ORFs have absolute

expression distributed over a dynamic range of nearly 10,000. In this study we only sample standard exponential growth conditions during circadian free-run in constant light conditions; both transcription density and the number of expressed ORFs are likely to be higher if multiple growth conditions are sampled.

RNA sequencing affords high-resolution determination of the 5′ and 3′ ends of each transcription unit. Transcription units were defined using *a priori* knowledge of JGI ORF, tRNA, and rRNA annotations (Materials and methods). A total of 1,473 transcription units were identified, 1,415 of which were designated as mRNA transcripts as they are devoid of tRNA or rRNA and contain at least one JGI annotated ORF. 5′ and 3′ ends were determined for all transcripts and all subsequent analysis is performed on the subset defined as mRNA transcripts (Figure S3.1, Materials and methods). Highly expressed transcripts show particularly clear 5′ and 3′ boundaries of transcription, each with an associated peak in RNA pol occupancy as measured by RNA pol ChIP (Figure 3.1B). The RNA pol ChIP data are characterized by the presence of several large peaks that tend to be located near the 5′ end of transcripts, and many smaller peaks that tend to be located either at the 3′ end of highly expressed transcripts or within transcripts (Figure S3.2). Surprisingly, most 5′ RNA pol peaks are situated within the transcript rather than at the promoter. Sequence analysis of RNA pol peak positions reveals enrichment for the central AT nucleotides of the highly iterated palindrome 1 (HIP1) site, 5′ GCGATCGC 3′, at the RNA pol peak maximum (P < 1e-10, binomial cumulative distribution). The HIP1 palindrome is highly over-represented in many cyanobacteria, including *S. elongatus* - it appears 185 times more frequently in the *S. elongatus* chromosome than expected for a random 8-mer sequence, but its function is unknown

(Robinson et al., 1995). It is known that the HIP1 motif is a target of methylation in some cyanobacteria (Scharnagl et al., 1998), raising the possibility of an intriguing link between DNA methylation and transcription. Although RNA pol peaks are enriched at the HIP1 site, fewer than 1% of HIP1 sites (41 of 7,402) are situated at an RNA pol peak, and fewer than 2% of RNA pol peaks (41 of 2,159) are situated at HIP1 sites. Despite the fact that only 41 HIP1 sites are occupied by RNA pol, the probability of having at least this many sites occupied by chance is less than 1e-10 (binomial cumulative distribution).

One of the benefits of RNA sequencing is the ability to infer absolute mRNA transcript levels (Figure 3.2A). We calculated the absolute expression of each mRNA per cell, assuming a total of 1,500 mRNAs per cell (Ingraham et al., 1983; Raniguchi et al., 2010) (Materials and methods). We find that using this estimate, over 80% of mRNA transcripts are present at fewer than one copy per cell, suggesting an enormous diversity in single-cell transcriptome profiles and the potential for stochastic effects to play a substantial role in bacterial gene expression. Even if the estimated number of mRNAs per cell is four times larger (6,000 per cell), still nearly half (46%) of mRNAs are present at less than one copy per cell. Although an enormous amount of diversity in mRNA exists in each cell at any given time, the relatively rapid mRNA decay rates in cyanobacteria (Steglich et al., 2010) - median 2.4 minutes in *Prochlorococcus* MED 4 - allow for rapid transcriptome turnover. The distribution of mRNAs per cell appears approximately log-normal with a dynamic range of almost 10,000. Most mRNAs fall within a smaller dynamic range of approximately 100, with a tail of higher expressed transcripts. The bottom part of the distribution was cut at 2^{-4} because transcripts below

this level are almost undetectable at our sequencing coverage (Materials and methods). The highest expressed KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2002) categories include photosynthesis, ribosome, and RNA polymerase, with *P*-values of 2.6e-20, 1.3e-20, and 0.001, respectively (two-sided Wilcoxon rank sum test). The lowest expressed KEGG categories include mismatch repair, homologous recombination, and nucleotide excision repair - ORFs that may not be expressed in standard growth conditions (all $P < 0.002$, two-sided Wilcoxon rank sum test). Absolute transcript levels are generally correlated (Pearson correlation, $r = 0.65$) with RNA pol occupancy (Figure 3.2B), suggesting that transcription and not decay is the primary determinant for setting absolute transcript abundance. The variation (approximately one order of magnitude scatter) observed is roughly proportional to the expected distribution of mRNA decay rates in cyanobacteria (Steglich et al., 2010). However, this variation may also arise from: (1) different RNA pol elongation rates for different transcripts; (2) variable amounts of RNA pol pausing for different transcripts; and/or (3) lack of strand-specific information in the RNA pol ChIP data.

Of the 1,415 mRNA transcripts identified, many (approximately 38%) have more than one ORF per transcript (Figure 3.2C). Most mRNAs contain only one or two ORFs, but the ribosomal protein operon presents an extreme case of 31 ORFs on a transcript spanning over 17,000 nucleotides. Our operon identification via RNA sequencing shows good correlation with bioinformatic operon predictions from MicrobesOnline (Dehal et al., 2010; Price et al., 2005) (Figure 3.2D), which are based on: (1) distance between ORFs; (2) conservation of synteny in other genomes; and (3) commonality of Gene Ontology or COG category. The relatively high correspondence between RNA

sequencing and bioinformatic predictions suggests that the operon structure in *S. elongatus* may be used to infer the operon structure in other cyanobacterial genomes. The median operon size is 1,320 nucleotides (Figure 3.2E), approximately twice the median size of an ORF (776 nucleotides) in *S. elongatus.*

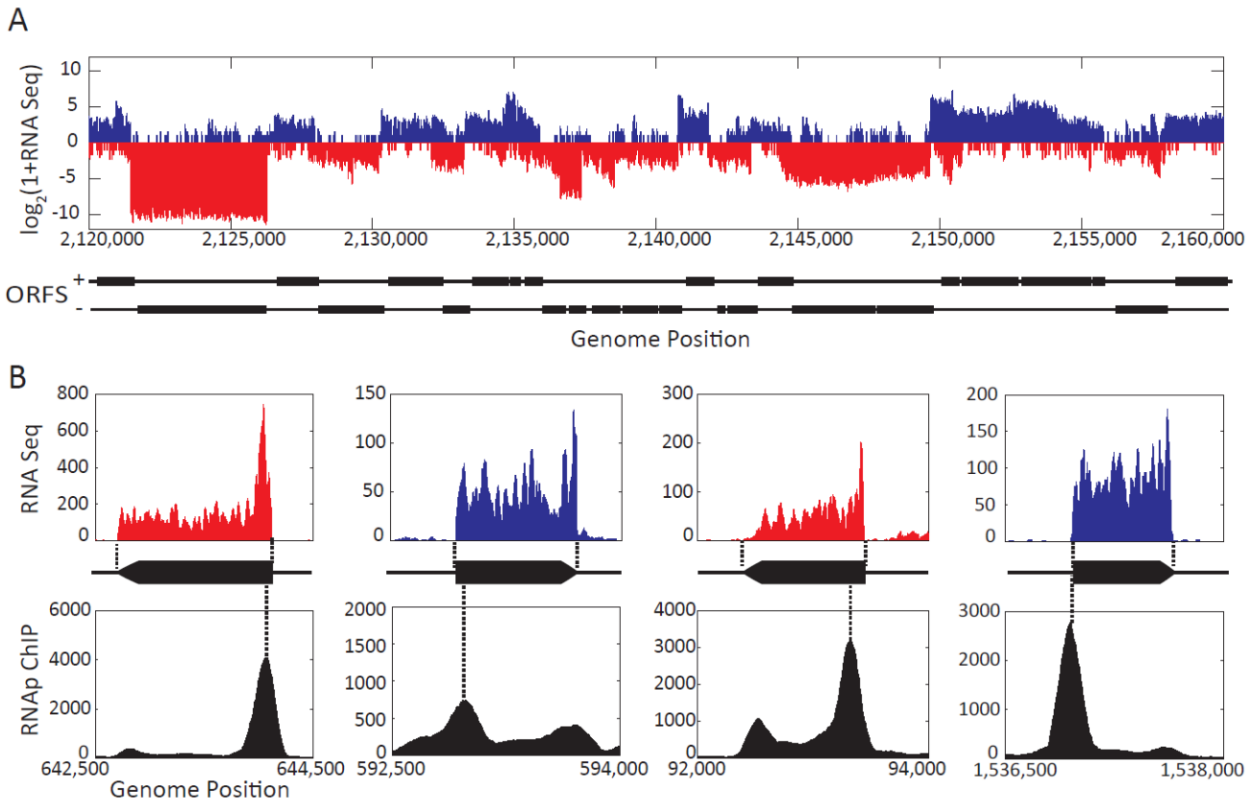


Figure 3.1: RNA sequencing and RNA pol ChIP in *S. elongatus*.

(A) Strand-specific RNA sequencing over a representative 40-kb region in the *S. elongatus* chromosome. Positive strand transcription is shown in blue (positive y-axis), and negative strand transcription in red (negative y-axis). For visualization over full dynamic range, the y-axis shows $\log_2$ transformed reads per nucleotide of RNA sequencing coverage. The position of Joint Genome Institute predicted ORFs for each strand are shown below in black. High RNA sequencing signal is present at nearly all ORFs and anti-sense transcription is extensive.

(B) RNA sequencing and RNA pol ChIP sequencing for representative highly expressed transcripts. Top panel: zoomed in view of RNA sequencing coverage of particular mRNA transcripts. Transcripts are color coded by strand as in (a). Transcription units with precise 5′ and 3′ ends are defined from RNA sequencing data for all mRNAs (black arrow) (Figure S3.1, Materials and methods). Bottom panel: RNA pol ChIP sequencing associated with the transcripts from the top panel. The y-axis is normalized such that the genome average is 200 units per nucleotide. Peaks in RNA pol occupancy are often found near the 5′ end of the transcript and occasionally smaller peaks in RNA pol occupancy are located near the 3′ ends or inside the transcript. 5′ peaks tend to be located within the transcript as opposed to within the promoters.

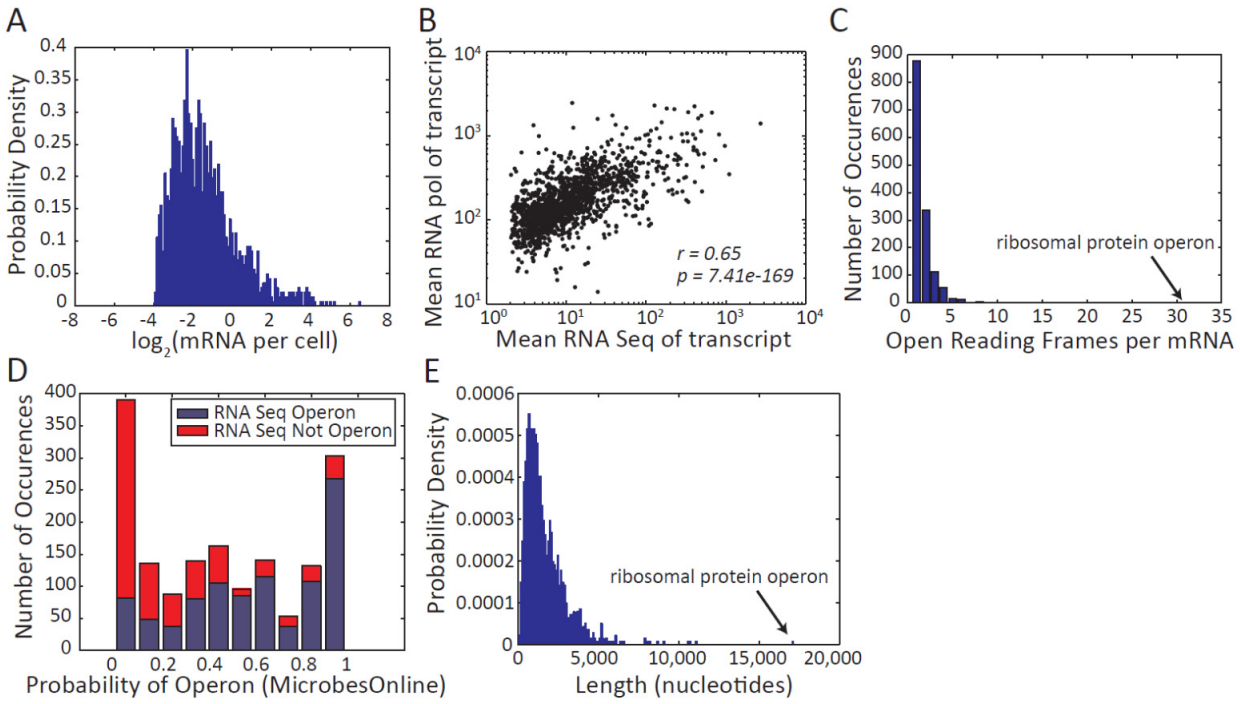


Figure 3.2: Basic features of the *S. elongatus* transcriptome.

(A) Distribution of absolute transcript abundance per cell. Only transcripts with mean coverage of over two reads per nucleotide (corresponding to approximately 1 mRNA per 15 cells) are shown, and a total of 1,500 mRNA per cell is assumed (Ingraham et al., 1983; Taniguchi et al., 2010) (Materials and methods).

(B) RNA sequencing versus RNA pol ChIP. Absolute transcription (RNA sequencing averaged over transcript) and absolute RNA pol occupancy (RNA pol ChIP averaged over transcript) are generally correlated (Pearson correlation, r = 0.65). The probability of getting a correlation as large by random chance (*P*-value) is 7.41e-169.

(C) Distribution of ORFs per mRNA. Most mRNAs contain one to two ORFs. The extreme case is that of an operon composed primarily of ribosomal proteins that includes 31 ORFs and is 17,158 nucleotides in length.

(D) Operon estimations based on RNA sequencing versus bioinformatic predictions. Comparison of RNA sequencing based operon determination and bioinformatic predictions from MicrobesOnline (Dehal et al., 2010; Price et al., 2005)

(E) Distribution of mRNA lengths. The median mRNA length is 1,320 nucleotides, approximately twice the median size of an ORF (776 nucleotides) in S*. elongatus*.

Transcription start

Identification of the 5′ ends of all mRNAs allows for more detailed characterization of the promoter and initial steps in transcription. When we align all mRNAs by their 5′ transcription start and average their AT content, we observe an increase at the -10 element, also known as the Pribnow box (Pribnow, 1975; Schaller et al., 1975) (Figure 3.3A). At this same location we observe a large drop in DNA melting temperature, a signature of bacterial promoters (Figure S3.4A). Downstream of the -10 element, we detect a peak in AT content at the first nucleotide of the transcript, indicative of a preference for incorporating adenine (Figure S3.4BC). We computed the sequence alignment of the 30 nucleotides prior to the transcription start and find a -10 element similar to that found in a genome-wide map of transcription start sites in *Synechocystis* PCC 6803 and 25 experimentally determined promoters in *Prochlorococcus* MED4 (Vogel et al., 2003; Mitschke et al., 2011; Harley et al., 1987) (Figure 3.3B, Materials and methods). Sequence alignment or motif analysis at the expected location of the -35 element or spacer does not reveal a strong consensus or motif. The absence of a strong -35 element signature has been observed in *Prochlorococcus* MED4 and in the *psbA* transcripts of many cyanobacteria (Vogel et al., 2003; Shibato et al., 1998), suggesting that the -35 elements in cyanobacteria may be very diverse in sequence. This diversity in -35 element may be related to the extensive control of gene expression by sigma factors in cyanobacteria (Imamura et al., 2009).

To investigate the presence of RNA pol peaks near the transcription start site, we aligned the top 500 expressed transcripts by their 5′ ends, and averaged the normalized RNA pol occupancy profiles (Figure 3.3C). On average, the maximum of the RNA pol

peak is situated 63 nucleotides downstream of the transcription start site. The exact peak position varies from transcript to transcript; this peak can be located either in the 5′ UTR or within the first ORF, with the majority of peaks occurring at the beginning of the ORF. This is in stark contrast to previous bacterial RNA pol ChIP-chip studies in which the RNA pol peaks are observed at the promoter (Grainger et al., 2005; Wade et al., 2008; Wade et al., 2004; Reppas et al., 2006), possibly due to a lack of resolution. A more recent high-resolution RNA pol ChIP-chip study in *Escherichia coli* was able to localize these RNA pol peaks to within the transcript (Mooney et al., 2009).

To assess a potential functional role for these RNA pol peaks, we calculated the RNA pol retention as the ratio of RNA pol occupancy at the 5′ end to the RNA pol occupancy in the middle of the first ORF for all mRNAs. We find that over 80% of transcripts have a retention ratio greater than one and that this retention ratio is variable from transcript to transcript (Figure 3.3D), allowing for the possibility that bacteria can tune the amount of retained RNA pol to affect gene expression.

One possible explanation for these RNA pol peaks is RNA pol pausing due to RNA secondary structure in nascent transcribed RNA, which may cause the RNA pol to pause or pause and subsequently terminate (Yanofsky, 1981). To determine if RNA secondary structure may be involved in pausing RNA pol, we selected a subset of 183 RNA pol peaks that were located 100 to 300 nucleotides within the transcript and were closer to a 5′ end than a 3′ end. This subset was chosen to specifically isolate the RNA pol peaks from features at the promoter or terminus, which may bias the analysis. The minimum free energy of 60-nucleotide RNA fragments from the transcribed strand, in 10-nucleotide increments, was calculated around each RNA pol peak and averaged,

revealing a steep drop in the minimum free energy slightly prior to the RNA pol peak (Figure S3.5A, Materials and methods). However, this decrease in free energy is still observed in dinucleotide shuffled sequences, suggesting that a specific stem loop structure is not formed in this region. Instead, we observe a shift in sequence bias from low to high GC content at the RNA pol peak (Figure S3.5B), which may be influencing the RNA minimum free energy calculation. Thus, the mechanism underlying the global accumulation of RNA pol at the 5′ end of transcripts remains unclear.

According to this hypothesis of 5′ proximal RNA pol pausing, we should also observe enrichment of RNA sequencing reads at the 5′ end of transcripts. Indeed, over 80% of transcripts have more RNA sequencing reads recovered at their 5′ ends than in the middle of their first ORF (Figure S3.6A), and a small but significant correlation exists between enrichment in RNA sequencing at 5′ ends and RNA pol retention ratio (Figure S3.6B).

Our genome-wide observations of 5′ RNA pol peaks suggest that this may be a more important and widespread phenomenon in bacterial gene expression than previously appreciated. Our observations of RNA pol pausing may be different from the canonical examples of transcriptional attenuation observed in amino acid biosynthetic operons of *E. coli* where specific terminator structures attenuate transcription [Yanofsky, 1981), although the peaks in RNA pol we observe are qualitatively similar to the peaks at the *trp* and *pyrBI* operons observed by tiling microarray in *E. coli* (Mooney et al., 2009).

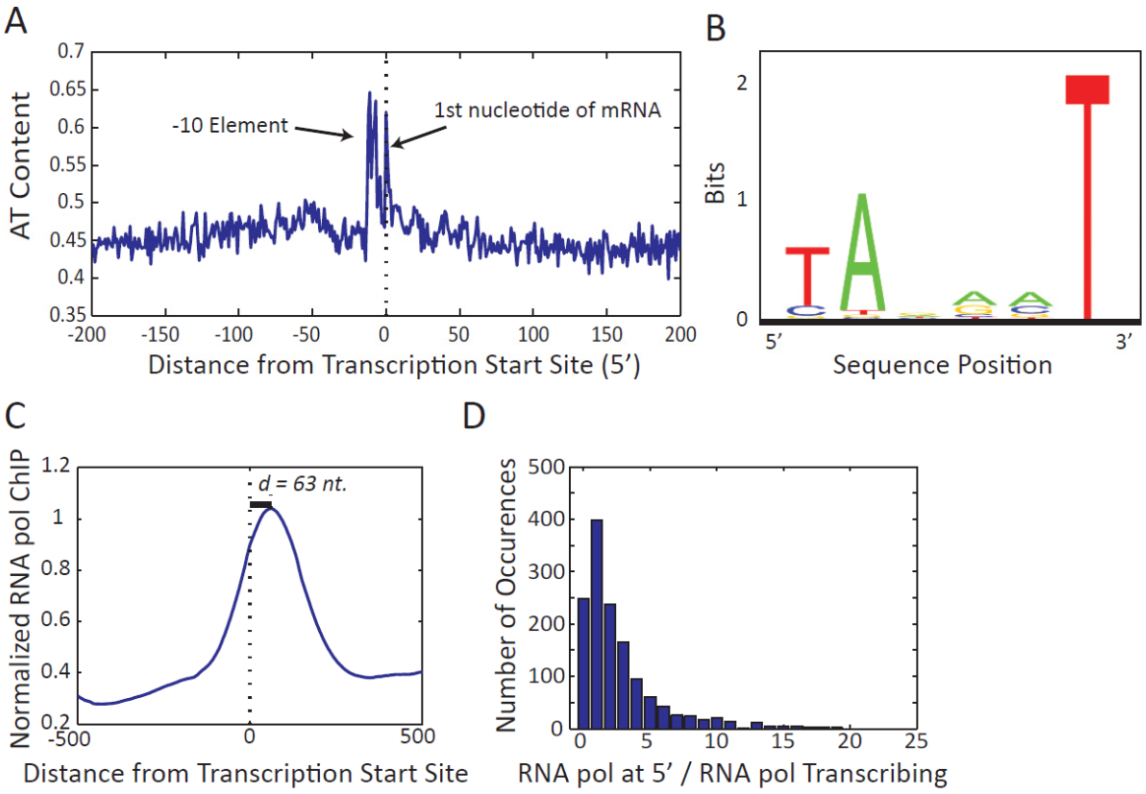


Figure 3.3: Transcription initiation in *S. elongatus*.

(A) AT content of the transcription start. The AT content from -200 to +200 from the start site of transcription was averaged for all mRNAs. A strong enrichment in AT content is observed at the -10 element as well as a strong preference for adenine at the first nucleotide of a transcript.

(B) -10 element consensus logo. A consensus -10 element similar in sequence to that determined for *E. coli* was identified through sequence alignment (Materials and methods).

(C) Normalized RNA pol occupancy at promoter. For each of the top 500 expressed mRNAs, the RNA pol occupancy was normalized to a mean occupancy of 0.5 per nucleotide, and then averaged across mRNAs from -500 to +500. A peak in RNA pol occupancy is observed, on average, 63 nucleotides within the RNA transcript, suggesting potential stalling of RNA pol after initiation of transcription rather than at the promoter.

(D) RNA pol retention ratio at the promoter is variable. The relative amount of RNA pol at the 5′ end versus RNA pol in the ORF varies from transcript to transcript. RNA pol at the 5′ end was calculated as the mean occupancy in a 200-nucleotide window centered at the +63 nucleotide. RNA pol transcribing was calculated as the mean occupancy in a 200-nucleotide window centered in the middle of the first ORF.

Transcription termination

In addition to analysis of the transcription start, our catalog of 3′ ends allows analysis of transcription termination. Two signals for transcription termination have been previously identified in bacteria: intrinsic Rho-independent terminators, typically low energy RNA hairpins; and Rho-dependent terminators, whose activity relies on the binding of the Rho protein to particular sites on the nascent transcript (Hoon et al., 2005). The majority of bacteria have a homolog of the *E. coli* Rho protein, but notable exceptions include the cyanobacteria *S. elongatus* and *Synechocystis* PCC 6803 (Hoon et al., 2005). A previous study analyzing the 3′ ends of ORFs in *Synechocystis* PCC 6803 found no noticeable drop in RNA minimum free energy, suggesting the potential for a previously uncharacterized mechanism for transcription termination in this organism (Washio et al., 1998). With knowledge of the actual 3′ positions of transcripts, a more accurate analysis of transcription termination in *S. elongatus* is possible.

To analyze the secondary structure at the 3′ end of transcripts, we averaged the minimum free energy of all transcripts aligned by the 3′ end (Figure 3.4A). We observe a dip in minimum free energy slightly prior to the transcript terminus, indicative of a stem-loop structure involved in Rho-independent transcription termination. This dip in free energy is not present in dinucleotide shuffled sequences, suggesting that a discrete stem-loop structure exists at the end of transcripts (Materials and methods, Figure S3.5C). To further assess the role of Rho-independent transcription termination in *S. elongatus*, we assembled all Rho-independent intrinsic terminators predicted in *S. elongatus* from TransTermHP (Kingsford et al., 2007). These predicted Rho-independent intrinsic terminators typically consist of short, often GC-rich hairpins

followed by sequence enriched in thymine nucleotides. We find these terminators tend to be significantly closer to 3′ ends than to random locations distributed at the same frequency (Figure 3.4B). Together, these analyses suggest that the classical Rho-independent termination plays a large role in cyanobacterial transcription termination.

Not all of the predicted intrinsic terminators cause complete transcription termination. The hairpin energy score (as calculated by TransTermHP (Kingsford et al., 2007) of those terminators that are within 100 nucleotides of a transcription terminus tend to be lower (more negative) than those that are located elsewhere (Figure 3.4C). These more stable hairpins may be more competent to cause transcription termination because they are either more likely to fold and/or more likely to cause termination after folding (Larson et al., 2008). In some cases, terminators that do not cause complete termination are involved in creating complex transcription structures. In several of these cases, terminators are found in between ORFs in the same operon, leading to lower transcription of the ORFs proximal to the 3′ end (Figure 3.4D). This strategy could potentially be used to regulate the stoichiometry of transcript abundance of ORFs, and subsequently proteins, regardless of the state of the promoter. A potential physiological example is that of the phycocyanin operon where a terminator that causes incomplete termination sets the stoichiometry of mRNA for *cpcβ* and *cpcα* to phycobilisome rod linkers at 6:1 - the same stoichiometry as in the organized phycobilosome (Belknap et al., 1987) (Figure S3.7).

Figure 3.4: Transcription termination in *S. elongatus*.

(A) Minimum RNA free energy at the end of transcripts. The minimum free energy of 60-nucleotide RNA fragments with 10-nucleotide spacing was calculated and averaged for all mRNAs (Materials and methods). A drop in minimum free energy at the 3′ end is indicative of Rho-independent transcription termination.

(B) Distance between TransTermHP bioinformatically predicted terminators and 3′ ends. Predicted intrinsic terminators (from TransTermHP (Kingsford et al., 2007)) tend to be much closer to the 3′ end of transcripts than to random positions occurring at the same frequency as 3′ ends. Blue bars show distance from a predicted terminator to the closest 3′ end. As a control, we randomized the location of 3′ ends in the genome. Grey bars show distance from a predicted terminator to the closest randomized 3′ end.

(C) Energy distributions of TransTermHP terminators. Not all predicted TransTermHP terminators cause transcription termination. Several terminator-like structures are located in non-transcribed regions or in the middle of transcripts. The free energy of terminators that cause transcription termination tends to be lower than the free energy of those that do not. The *P*-value is 3.00e-20 by two-sided Wilcoxon rank sum test.

(D) Partial transcription termination creates complex transcriptional structures. Positive strand transcription is shown in blue and negative strand transcription in red. The positions of predicted terminators (from TransTermHP) are shown in green, and the position of JGI predicted ORFs are shown in black. Terminators located within transcripts often result in a decrease in the transcription of downstream ORFs.

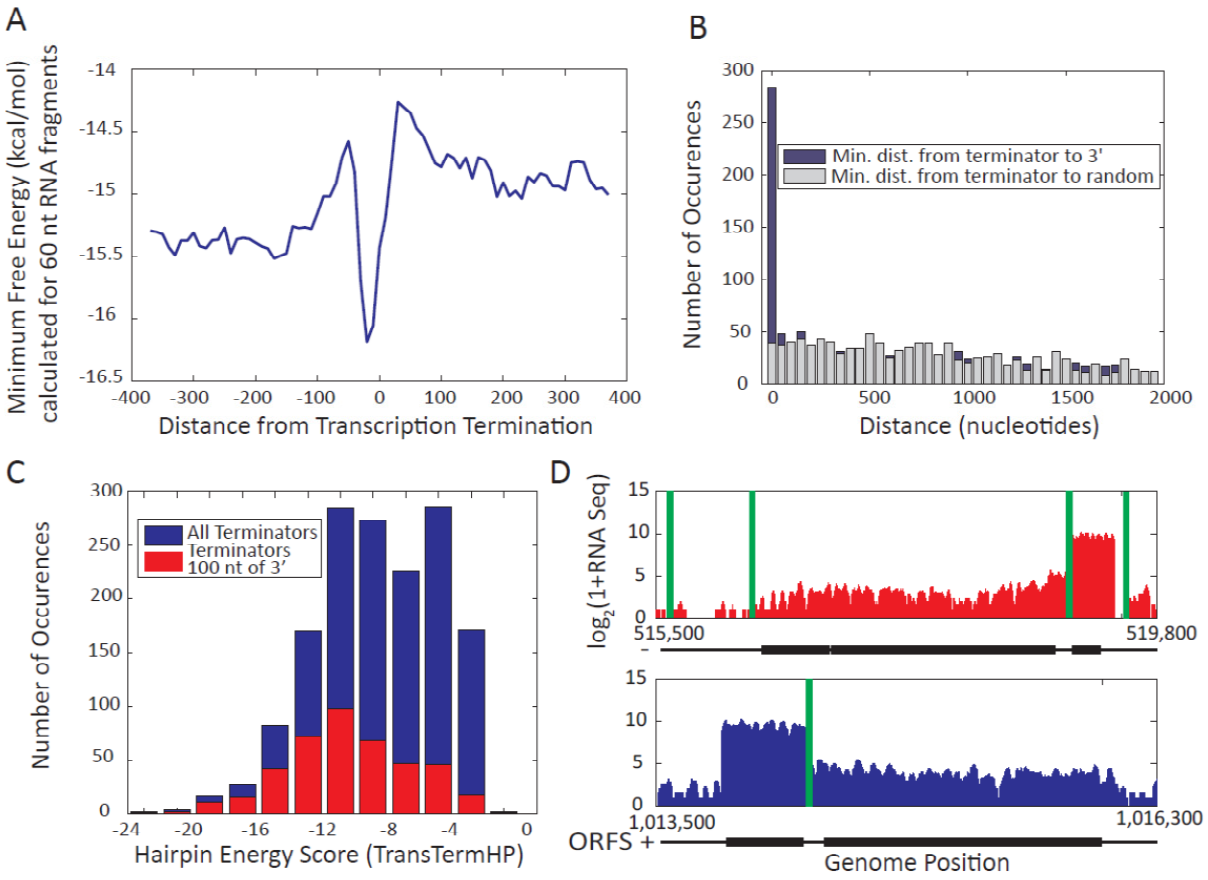


Figure 3.4 Continued

Putative non-coding transcripts and 5' UTRs

One particularly interesting feature of the *S. elongatus* transcriptome is the presence of widespread non-coding transcription. We identify 1,579 putative non-coding transcripts from RNA sequencing, 983 of which are considered high-confidence after verification by tiling microarray, and annotate their 5′ and 3′ ends (Materials and methods). The number of non-coding transcripts is comparable to the number of annotated protein-coding transcripts (1,415). It is possible that some of the transcripts designated as non-coding may have a protein coding region that was not identified in the JGI annotation. Those putative non-coding transcripts that have any overlap with annotated transcripts on the opposite strand were considered anti-sense and the remaining were considered not anti-sense.

Several hundred non-coding RNAs have previously been identified in *E. coli* and *Bacillus subtilis* (Altuvia, 2007) and recently 276 novel transcriptional units were identified in *Prochlorococcus* MED4 by tiling microarray (Georg et al., 2009), 117 in *Mycoplasma pneumonia* by tiling microarray and transcriptome sequencing (Guell et al., 2009), 390 in *Sulfolobus solfataricus* P2 by transcriptome sequencing (Wurtzel et al., 2010), and 137 in *Salmonella* Typhi by transcriptome sequencing (Perkins et al., 2009). As RNA from these and other genomes are sequenced at further depth, we may find that non-coding transcription is more prevalent in bacteria than previously thought (Sharma et al., 2010; Passalacqua et al., 2009; Co et al., 2009). A recent RNA sequencing-based map of transcription start sites in another unicellular cyanobacterium, *Synechocystis* PCC 6803, identified 1,541 potential non-coding transcription start sites, making up 64% of all transcription start sites in the organism (Mitschke et al., 2011).

We find that some of the non-coding transcripts in *S. elongatus* display differential expression in the subjective dawn and subjective dusk timepoints, indicative of circadian expression, as assayed by tiling microarray (Figure S3.8, Materials and methods). Although several non-coding RNAs appear to exhibit circadian oscillations in expression, the physiological role for circadian gene expression remains unclear and no expression correlation exists between anti-sense circadian non-coding RNAs and the transcripts on the opposite strand.

Very few well-described examples of non-coding RNAs have been noted in cyanobacteria. One previously identified functional non-coding RNA, Yfr1, is required for growth under several stress conditions (Nakamura et al., 2007) (Figure 3.5A). In *S. elongatus*, there appears to be occasional co-transcription of Yfr1 with the neighboring ORF *guaB*, but the extent of co-transcription is negligible compared to the expression of Yfr1. In the same genomic region as Yfr1, we observe several previously unidentified transcripts anti-sense to the *trxA* and *guaB* coding regions. The Yfr1 non-coding transcript is approximately 60 nucleotides in length, and the median size of all identified non-coding transcripts is approximately 200 nucleotides, roughly 15% of the size of mRNA transcripts (Figure 3.5B).

We find that most non-coding transcripts are at least partially anti-sense to an mRNA transcript (Figure 3.5B). These transcripts have the potential for base pairing with the transcript on the opposite strand. One such functional RNA, IsrR, has been identified in the cyanobacterium *Synechocystis* PCC 6803 (Duhring et al., 2006). This 177-nucleotide RNA is down-regulated in iron stress and base-pairs with the iron stress-induced *isiA* transcript, subsequently decreasing its levels. IsiA enhances

photosynthesis by forming a ring around photosystem I, and IsrR is currently the only RNA known to regulate a photosynthesis component (Duhring et al., 2006). We find a transcript anti-sense to *isiA* in *S. elongatus* that shows significant similarity to IsrR in *Synechocystis* PCC 6803 (RNA Families (RFAM) bit score 97.96) (Griffith-Jones et al., 2003). This transcript may have a similar role in modulating photosynthesis in *S. elongatus*.

To identify if any other known RNA families are present within our set of non-coding RNAs, we queried the RFAM database (Griffith-Jones et al., 2003). In addition to Yfr1, IsrR, and RNase P, we identify a non-coding RNA containing a putative group I intron (Nielsen et al., 2009). Group I introns are ribozymes capable of catalyzing their own excision from an RNA, and ligating the upstream and downstream exons.

To extend our analysis of potential RNA-based regulators in *S. elongatus*, we queried our set of 5′ UTRs against RFAM and identified metabolite-binding riboswitches for thiamine (vitamin B_1) and coenzyme B_{12} (vitamin B_{12}). The 5′ leader of the *thiC* mRNA in *S. elongatus* contains a ‘thi box’ riboswitch domain that undergoes a structural change that has been shown to cause both a reduction in translation and transcription when bound to thiamine or its pyrophosphate derivative (Winkler et al., 2002). Similarly, the 5′ leader of a putative cobalt transporter (JGI 637799805, *Synpcc7942_1373*) contains the cobalamin riboswitch domain, which represses expression in the presence of coenzyme B_{12} (Nahvi et al., 2002). Both of these mRNA transcripts have unusually large 5′ UTRs of 210 and 153 nucleotides, respectively, compared to a median 5′ UTR size in *S. elongatus* of 30 nucleotides. Although most 5′ UTRs are small, 12% are longer than 100 nucleotides and 6% are longer than 150 nucleotides. Transcripts with long 5′

UTRs may be good candidates for riboswitches or RNA-based regulators. Interestingly, both riboswitch-containing mRNAs show large RNA pol occupancy peaks near the riboswitch domain in the 5′ UTR, suggesting that these riboswitches - likely when in their bound configuration - can cause RNA pol pausing or termination. These peaks in RNA pol are qualitatively similar to the peaks we observe globally, although mechanisms likely differ, as most RNA pol peaks are situated within the beginning of the ORF.

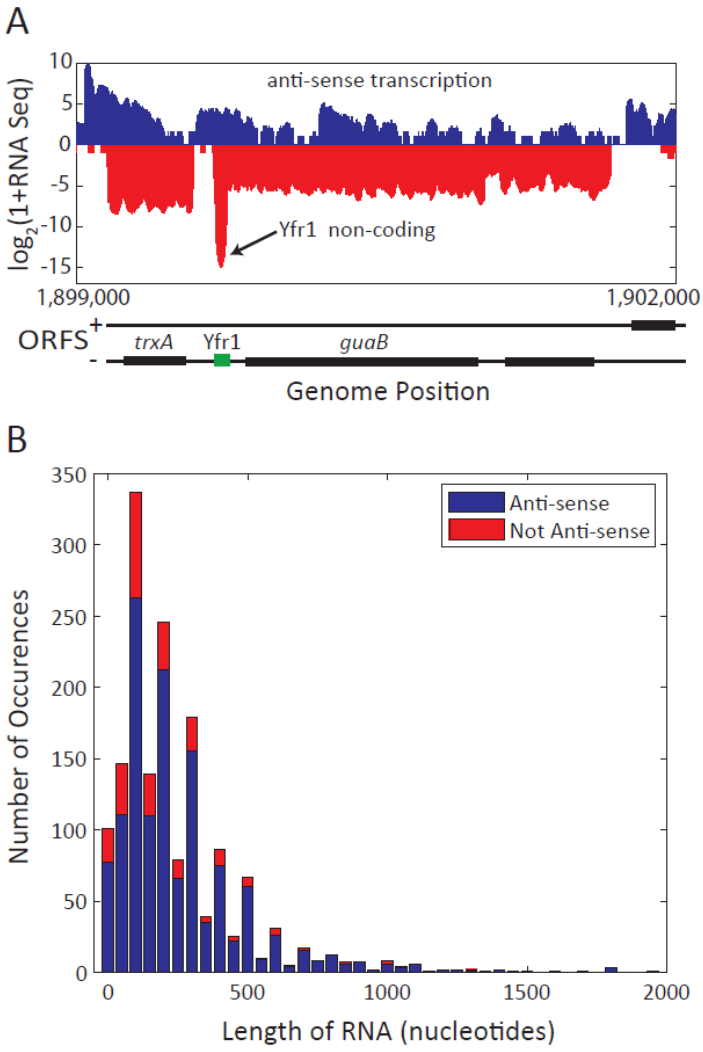


Figure 3.5: Non-coding transcripts.

(A) Extensive non-coding transcription. Positive strand transcription is shown in blue (positive y-axis), and negative strand transcription in red (negative y-axis). The position of JGI predicted ORFs on the plus and minus strand are shown in black, and the position of the Yfr1 non-coding RNA is shown in green. In this same region, there is anti-sense transcription opposite to the *trxA* and *guaB* ORFs.

(B) Length distribution of non-coding transcripts. Transcripts that have any overlap with an annotated transcript on the opposite strand are designated anti-sense. Most non-coding transcripts are anti-sense by this designation. The median size for a non-coding transcript is approximately 200 nucleotides.

Conclusions

Here we combine three high-resolution data sets - RNA sequencing, tiling expression microarray, and RNA pol ChIP sequencing - to present a characterization and analysis of the *S. elongatus* transcriptome. We report absolute transcript levels, operon identification, and high-resolution mapping of 5′ and 3′ transcript ends. At the 5′ end of transcripts, we characterize promoter sequence and find widespread peaks in RNA pol occupancy. At 3′ ends we observe significant Rho-independent transcription termination and occasional incomplete termination resulting in interesting transcriptional structures. In addition, we find extensive non-coding transcription, suggesting a larger role for these non-coding RNAs in bacteria, and cyanobacteria in particular, than previously anticipated. The presence of numerous non-coding RNAs and 5′ proximal pausing of RNA pol suggest that post-transcriptional regulation - regulation after binding of RNA pol at the promoter - may be more widespread in bacteria than expected. We hope this work will serve as a catalog and primer for further studies of bacterial and cyanobacterial transcription.

Materials and methods

Continuous culture of cyanobacteria

Cyanobacteria were cultured as previously described (Vijayan et al., 2009). A continuous culture apparatus kept cells in constant light and growth conditions and provided real-time bioluminescence readings. *S. elongatus* (stain AMC 408 (Min et al., 2004)): *psbAI::luxCDE* fusion in NS1 (Andersson et al., 2000) (spectinomycin and streptomycin) and *purF::luxAB* fusion in NSII (Andersson et al., 2000) (chloramphenicol)

was grown in a 6-L cylindrical spinner flask (Corning, Corning, NY, USA) at a volume of 4.5 L. Cells were grown in BG-11 medium [0] with the following modifications: 0.0010 g/L $FeNH_4$ citrate was used instead of 0.0012 g/L $FeNH_4$ citrate and citric acid was supplemented at 0.00066 g/L. Cells were initially inoculated in the presence of antibiotics (5 µg/ml spectinomycin and 5 µg/ml chloramphenicol), and subsequently diluted with modified BG-11 lacking antibiotics. Cells were exposed to surface flux of approximately 25 µmol photons m^{-2} s^{-1} cool white florescent light, bubbled with 500 ml/minute 1% CO_2 in air, maintained at 30°C, and stirred at one rotation per second. Constant optical density (OD_{750} 0.15) and volume are achieved via a two state controller. OD does not fluctuate greater than 8% during an experiment. Cells are exposed to two 12-hour light-dark cycles for entrainment before release into continuous light.

RNA preparation

Total RNA was prepared as previously described (Vijayan et al., 2009). Cells (120 ml) from continuous culture were collected by vacuum filtration, snap frozen in liquid nitrogen, and stored at -80°C for no more than 1 week prior to RNA extraction. RNA was extracted from frozen cells in two steps. First, cells were lysed in 65°C phenol/SDS by vortexing and total RNA was purified by phenol/chloroform extraction. Second, total RNA was subjected to DNase I (Promega, Madison, WI, USA) treatment followed by a second phenol/chloroform extraction. Total RNA was analyzed on agarose gel and an Agilent Bioanalyzer to assess integrity.

Strand-specific RNA sequencing

Total RNA was prepared for timepoints collected at 4-hour intervals from 76 to 96 hours after release into continuous light and mixed in equal proportions. Mixed total RNA was supplemented with RNase Out (Invitrogen, Carlsbad, CA, USA) to a final concentration of 2 units/µl and depleted of 23S and 16S ribosomal subunits using the MICROBExpress Bacterial mRNA Enrichment Kit (Ambion, Austin, TX, USA) according to manufacturer's instructions.

RNA sequencing libraries were prepared from total RNA depleted of 16S and 23S rRNA with modifications to a previously described procedure (Ingolia et al., 2009). RNA (8 µg) was fragmented for 40 minutes at 95°C in fresh 2 mM EDTA, 100 mM $NaCO_2$, pH 9.2. Fragmentation reactions were immediately precipitated in 300 mM NaOAc, pH 5.2, glycogen, and isopropanol. Fragmented RNA was resuspended in RNA loading buffer (Fisher, Pittsburg, PA, USA), briefly denatured, and loaded in a 15% TBE-Urea polyacrylamide gel (BioRad, Hercules, CA, USA) for size selection. Gels were stained with Sybr Gold (Invitrogen) and a 25- to 30-nucleotide band was excised using a synthesized 28-nucleotide RNA and denatured 10-bp DNA ladder (Invitrogen) as standards. The gel slice was physically disrupted and RNA was recovered in 300 mM NaOAc, 1 mM EDTA, 0.1 units/µl SUPERase·In (Ambion) overnight at room temperature. Solution was transferred to a Spin-X cellulose acetate filter (Corning) to remove gel debris and precipitated with glycogen and isopropanol. Size selected fragmented RNA was denatured briefly and dephosphorylated in a 30 µl reaction with 1× T4 polynucleotide kinase buffer without ATP (NEB, Ipswich, MA, USA), 20 units SUPERase·In, and 15 units T4 polynucleotide kinase (NEB) at 37°C for 1 hour. The

reaction was precipitated, resuspended, briefly denatured, and poly-(A) tailed in a 25 µl reaction with 1× poly-(A) polymerase buffer (NEB), 5 units SUPERase·In, 1 mM ATP, and 1.25 units *E. coli* poly-(A) polymerase (NEB) at 37°C for 10 minutes. Reactions were quenched with 80 µl of 5 mM EDTA and precipitated.

Reverse transcription was carried out from the introduced poly-(A) tail anchor of denatured RNA using primer oNTI255 (Ingolia et al., 2009) with the SuperScript III reverse-transcriptase system (Invitrogen) supplemented with 2 units/µl of SUPERase·In at 48°C for 30 minutes. RNA was subsequently hydrolyzed in 0.1 M NaOH at 98°C for 15 minutes and loaded in a 10% TBE-Urea polyacrylamide gel (BioRad) and the extended first-strand product was excised and recovered as above in 300 mM NaCl, 10 mM Tris, pH 7.9, 1 mM EDTA. First-strand cDNA was circularized in a 20 µl reaction with 1× CircLigase buffer (Epientre, Madison, WI, USA), 50 µM ATP, 2.5 mM $MnCl_2$, and 1 µl CircLigase (Epientre) for 1 hour at 60°C, and then heat-inactivated for 10 minutes at 80°C.

Circularized cDNA template (1 µl) was amplified using Phusion Hot Start High-Fidelity enzyme (NEB) and primers oNTI230 and oNTI231 (Ingolia et al., 2009) to create DNA with Illumina cluster generation sequences on each end along with the Illumina small RNA sequencing primer binding site. PCR was carried out with an initial 30 second denaturation at 98°C, followed by 8 cycles of 10 second denaturation at 98°C, 10 second annealing at 60°C, and 5 second extension at 72°C. PCR product was loaded in a non-denaturing 10% TBE polyacrylamide gel (BioRad) and a 113- to 125-nucleotide band was excised using a 10-bp ladder as standard. DNA was recovered as previously described. Libraries were quantified using an Agilent Bioanalyzer and 4 to 6

pM of template was used for cluster generation and sequenced on Illumina Genome analyzer II with the Illumina small RNA sequencing primer. Sequence tags were stripped of the terminal poly-(A) sequence and aligned to the *S. elongatus* genome with Bowtie (Langmead et al., 2009). Stripping of terminal poly-(A) sequence at the end of each read will remove the introduced poly-(A) tail but will also remove any trailing adenines at the 3′ end of the reverse-transcribed RNA fragment, biasing the 3′ end determination of RNAs that end in trailing adenines. GenBank CP000100, CP000101, and S89470 were used to align reads to the chromosome and endogenous plasmids. Uniquely mappable reads with a maximum of three mismatches were mapped to the genome and extended by the length of the individual read.

A total of 22,375,035 uniquely mappable reads were mapped to the genome with approximately 624 million bases of sequences covering each nucleotide strand-specifically an average of approximately 115 times. These uniquely mappable reads exclude any reads from rRNA since multiple copies of each rRNA exist in the genome. Technical replicates showed very high Pearson correlation coefficients ($r > 0.99$). RNA sequencing data are displayed and analyzed as coverage per nucleotide - defined as the number of times a given nucleotide position was observed in all the sequencing reads. Absolute transcript levels are assumed to be equal to the average coverage per nucleotide across the length of the transcript. All analysis was performed on the chromosome, although raw data for both endogenous plasmids are available.

Strand-specific expression tiling microarray

Expression was measured using two separate custom designed two-color 244 k microarrays - one for the forward strand and another for the reverse strand (forward strand tiling array, Agilent Array ID 022715; reverse strand tiling array, Agilent Array ID 022716). Arrays were designed using eArray software (Agilent). Forward and reverse strand sequence is as defined by GenBank CP000100, CP000101, and S89470 - which define the chromosome and two plasmid sequences, respectively.

All tiling probes were 60 nucleotides in length with 12-nucleotide spacing between probe starts such that $probe_i$ and $probe_{i+1}$ overlapped by 48 nucleotides. A 6-nucleotide offset of the tile between strands allows for 6-nucleotide resolution of double stranded targets and 12-nucleotide resolution for strand-specific targets. In addition, each array included four temperature matched probes (80°C) against each JGI predicted ORF, *luxA* through *luxE*, and *Arabidopsis* spike-in controls (Ambion). These additional probes are identical to those in Agilent Array ID 020846, as previously described (Vijayan et al., 2009).

cDNA was prepared for each individual timepoint (foreground channel) as well as for a pool of all timepoints (background channel). The background channel consisted of a pool of samples collected at 4-hour intervals from 24 to 84 hours after release into continuous light. The foreground channel consisted of individual timepoints 60, 68, 72, and 80 hours after release into continuous light. The same samples were analyzed by non-tiling microarray in Vijayan et al 2009. Spike-in RNA was introduced at different concentrations and ratios to the foreground and background channels before reverse transcription to ensure proper ratio detection over a wide dynamic range. Total RNA (5 µg; plus spike-ins) was reverse-transcribed with random 15-mer primers (Operon,

Huntsville, AL, USA) and a 2:3 ratio of amino allyl-UTP:dTTP (Sigma, St. Louis, MO, USA) using the SuperScript III reverse-transcriptase system without amplification. RNA was hydrolyzed and cDNA was purified using Microcon 30 spin column (Millipore, Billerica, MA, USA).

First-strand cDNA was labeled with *N*-hydroxysuccinimide-ester cyanine 3 (Cy3, foreground) or cyanine 5 (Cy5, background) (GE Biosciences, Uppsala, Sweden) in 0.1 M sodium bicarbonate pH 9.0 for 6 hours. Labeled cDNA was purified (Microcon 30) in preparation for hybridization. Each array was hybridized with approximately 750 ng Cy3 and approximately 750 ng Cy5 labeled cDNA and rotated (five rotations per minute) at 60°C for 17 hours in SureHyb chambers (Agilent). Arrays were subsequently washed in 6.7× SSPE and 0.005% N-lauryl sarcosine buffer for at least 1 minute, 0.67× SSPE and 0.005% N-lauryl sarcosine buffer for 1 minute, and then Agilent drying and ozone protection wash for 30 seconds at room temperature (1× SSPE = 0.15 M NaCl, 10 μM sodium phosphate, 1 mM EDTA, pH 7.4). The arrays were immediately scanned using an Axon 4000B scanner at 5-μm resolution. The median intensity of the Cy3 and Cy5 florescence at each spot was extracted using GenePix software (Molecular Devices, Sunnyvale, CA, USA). For calculation of logarithmic ratios, Loess and quantile normalization were performed in succession using the MATLAB (MathWorks, Natick, MA, USA) bioinformatics toolbox.

ChIP sequencing of RNA polymerase

We crosslinked 250 ml of cells from continuous culture (OD_{750} 0.15) with 1% formaldehyde for 15 minutes and then quenched them with 125 mM glycine for 5

minutes at room temperature. Cells were collected by centrifugation and washed twice with cold phosphate-buffered saline buffer, pH 7.4. The cell pellet was snap frozen in liquid nitrogen and stored at -80°C. Samples were collected 32 to 52 hours after release into continuous light at 4-hour intervals. At the same time, samples were collected and processed for non-tiling microarray as described in (Vijayan et al., 2009).

ChIP was performed in a manner similar to that previously described (hecht et al., 1999; Lam et al., 2008). Cells were mechanically lysed by beating with 0.1 mm glass beads in cold lysis buffer A (50 mM HEPES, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-Deoxycholate) with protease inhibitors (Roche, Basel, Switzerland). Chromatin was fragmented by sonication of the lysate to a median of approximately 300 bp and the protein concentration of the supernatant was measured by BCA (bicinchoninic acid) (Thermo, Rockford, IL, USA) using bovine serum albumin as standard. Lysate (750 µg) was incubated with 30 µg of antibody - RNA polymerase β subunit antibody WP023 (Neoclone, Madison, WI, USA) or mouse whole IgG mock (Jackson ImmunoResearch, West Grove, PA, USA) - and incubated overnight at 4°C. We verified that the monoclonal RNA polymerase β subunit antibody WP023 reacts with *S. elongatus* RNA polymerase β by western blot analysis of whole cell extract, where it produces a single band of the expected size. Lysate was supplemented with Protein G Sepharose Fast-Flow beads (Invitrogen) and incubated for an additional 2 hours at 4°C. After incubation, sepharose beads were washed in cold buffer at room temperature: 2 × 5 minutes lysis buffer A; 1 × 5 minutes lysis buffer B (50 mM HEPES, pH 7.5, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate); 1 × 5 minutes wash buffer (10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% NP-40, 0.1% Na-

deoxycholate); 1 × 5 minutes TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA). Protein-DNA was eluted from beads by incubation of samples at 65°C for 1 hour in elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1.0% SDS). Crosslinks were reversed in supernatant by incubation of samples at 65°C overnight in elution buffer. Western blotting of supernatant of mock versus immunoprecipitation shows 45% efficiency pull-down of the β subunit and 25% co- immunoprecipitation of the β' subunit in the immunoprecipitation using Neoclone antibodies WP023 and WP001, respectively. Proteins were digested with 0.2 mg/ml proteinase K for 2 hours at 37°C. Nucleic acid was then purified with phenol/chloroform extraction and precipitated with ethanol and LiCl. Nucleic acid was re-suspended in TE buffer and RNA was digested in 20 µg/ml RNase and subsequently phenol/chloroform purified. For input control, 5% of the volume of cell lysate was removed after sonication and used to prepare the input DNA. The ChIP DNA concentration was estimated with the Pico-green DNA detection kit (Invitrogen).

ChIP sequencing libraries were prepared for samples zeitgeber time (ZT) 32 (subjective dusk) and ZT 44 (subjective dawn) as these timepoints showed maximal/minimal gene expression for canonical circadian mRNAs *kaiC* and *purF* by microarray. Mock ChIP sequencing libraries were prepared for an equal mix of lysate from ZT 32 through ZT 52 (collected at 4 hour intervals). A total of six sequencing libraries were prepared (Table 3.1).

Table 3.1: RNA polymerase ChIP samples

Sample	Total aligned reads
ZT 32 RNA pol ChIP	8,815,678
ZT 32 input	20,203,310
ZT 44 RNA pol ChIP	11,201,620
ZT 44 input	19,864,425
ZT 32 through 52 mock	10,595,684
ZT 32 through 52 input	16,712,868

ZT, zeitgeber time.

Sequencing libraries were prepared from 10 ng DNA following the Illumina ChIP protocol (revision A) and libraries sized between 200 and 300 bp were selected for amplification. Libraries were assayed with the Agilent Bioanalyzer and 8 pM of template was used for cluster generation. Libraries were sequenced using Illumina primers on an Illumina Genome analyzer II, and each sequence tag was aligned to the *S. elongatus* genome with Bowtie (Langmead et al., 2009). GenBank CP000100, CP000101, and S89470 were used to align reads to the chromosome and endogenous plasmids. Uniquely mappable reads with a maximum of three mismatches were mapped to the genome. Reads were then extended 150 bp to cover the average length of insert DNA between sequencing adaptors as determined by the Agilent Bioanalyzer.

A comparison of change in RNA pol ChIP versus change in gene expression (measured by non-tiling microarray) at timepoints ZT 32 and ZT 44 is shown in Figure S3.3. All other analysis was performed on the sum of the normalized libraries from ZT 32 and ZT 44, which was normalized to a mean coverage of 200 reads per nucleotide. Additional normalization by input does not change conclusions (Figure S3.2). A representative region of the genome is presented in Figure S3.2. All analysis was performed on the chromosome, although raw data for both endogenous plasmids is available.

Calculation of percent of genome transcribed

The percent of transcription along the *S. elongatus* chromosome was calculated by imposing a coverage cutoff for transcription of two reads per nucleotide. If a nucleotide is expressed at or over this cutoff it is regarded as transcribed.

This conservative cutoff indicates that only approximately 84.7% of the nucleotides within annotated JGI chromosomal ORFs are transcribed. Of the approximately 15.3% of nucleotides within annotated ORFs that do not pass this cutoff, approximately 41.4% are within an ORF that has an average number of reads per nucleotide of less than 2, which corresponds to approximately 1 RNA per 15 cells when we assume a total of 1,500 mRNAs per cell (245 of 2,665 chromosomal ORFs have <2 reads per nucleotide).

Using this cutoff we find 54.7% of each strand is transcribed and 88.0% of the chromosome is transcribed on either the plus or minus strand. That is, on any given strand and at any given chromosomal position, there is a 54.7% chance that the nucleotide is transcribed. Similarly, at any given chromosomal position, there is an 88% chance that the nucleotide is transcribed on either the plus or minus strand. Eighty-two percent of non-coding regions are transcribed on either the plus or minus strand.

Identification of 5′ and 3′ ends of Joint Genome Institute predicted ORFs and definition of operons

The 5′ and 3′ ends of all JGI predicted ORFs (and rRNA and tRNA) with an average coverage of at least two reads per nucleotide were identified using a probability-based approach using *a priori* knowledge of translation start and stop positions. Of 2,665 chromosomal ORFs (and rRNA and tRNA), 2,420 had an average number of reads per nucleotide of ≥2. For every predicted translation start, we searched for the first upstream nucleotide (*i* - 1 is upstream of *i*) on the same strand *i* that was not within a JGI predicted ORF and that satisfied one of the following three criteria: (1)

$binomial_{cdf}$ ($reads_{i-1}$, $reads_i + reads_{i-1}$, 0.5) < 0.01 and $reads_i/reads_{i-1} \geq 2$; (2) $binomial_{cdf}$ ($reads_{i-2}$, $reads_i + reads_{i-2}$, 0.5) < 0.01 and $reads_i/reads_{i-2} \geq 2$; and (3) $reads_{i-1} < 2$.

Where $binomial_{cdf}$ (k, n, p) is the probability of getting at least k success in n trials when p is the success probability of each trial. This *i* was designated the 5′ transcription start site. The distance of predicted 5′ ends to those published in previous studies is reported in Table S4 of the associated publication (Vijayan et al., 2011) and examples are shown in Figure S3.1. Similarly, for every predicted translation stop codon, we searched for the first downstream nucleotide *i* that was not within a JGI predicted ORF and that satisfied one of the same criteria. This *i* was designated the 3′ transcription end. 5′ Ends tend to be better defined than 3′ ends, possibly related to the biology of transcription termination. ORFs that shared the same 5′ transcription start site were defined as being on the same operon. We observed 43 cases of multiple transcription start sites - the presence of a 5′ transcription start within another transcript. All identified transcripts are reported in Table S1 of the associated publication (Vijayan et al., 2011). A total of 1,473 transcripts were identified. All analysis was performed on the subset of 1,415 transcripts defined as mRNA transcripts as they do not contain any tRNA or rRNA. Note, in some cases a tRNA was predicted to be on the same transcript as an ORF because the high expression of the tRNA obscures the transcription boundary.

Identification of non-coding transcripts

Non-coding transcripts were identified using a multi-tiered approach that first identifies transcribed regions and then estimates their 3′ and 5′ positions.

First, 15,000 nucleotide intervals of the chromosome (with overlap of 5,000 nucleotides) were optimally segmented into 30 segments of approximately constant signal, yielding a total of 8,070 segments per strand. Segmentation was performed in MATLAB to minimize the cost function:

$$\sum_{s=1}^{S=30} \sum_{i \geq t_s}^{i<t_{s+1}} (y_i - \bar{y}_s)^2 ,$$

where y_i is the $\log_2(1 + \text{reads}_i)$ at nucleotide *i*, $\bar{y}_s$ is the arithmetic mean of $\log_2(1 +$ reads) along segment *s*, and $t_1, \ldots, t_s$ are segment boundaries (Picard et al., 2005; David et al., 2006; Huber et al., 2006). This change-point approach more accurately discriminates transcribed and non-transcribed segments than the running window approach and requires only one user-defined parameter - the total number of transcribed segments - which we set at 1 per 500 nucleotides strand-specifically.

Next, all segments that correspond to non-transcribed regions - mean coverage less than two reads per nucleotide - were removed. Segments that overlapped with an annotated transcript (see previous section) were removed and the remaining segments were consolidated. The exact 5′ and 3′ end of each segment was determined using the same algorithm described in the previous section except 5′ and 3′ ends were not allowed to overlap with an annotated operon. A total of 1,579 non-coding transcripts were detected using this method. All non-coding transcripts are reported in Table S2 of the associated publication (Vijayan et al., 2011).

Identification of high-confidence non-coding transcripts

Tiling microarray ratios were utilized to identify a set of high-confidence non-coding transcripts. We took advantage of the fact that transcripts have high Pearson cross-correlation among internal probes (probes that are fully internal to the transcript) across all circadian timepoints (McGrath et al., 2007). That is, when the ratio of one probe changes at a particular circadian time, the ratio of the other probes within the transcript is similarly affected. First, we assembled the distribution of mean cross-correlation values among internal probes for all predicted JGI ORFs. This formed the expected cumulative distribution for mean cross-correlation of transcribed regions. All non-coding transcripts whose mean cross-correlation was above the 5% cutoff of the expected distribution were considered high-confidence. This assumes that all non-coding transcripts with mean cross-correlation larger than the bottom 5% of ORFs are high-confidence. Table S2 in the associated publication (Vijayan et al., 2011) indicates whether a non-coding transcript was designated as high-confidence. Of the 1,579 non-coding transcripts, 157 could not be assayed because they were smaller than the probe width of 60 nucleotides. Of the remaining 1,422 non-coding transcripts, 983 (approximately 70%) passed this cutoff.

Identification of high-confidence circadian non-coding transcripts

Circadian transcripts corresponding to annotated JGI ORFs have been previously described (Vijayan et al., 2009; Ito et al., 2009). To identify potential non-coding circadian transcripts, we first calculated the relative gene expression of each non-coding transcript at each timepoint by taking the arithmetic mean of gene expression ratios across all microarray probes internal to the transcript. This gives us the relative expression of each non-coding transcript at each timepoint relative to the

background. Then we calculated the gene expression ratio between the two most extreme (in gene expression) circadian timepoints (circadian time (CT) 12 (subjective dusk) and CT 20 (subjective dawn), corresponding to ZT 60 and ZT 72, respectively) (Vijayan et al., 2009). Large negative ratios are indicative of dawn-peaking transcripts and large positive ratios are indicative of dusk-peaking transcripts. To assign a designation of circadian behavior to each non-coding transcript, we calculated the same ratios for all annotated ORFs - where the circadian behavior is already known from (Vijayan et al., 2009). We found the ratio for annotated ORFs at which a cumulative 10% false positive rate existed for dawn or dusk genes, and used these cutoffs to identify potential circadian non-coding transcripts. Expression ratios and indication of potential circadian behavior are shown in Table S2 in the associated publication (Vijayan et al., 2011). The timecourse expression of all high-confidence circadian non-coding RNAs is shown in Figure S3.8. Although only 106 of 1,579 non-coding transcripts pass this strict cutoff (10% false-positive rate), by comparing the distribution of ratios for annotated and non-coding transcripts, we estimate that a total of 817 non-coding transcripts are circadian.

Identification of RNA polymerase peaks

RNA pol ChIP peaks were identified in the sum of timepoints ZT 32 and ZT 44 hours using a maxgap/minrun approach similar to the first pass of PeakSeq (Rozowsky et al., 2009). All peaks larger than 100 nucleotides and separated by at least 20 nucleotides in the ChIP sample were assembled for thresholds starting from the mean coverage to ten times the mean coverage with increments of one-twentieth mean coverage. The unique peaks were selected and consolidated such that no peak

maximums are within 150 nucleotides of each other. This method accurately captures the wide dynamic range of peaks present in the data. All RNA pol peaks and their enrichment over mock are reported in Table S3 in the associated publication (Vijayan et al., 2011); 87% of RNA pol peaks are enriched over the mock ($P < 0.1$). Those peaks that are not enriched over mock appear to be actual peaks in RNA pol ChIP, but these RNA pol ChIP peaks are smaller than the mock background, which is elevated with respect to the ChIP background after both data sets are normalized for the number of reads (Figure S3.2). All RNA pol peaks were used in analysis and results do not change when only peaks enriched over the mock are used. Figure S3.2 shows peak identification over a representative genomic region.

Distribution of mRNA per cell

The distribution of mRNA per cell was calculated by assuming a total of 1,500 mRNAs per cell (Ingraham et al., 1983; Taniguchi et al., 2010). For each mRNA species $m_1,\ldots, m_{1415}$, the abundance of the species m_i per cell was given by:

$$m_i = r_i \times \frac{1500}{\sum_{i=1}^{i=1415} r_i},$$

where r_i is the mean number of reads per nucleotide within the mRNA species *i*. All mRNA-per-cell estimates are reported in Table S1 in the associated publication (Vijayan et al., 2011). Only mRNAs with r_i greater than 2 are shown in Figure 3.2A.

Calculation of minimum free energy of secondary structure of RNA

Minimum free energy of secondary structure of RNA was calculated with MATLAB Bioinformatics Toolbox command *rnafold* - minimum free energy is calculated

using a thermodynamic nearest-neighbor approach (Wuchty et al., 1999; Matthews et al., 1999) and is reported in kcal/mol. All free energies are calculated on 60-nucleotide RNA fragments using a sliding window of 10 nucleotides.

To test whether minimum free energy changes were dependent on dinucleotide frequency of the RNA, dinucleotide shuffled sequences with the same overall dinucleotide content distribution were generated using a first order Markov model. That is, for each position in the sliding window, the dinucleotide content of all sequences was assembled. Then an equal number of dinucleotide shuffled sequences were randomly generated maintaining the same overall dinucleotide content distribution.

At the 3′ end of transcripts, a dip in minimum free energy was not observed in the dinucleotide shuffled sequences, but was observed in native sequences (Figure S3.5C). In addition, the minimum free energy at the dip in native sequences (mean = -16.11 kcal/mol) was significantly lower than that in dinucleotide shuffled sequences at the same position (mean = -13.95 kcal/mol; Z = -0.52, P = 1.66e-31). Z-scores were calculated as the difference in mean of native and dinucleotide shuffled sequences divided by the standard deviation of dinucleotide shuffled sequences and P-value was calculated using the two-sided Wilcoxon rank sum test. This suggests that a particular stem-loop feature, likely associated with transcription termination, is present at the end of transcripts.

At the RNA pol peaks at the 5′ ends of genes (Figure S3.5A), the change in minimum free energy in native and dinucleotide shuffled sequences was nearly identical, suggesting that changes in dinucleotide (or nucleotide) frequency and not a

discrete stem loop structure are responsible for the transition in free energy. A change in nucleotide content does occur at the position of the RNA pol peaks (Figure S3.5B), and may play a role in RNA pol pausing by an unknown mechanism. A drop in minimum free energy in native and dinucleotide shuffled sequences is also observed globally when all transcripts are aligned by their 5′ end (Figure S3.5D). A similar change in nucleotide content occurs approximately 100 nucleotides from the 5′ end of transcripts (Figure S3.4B). These global sequence changes proximal to the 5′ end of transcripts may coincide with our observation of global RNA pol pausing internal to the 5′ ends of transcripts.

Calculation of DNA melting temperature

Melting temperature was calculated with MATLAB Bioinformatics Toolbox command *oligoprop* - melting temperatures are calculated using a nearest-neighbor approach with default parameters (Sugimoto et al., 1996).

Identification of -10 element in promoters

All unique mRNA transcription start sites were aligned and the +1 to -30 sequences were input into CONSENSUS-V6C (Gertz et al., 1999), which finds a consensus pattern of defined width (width = 8 nucleotides) in unaligned sequences. This procedure identified 5′ --Ta-aaT 3′ motif, corresponding to the -10 element (Pribnow box), with ln(p) = -4092.23 where p is the probability of identifying a motif with the same or higher information content in an arbitrary alignment. This motif was found at slightly different positions in each of the sequences. To identify the true -10 element while removing any potential false positives, the motif from the subset of alignments that

identified the initial nucleotide of the motif at -8 (285 of 1,416 transcripts) is shown in Figure 3.3B. In subsequent searches using CONSENSUS-V6C or other motif algorithms, no motif was found downstream of the -10 motif where a -35 motif may be expected.

Abbreviations

bp, base pair; ChIP, chromatin immunoprecipitation; CT, circadian time; HIP1, highly iterated palindrome 1; JGI, Joint Genome Institute; OD, optical density; ORF, open reading frame; RFAM, RNA Families; RNA pol, RNA polymerase; UTR, untranslated region; ZT, zeitgeber time.

Data availability

All data sets have been uploaded to the Gene Expression omnibus under accession [GEO:GSE29264].

Acknowledgments

We thank members of the O'Shea laboratory for discussion and commentary. We thank Dr Susan Golden for the *S. elongatus* strain AMC 408. This work was supported by the Howard Hughes Medical Institute, National Defense Science and Engineering Fellowship (VV), and National Science Foundation Graduate Research Fellowship (VV).

References

Altuvia S (2007) Identification of bacterial small non-coding RNAs: experimental approaches. *Curr Opin Microbiol* 10:257-261.

Andersson CR, Tsinoremas NF, Shelton J, Lebedeva NV, Yarrow J, Min H, Golden SS (2000) Application of bioluminescence to the study of circadian rhythms in cyanobacteria. *Methods Enzymol* 305:527-542.

Belknap WR, Haselkorn R (1987) Cloning and light regulation of expression of the phycocyanin operon of the cyanobacterium *Anabaena. EMBO J* 6:871-884.

Bustos SA, Golden SS (1991) Expression of the psbDII gene in *Synechococcus* sp. strain PCC 7942 requires sequences downstream of the transcription start site. *J Bacteriol* 173:7525-7533.

Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol* 27:1043-1049.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* 139:5320-5325.

Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* 38:D396-D400.

Duhring U, Axmann IM, Hess WR, Wilde A (2006) An internal antisense RNA regulates expression of the photosynthesis gene isiA. *Proc Natl Acad Sci USA* 103:7054-7058.

Frenkel A, Gaffron H, Battley EH (1950) Photosynthesis and photoreduction by the blue green alga, *Synechococcus elongatus*, Nag. *Biol Bull* 99:157-162.

Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR (2009) Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Sys Biol* 5:305.

Gertz GZ, Stromo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-577.

Grainger DC, Hurd D, Harrison M, Holdsock J, Busby SJW (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Nat Acad Sci USA* 102:17693-17698.

Griffith-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Res* 31:439-441.

Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* 326:1268-1271.

Harley CB, Reynolds RP (1987) Analysis of *E. coli* promoter sequences. *Nucleic Acids Res* 15:2343-2361.

Hecht A, Grunstein M (1999) Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. *Methods Enzymol* 304:399-414.

Hoon MJL, Makita Y, Nakai K, Miyano S (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* 1:e25.

Huber W, Toedling J, Steinmetz LM (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* 22:1963-1970.

Imamura S, Asayama M (2009) Sigma factors for cyanobacterial transcription. *Gene Regul Syst Biol* 3:65-87.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324:218-223.

Ingraham, JL, Maaloe O, Neidhardt FC (1983) *Growth of the Bacterial Cell.* Sunderland, MA: Sinauer Associates.

Ito H, Mutsuda M, Murayama Y, Tomita J, Hosokawa N, Terauchi K, Sugita C, Sugita M, Kondo T, Iwasaki H (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.

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30:42-44.

Kingsford CL, Ayanbule K, Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 8:R22.

Kondo T, Strayer CA, Kulkarni RD, Taylor W, Ishiura M, Golden SS, Johnson CH (1993) Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression. *Proc Natl Acad Sci USA* 90:5672-5676.

Liu Y, Tsinoremas NF, Golden SS, Kondo T, Johnson CH (1996) Circadian expression of genes involved in the purine biosynthetic pathway of the cyanobacterium *Synechococcus* sp. Strain PCC 7942. *Mol Microbiol* 20:1071-1081.

Lam FH, Steger DJ, O'Shea EK (2008) Chromatic decouples promoter threshold from dynamic range. *Nature* 453:246-250.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Larson MH, Greenleaf WJ, Landick R, Block SM (2008) Applied force reveals mechanistic and energetic details of transcription termination. *Cell* 132:971-982.

Luque I, Flores E, Herrero A (1993) Molecular mechanism for the operation of nitrogen control in cyanobacteria. *EMBO J* 13:2862-2869.

Matthews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288:911-940.

McGrath PT, Lee H, Zhang L, Iniesta AA, Hottes AK, Tan MH, Hillson NJ, Shapiro L, McAdams HH (2007) High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol* 25:584-592.

Min H, Liu Y, Johnson CH, Golden SS (2004) Phase determination of circadian gene expression in *Synechococcus elongatus* PCC 7942. *J Biol Rhythms* 19:103-112.

Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J, Hess WR (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci USA* 105:2124-2129.

Mooney R, Davis S, Peters J, Rowland J, Ansari A, Landick R (2009) Regulator trafficking on bacterial transcription units. *Mol Cell* 33:97-108.

Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR (2002) Genetic control by a metabolite binding mRNA. *Chem Biol* 9:1043-1049.

Nakamura T, Naito K, Yokota N, Sugita C, Sugita M (2007) A cyanobacterial non-coding RNA, Yfr1, is required for growth under multiple stress conditions. *Plant Cell Physiol* 48:1309-1318.

Nielsen H, Johansen SD (2009) Group I introns: moving in new directions. *RNA Biol* 6:375-383.

Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH (2009) Structure and complexity of a bacterial transcriptome. *J Bacteriol* 191:3203-3211.

Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G (2009) A strand-specific RNA-Seq Analysis of the transcriptome of the typhoid bacilus Salmonella typhi. *PLoS Genet* 5:e1000569.

Picard F, Robin S, Lavielle M, Vaisse C, Daubin JJ (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics* 6:27.

Pribnow D (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci USA* 72:784-788.

Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 33:880-892.

Reppas NB, Wade JT, Church GM, Struhl K (2006) The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol Cell* 24:747-757.

Robinson NJ, Robinson PJ, Gupta A, Bleasby AJ, Whitton BA, Morby AP (1995) Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res* 23:729-735.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27:66-75.

Schaller H, Gray C, Herrman K (1975) Nucleotide Sequence of an RNA polymerase binding site from the DNA of bacteriophage. *Proc Natl Acad Sci USA* 72:737-741.

Scharnagl M, Richter S, Hagemann M (1998) The Cyanobacterium *Synechocystis* sp. strain PCC 6803 expresses a DNA methyltransferase specific for the recognition sequence of the restriction endonuclease *PvuI. J Bacteriol* 180:4116-4122.

Sharma CM, Hoffmann S, Darfeuille F, Reignier F, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen Heliobacter pylori. *Nature* 464:250-255.

Shibato J, Asayama M, Shirai M (1998) Specific recognition of the cyanobacterial psbA promoter by RNA polymerases containing principle sigma factors. *Biochim Biophys Acta* 1442:296-303.

Steglich C, Lindell D, Futschik M, Rector T, Chisholm SW (2010) Short RNA half-lives in the slow-growing marine cyanobacterium *Prochlorococcus*. *Genome Biol* 11:R54.

Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 24:4501-4505.

Taniguchi Y, Choi PJ, Li GW, Babu M, Hearn J, Emili A, Xie XS (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533-538.

Vijayan V, Zuzow R, O'Shea EK (2009) Supercoiling drives circadian gene expression in cyanobacteria. *Proc Natl Acad Sci USA* 106:22564-22568.

Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47.

Vogel J, Axmann IM, Herzel H, Hess WR (2003) Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. *Nucleic Acids Res* 31:2890-2899.

Wade JT, Struhl K (2008) The transition from transcriptional initiation to elongation. *Curr Opin Genet Dev* 18:130-136.

Wade JT, Struhl K (2004) Association of RNA polymerase with transcribed regions in *Escherichia coli*. *Proc Natl Acad Sci USA* 51:17777-17782.

Washio T, Sasayama J, Tomita M (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res* 26:5456-5463.

Winkler W, Nahvi A, Breaker RR (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 419:952-956.

Wuchty S, Fontana W, Hofacker I, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49:145-165.

Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R: A single-base resolution map of an archael transcriptome. *Genome Res* 2010, 20:133-141.

Yanofsky C (1981) Attenuation in the control of expression of bacterial operons. *Nature* 289:751-758.

CHAPTER 4

Sequence determinants of circadian gene expression phase in cyanobacteria

*This chapter contains text and figures from:

Vijayan V, O'Shea EK (2012). Sequence determinants of circadian gene expression phase in cyanobacteria (in review *Journal of Bacteriology*).

Abstract

The cyanobacterium *Synechococcus elongatus* PCC 7942 exhibits global biphasic circadian oscillations in gene expression. Class I genes are maximally expressed in the subjective dusk whereas class II genes are maximally expressed in the subjective dawn. Here we identify sequence features that encode the phase of circadian gene expression. We find that, for multiple genes, a ~70 nucleotide promoter fragment is sufficient to specify class I or II phase. We demonstrate that gene expression phase can be changed by random mutagenesis and that a single nucleotide substitution is sufficient to change the phase. Our study provides insight into how gene expression phase is encoded in the cyanobacterial genome.

Introduction

The cyanobacterium *Synechococcus elongatus* PCC 7942 (hereafter, *S. elongatus*) exhibits circadian oscillations in gene expression in continuous light conditions (Liu et al. 1995; Ito et al. 2009; Vijayan et al. 2009). Microarray analysis has shown that the expression of at least 30 to 65% of genes oscillate with ~24 hour periodicity (Ito et al. 2009; Vijayan et al. 2009), with two primary phases of gene expression – genes peaking in the subjective dusk (class I) or subjective dawn (class II).

But what determines whether a particular gene oscillates with class I or II phase? Previous studies of the class II *purF* (*synpcc7942_0004)* promoter identified an 89 nucleotide fragment that specifies class II phasing (Min et al. 2004), and analysis of the class I *kaiBC* (*synpcc7942_1217* and *synpcc7942_1216*) promoter identified a 56 nucleotide fragment which specifies class I phasing (Kutsuna et al. 2005). However,

neither study was able to identify the sequence within the fragments that specified phase information, nor were they able to identify mutations that switched the phase. Here, we investigate the sequence features responsible for circadian phase determination. These features may provide insight into the mechanism of circadian gene expression and may aide in understanding cyanobacterial promoter design.

Results and Discussion

Identification of a promoter region sufficient to encode circadian gene expression phase

A previous study analyzing the relationship between sequence and phase in *S. elongatus* identified a long range (~3 kilobase) statistically significant enrichment in AT content (~1%) in both the promoter and open reading frame of genes activated when the chromosome is relaxed versus those that are repressed (Vijayan et al. 2009). These AT content differences were similar in magnitude and location to those found in genes activated and repressed after induction of chromosomal relaxation in *Escherichia coli* (Peter et al. 2004). The concordance in sequence signature, combined with the observation of circadian changes in chromosome supercoiling (Smith and Williams 2006; Woelfle et al. 2007), suggested a role for supercoiling in circadian gene expression in *S. elongatus* (Vijayan et al. 2009).

Although a long range (~ 3 kilobase) enrichment in AT content exists between genes activated when the chromosome is relaxed versus those that are repressed (Vijayan et al. 2009), circadian transcripts (median length 1320 nucleotides (Vijayan et al. 2011)) of a given phase are randomly distributed along the densely transcribed

genome (Ito et al. 2009; Vijayan et al. 2009). This suggests that the relevant sequence information encoding phase is not long range, but more proximal to each transcript. Recent RNA sequencing and transcription start site identification in *S. elongatus* (Vijayan et al. 2011) allows analysis based on transcription start sites as opposed to translation start sites which were used in both of the previous bioinformatic studies (Peter et al. 2004; Vijayan et al. 2009). This added resolution enables a more detailed analysis of sequence content. In the region between -20 and -100 relative to the transcription start site, we find an enrichment of AT content in transcripts that are activated when the chromosome is relaxed (subjective dawn) (Figure 4.1A). To identify the location of the most statistically significant enrichment in AT content, we computed a p-value across the promoter and the transcript, and find a particularly significant p-value – corresponding to a 1 in 14 nucleotide GC to AT substitution – for the sequence between -20 and -30 often called the 'spacer' (Figure 4.1B). This spacer region is directly between the -10 and -35 elements at which the RNA polymerase complex makes its initial contacts (Pribnow 1975; Schaller et al. 1975).

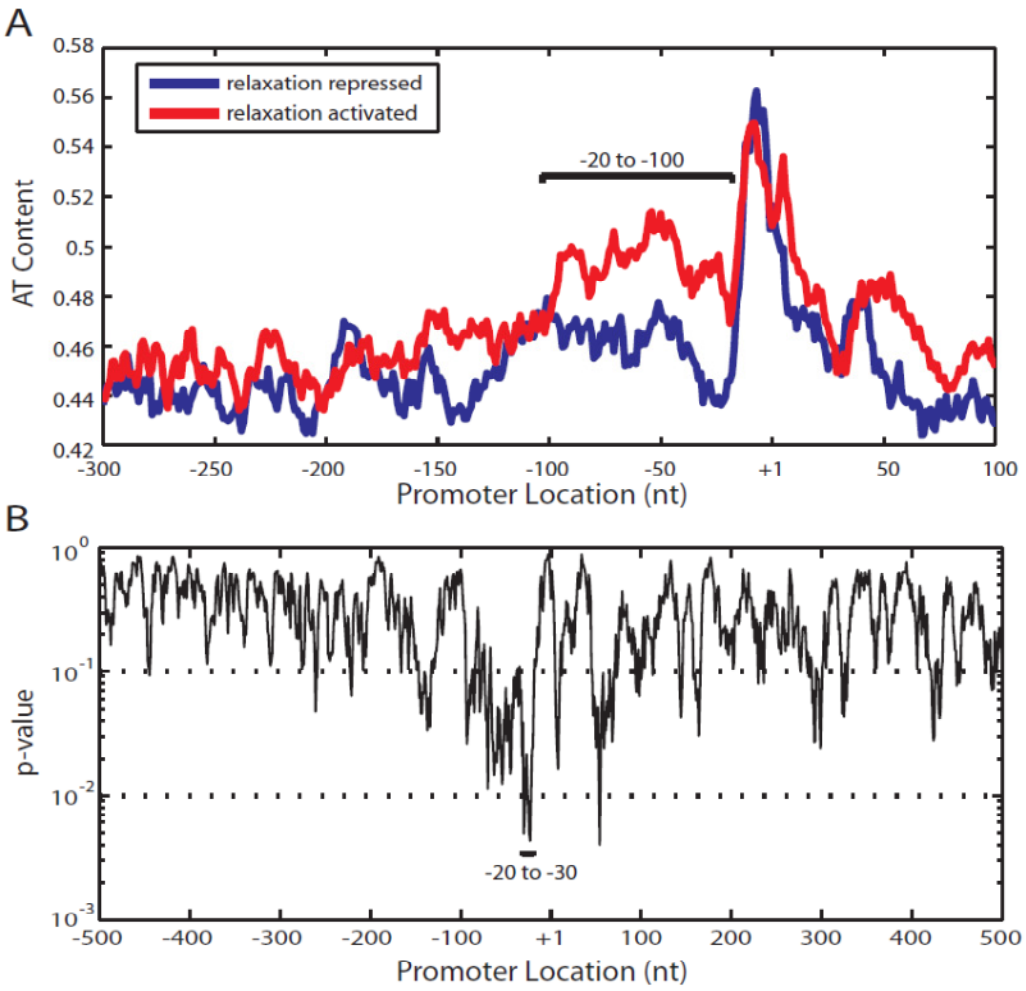


Figure 4.1: A local difference in AT content is observed between transcripts activated or repressed when the chromosome is relaxed.

(A) A comparison of the AT content of transcripts activated or repressed when the chromosome is relaxed. Transcripts are aligned by transcription start location (+1) (Vijayan et al. 2011) and their AT content is averaged. An 11 nucleotide smoothing window is applied to the average AT content. A large difference in AT content exists between -20 and -100. The relaxation repressed transcripts were defined as those with greater than 0.5 correlation with chromosomal supercoiling and the relaxation activated transcripts as those with less than -0.5 correlation (Vijayan et al. 2009).

(B) p-value for AT content difference calculated in 11 nucleotide bins. For every 11 nucleotide bin, the probability of having as extreme an AT enrichment as the relaxation activated set is calculated by 10,000 simulations with randomized sets of relaxation activated and relaxation repressed transcripts. The majority of the bins between -20 and -100 nucleotides are significant ($p < 0.1$), with the most significant bins in the spacer region (-20 to -30).

Identification of a local difference in AT content in the spacer suggested that a single fragment containing the -10, spacer, and -35 elements may be capable of both transcription and encoding circadian phase. To determine if this is the case, we asked if a ~70 nucleotide fragment encompassing these elements from four different circadian transcripts (Figure 2.1A) – two class I and two class II – could drive expression with the same phase as the endogenous transcript. Transcription start site and circadian phase for each of these transcripts were obtained from RNA sequencing (Vijayan et al. 2011) and microarray (Vijayan et al. 2009) experiments, respectively. These fragments (P1 through P4) were fused to a promoterless *luxAB* (luciferase) bioluminescence reporter and subsequently inserted into a defined chromosomal locus, NS 2.1 (Mackey et al. 2007), in the strain AMC 395 (Min et al. 2004) (Materials and Methods). AMC 395 expresses the luciferase substrate, *luxCDE*, using the highly expressed class I *psbAI* promoter (Kondo et al. 1993; Liu et al. 1995; Nair et al. 2001; Mackey et al. 2007). We assume that the luciferase substrate is in excess at all time points. The promoterless *luxAB* alone does not lead to any detectable bioluminescence, but when fused to a promoter fragment can recapitulate the phase of the endogenous transcript (Figure 2.1B). To verify that the bioluminescence reporter accurately reports phase, we confirmed that the phase of mRNA accumulation is also preserved by measuring the abundance of the *luxAB* transcript in strains with the P1 fragment by quantitative PCR (qPCR) (Figure S4.1A). Our results indicate that the information required to encode phase is at least partially contained in a short fragment surrounding the spacer region of the promoter. Although the tested promoter fragments are able to reproduce the phase of circadian gene expression, they do not always preserve the overall level of

bioluminescence. Cells with a much larger ~900 nucleotide version of the P1 fragment (AMC 408 (Liu et al. 1996; Min et al. 2004)) have much higher overall expression than the P1 fragment, even though the phase and amplitude (peak to trough ratio) are identical (Figure 4.2B).

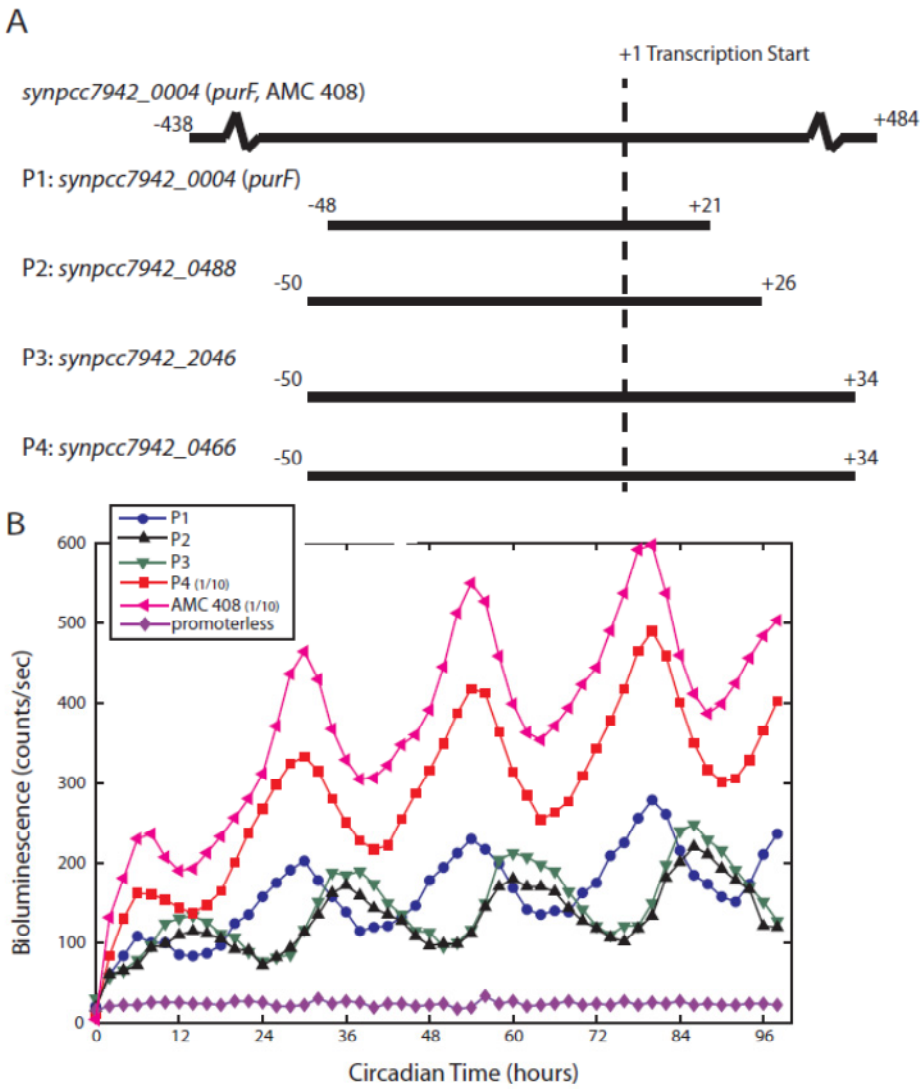


Figure 4.2: A short promoter fragment is sufficient to encode circadian gene expression phase.

(A) Four promoter fragments P1 through P4 were fused to a promoterless *luxAB* cassette in the *S. elongatus* strain AMC 395. P2 and P3 are class I genes (*synpcc7942__0488* and *synpcc7942__2046*); P1 and P4 are class II genes (*synpcc7942__0004* and *synpcc7942__0466*). The promoter fragment from a control class II strain (AMC 408 (Liu et al. 1996; Min et al. 2004)) is shown as a reference. This control strain uses the full length version (~900 nucleotides) of the P1 fragment.

(B) Bioluminescence data collected every two hours indicates that each fragment is sufficient to reconstitute the phase of the endogenous gene. The bioluminescence from a control class II strain (AMC 408) and a promoterless strain are shown as positive and negative controls, respectively. P4 and AMC 408 bioluminescence were scaled by $1/10^{th}$ as indicated.

Random mutagenesis of promoter fragments can change the phase of gene expression

Since we found that the information encoding phase is contained in a ~70 nucleotide fragment, we asked if mutagenesis of this fragment could alter the phase of gene expression. Promoter fragments P1 (class II), P2 (class I), and P3 (class I) were synthesized with a 15% per base substitution rate (5% chance that each of three non-endogenous nucleotides replaces the endogenous nucleotide at each position), fused to the promoterless *luxAB* cassette, and integrated into the NS 2.1 chromosomal locus of AMC 395 (Materials and Methods). A 15% substitution rate was chosen so that at least one substitution could be expected in the spacer region of the promoter. Approximately 200 individual colonies from each library (P1, P2, and P3), each with a unique mutagenized promoter fragment, were assayed for bioluminescence.

Nearly half of the colonies in each library had bioluminescence above background, and all of these colonies also exhibited circadian gene expression oscillations (Figure 4.3 and Figure S4.2). This suggests that the transcription of all transcribed genes oscillates with 24 hour period, in agreement with a previous bioluminescence promoter trap experiment (Liu et al. 1995). Previous microarray measurements reporting that expression of 30 to 65% of the genes oscillates may not have the resolution to detect all oscillations in mRNA abundance (Ito et al. 2009; Vijayan et al. 2009) or there may be an additional translational aspect to the circadian rhythms.

Over 20% of colonies with bioluminescence above background from P1, P2, and P3 exhibited a change in gene expression phase after mutagenesis (Figure 4.3 and Figure S4.2). In addition to the phase of expression, several other characteristics including shape, amplitude, and expression level were affected. To determine which mutations may cause the change in phase, the promoter fragment of each clone from the P1 and P2 libraries with bioluminescence above background was sequenced (Data Set S1 in associated publication). Mutations in clones with altered phase were very diverse in sequence and location. Since on average nearly 1 in every 7 nucleotides is substituted, and a large fraction of colonies changed phase, we expect the majority of the substitutions in phase changing clones to be non-causal.

Figure 4.3: Random mutagenesis of promoter fragments can alter the phase of gene expression from class I to class II and vice-versa.

(A) Random mutagenesis of the promoter fragment P1 from the class II gene, *synpcc7942__0004* (*purF*), yields class I mutants. Top panel: bioluminescence from two biological replicates of P1 and P3 are shown as class II and class I controls, respectively. Bottom panel: bioluminescence from mutant clones with gene expression ordered by phase. Mutants with a phase change are marked with the double lines on the y-axis. All bioluminescence traces have been linearly de-trended and normalized such that the minimum and maximum bioluminescence are 0 and 1 units, respectively.

(B) Random mutagenesis of the promoter fragment P2 from the class I gene, *synpcc7942__0488*, yields class II mutants. Top panel: bioluminescence from two biological replicates of P1 and P2 are shown as class II and class I controls, respectively. Bottom panel: bioluminescence from mutant clones with gene expression ordered by phase. Mutants with a phase change are marked with the double lines on the y-axis. All bioluminescence traces have been linearly de-trended and normalized such that the minimum and maximum bioluminescence are 0 and 1 units, respectively.

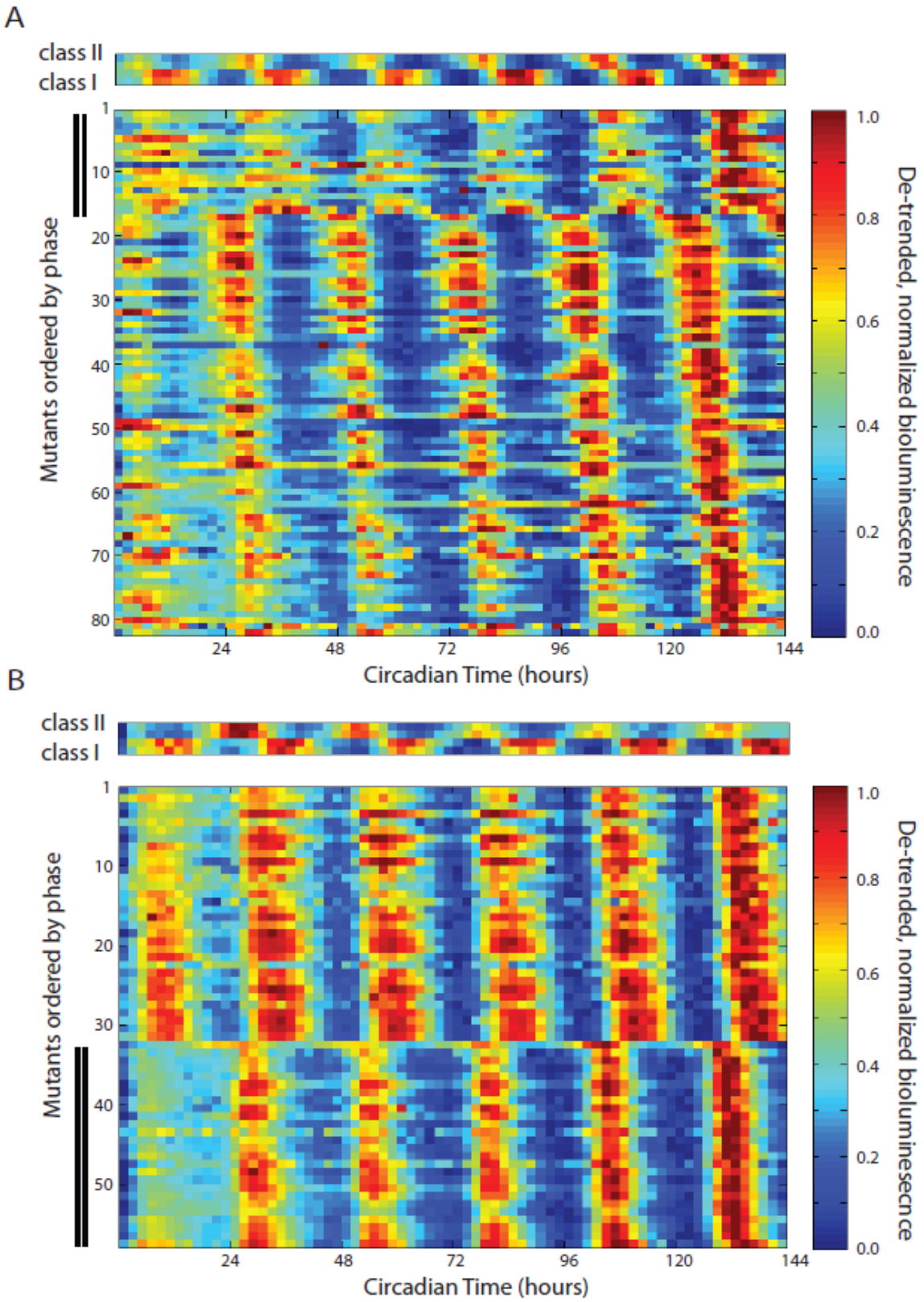


Figure 4.3 Continued

Single nucleotide substitutions are sufficient to change phase of gene expression

Since each promoter fragment contained many substitutions, further subcloning was used to identify which mutations caused the change in phase. Substitutions from two mutagenized promoters with class I phase, M1-1 and M2-1, both from the parent class II P1 library were subcloned to identify the causal substitutions (Figure 4.4). M1-1 has substitutions in 11 of 69 nucleotides and M2-1 in 9 of 69 nucleotides. For M1-1, all strains which retained substitutions at either -2 or-5 or both locations maintained the phase change to class I (see M1-6, M1-7, and M1-8 in Figure 4.4A). The T to C substitution at either -2 or -5 is sufficient to change the phase of the P1 parent promoter from class II to class I. A strain that does not contain either substitution does not change the phase of the P1 parent promoter (see M1-5 in Figure 4.4A). A similar result was observed for M2-1. All strains which retained substitutions at either -12 or -13 or both locations, maintained the phase change to class I (see M2-4, M2-8, and M2-9 in Figure 4.4B). The T to G substitution at either -12 or -13 is sufficient to change the phase of the P1 parent promoter. Three other substitutions downstream of -12 were also sufficient to change phase of P1 to class II, but this strain exhibited very weak rhythmicity (see M2-7 in Figure 4.4B). Quantification of mRNA by qPCR shows altered temporal dynamics of mRNA abundance in all four strains with single nucleotide substitutions (Figure S4.1).

Albeit a small sample size, all four of the causal substitutions in M1-1 and M2-1 increased GC content, consistent with our genome-wide observations (Figure 4.1A). The class II P1 promoter is highly expressed when the chromosome is relaxed (Vijayan et al. 2009), and substitutions increasing the GC content may switch the promoter to be repressed, resulting in a change in phase. Although only two of the four identified single

nucleotide substitutions fall near or within the spacer between the -10 and -35 elements, all substitutions are located proximal to where the RNA polymerase holoenzyme makes initial contacts.

Figure 4.4: Single nucleotide substitutions can change circadian gene expression phase.

(A) A particular clone (M1-1) from random mutagenesis of P1 was analyzed to determine the causal mutations. Top panel: mutations in M1-1 were subcloned in the P1 background (M1-2 through M1-8). All mutations are shown in red, bold, underline. The +1 position is determined from RNA sequencing data (Vijayan et al. 2011). All clones with the T to C substitution at either -2 or -5 change phase from class II to class I. Bottom panel: bioluminescence time-course for promoter fragments shown in top panel. M1-2, M1-3, and M1-4 bioluminescence were scaled as indicated.

(B) A particular clone (M2-1) from random mutagenesis of P1 was analyzed to determine the causal mutations. Top panel: mutations in M2-1 were subcloned in the P1 background (M2-2, M2-4, M2-6, M2-7, M2-8 and M2-9). All mutations are shown in red, bold, underline. The +1 position is determined from RNA sequencing data (Vijayan et al. 2011). All clones with the T to G substitution at either -12 or -13 change phase from class II to class I. Bottom panel: bioluminescence time-course for promoter fragments shown in top panel. M2-6 bioluminescence was scaled as indicated.

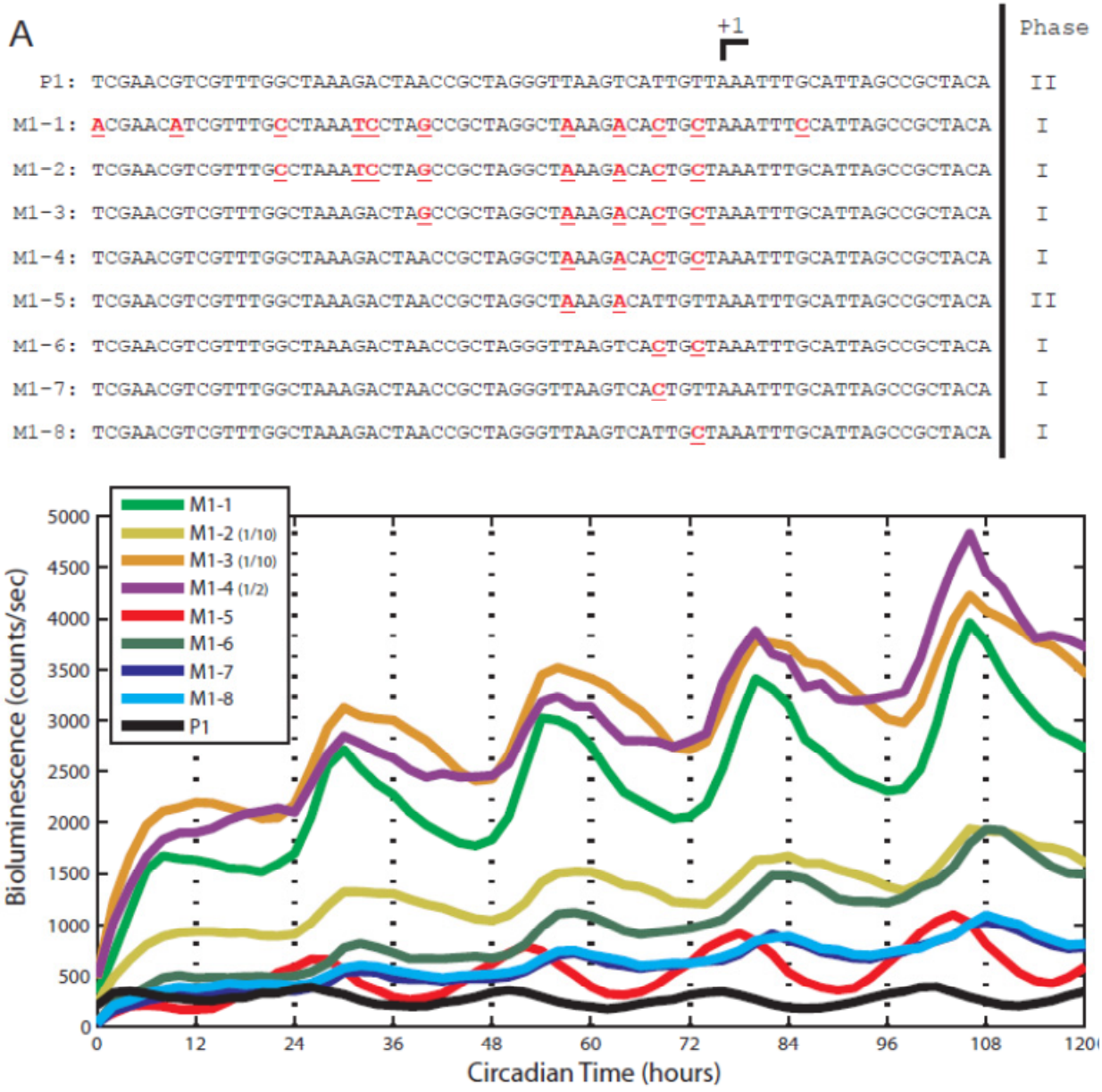


Figure 4.4 Continued

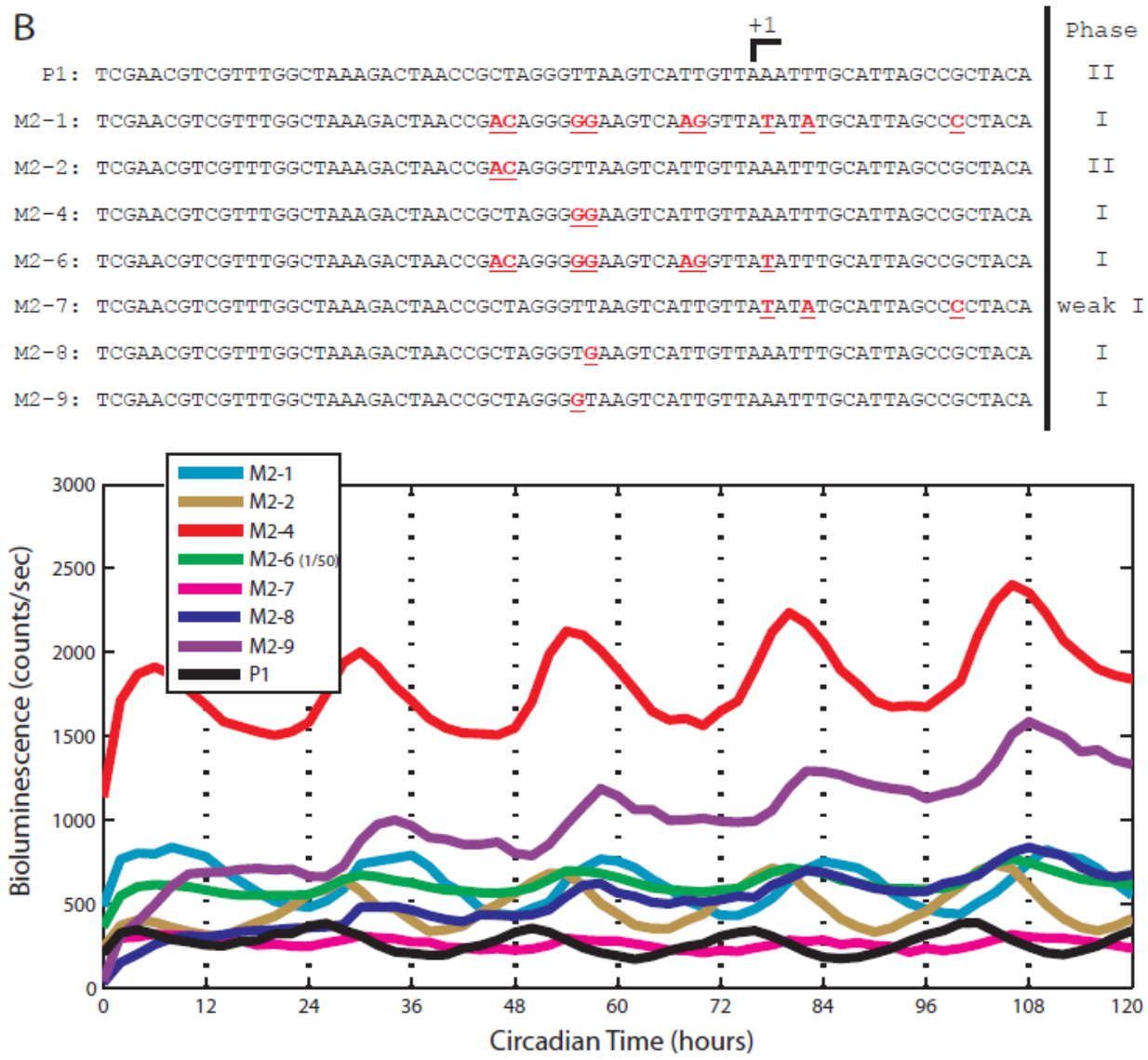


Figure 4.4 Continued

Concluding remarks

Here we have shown that short promoter fragments centered around the spacer region – the region between the -10 and -35 elements – are sufficient to encode circadian phase for multiple circadian genes. Furthermore, we show using random mutagenesis of these fragments that single nucleotide substitutions are sufficient to change circadian gene expression phase.

Previous studies suggested a role for chromosome supercoiling in controlling circadian gene expression in cyanobacteria (Smith and Williams 2006; Woelfle et al. 2007; Vijayan et al. 2009). However, very little is known about the general relationship between sequence and supercoiling sensitivity of promoters in any organism. A genome-wide study in *Escherichia coli* observed a long-range (several kilobase) AT content enrichment in genes repressed by chromosomal relaxation compared to those activated by this perturbation (Peter et al. 2004). Here we identify a local difference in AT content in the spacer of the promoter using transcription start site information, and find that a ~70 nucleotide promoter fragment encompassing this region is sufficient to encode circadian gene expression phase. This promoter fragment has the potential to affect the binding, open complex formation, or even promoter clearance of RNA polymerase complex. Several studies on individual promoters in other organisms have found that the region between -35 and +1 is critical for a promoter's sensitivity to changes in supercoiling (Menzel and Gellert 1987; Straney et al. 1994; Figueroa-Bossi et al. 1998; Niehus et al. 2008), but no consensus mechanism has been identified. Our analysis of the relationship between sequence and phase in *S. elongatus* provides an

entry point for studying mechanism and sequence dependence of supercoiling-mediated gene expression changes.

Our findings suggest that the phase of circadian gene expression is not firmly encoded in the *S. elongatus* genome. Strikingly, even a single nucleotide substitution can dramatically alter the phase of gene expression. Although this lack of sequence structure makes it difficult to design a class I or class II promoter *de novo*, it may serve a role in the fine-tuning of circadian gene expression during the course of evolution. Even though cyanobacteria did not evolve in continuous light conditions, the phase in continuous light is indicative of a gene's expression dynamics in the first twelve hours of light under the more natural light/dark conditions (see first 12 hours of Figure 4.2B). Random mutations in the promoter region have the potential to switch the phase of a gene, and if this phase change is beneficial, it may fix in the population. Since the mutations required to change the phase of a gene are minimal, each gene may be able to sample a different phase in a relatively short period of time. This may explain why almost all of the circadian genes involved in the photosynthesis pathway are more highly expressed in the dawn (Vijayan et al. 2009). This strategy of non-stringent sequence encoding may be applicable to other genome-wide responses where fine-tuning may be beneficial.

Materials and Methods

Cell culture

S. elongatus cells were grown in modified BG-11 medium (hereafter, BG-11M) (Mackey et al. 2007) containing antibiotics at 30° C with cool-white fluorescent

illumination of ~60 μE s m^{-2} (Phillips). Antibiotic concentrations were 2.5 μg ml^{-1} each spectinomycin/streptomycin (Sp/Sm) and 5 μg ml^{-1} chloramphenicol (Cm). Transformations were performed with a few modifications to standard protocols (Mackey et al. 2007). To reduce false-positive colonies, transformations were plated onto a sterile nitrocellulose membrane placed on top of a BG-11M agar plate and kept in low light (~20 μE s m^{-2}) for two days prior to transfer to normal light conditions. On the third and fifth days the nitrocellulose membrane was moved to a new BG-11M agar plate with antibiotics to ensure continuous selection. After ten days, individual colonies were isolated and patched.

Bioluminescence measurements and data analysis

Patched colonies were directly transferred to a transparent 96-well plate with 200 μl of liquid BG-11M containing antibiotics. Multiple independent colonies were selected and assayed. Cells were grown in a clear 96-well plate at ~60 μE s m^{-2} illumination for at least two days. Cells were diluted to OD_{750} ~0.5 and transferred to a black opaque 96-well plate covered with punctured TopSeal (Perkin Elmer) to allow air exchange. Cells were grown in ~60 μE s m^{-2} illumination for one day prior to two consecutive entrainments with 12 hour dark-12 hour light. Cells were then released into continuous light (~60 μE s m^{-2} illumination) and bioluminescence measurements were made every two hours on a TopCount (Perkin Elmer). Prior to each individual bioluminescence measurement, cells were maintained in the dark for three minutes. Five consecutive bioluminescent measurements were made for each well (each integrating incident photons over a one second interval) and subsequently averaged.

Raw bioluminescence data is shown everywhere except in Figure 4.3 and Figure S4.2. For Figure 4.3 and Figure S4.2, bioluminescence data was linearly de-trended and normalized such that minimum and maximum bioluminescence was 0 and 1, respectively. Phase was extracted from the first Fourier component calculated with assumption of a 24 hour period. All mutant promoter fragments are ordered from phase of 0° to 360°. Mutants marked as phase changing were determined by visual inspection. All raw bioluminescence data is provided in Data Set S1 in associated publication.

Cloning and library preparation

Promoter fragments were synthesized as oligonucleotides (Eurofins MWG Operon) with 5'-GCTCTAGA-3' appended to the 5' and 5'-AGGCCTTC-3' appended to the 3'. Sequence of the promoter fragments without appended sequences is: P1 (5'-TCGAACGTCGTTTGGCTAAAGACTAACCGCTAGGGTTAAGTCATTGTTAAATTTGCATTAGCCGCTACA-3'), P2 (5'-TTCCCCGCCTCGCTGACTGAATCTCATTGCCAATCGCTTGCTGCCTCGCCTAGGCTCGGCATAGCACGTGGAAAGG-3'), P3 (5'-TCTCGGCTGGCCCCCCTGTTGTTCCGGACGGGCAGCGGGCAAACTGAAAGCGTCCTCTCTACTTTGTTGCGATGGCGCTGATCT-3'), and P4 (5'-AGCATCACATGGGGCGGATGATAACGGCCCCGTCACGTTAATGTGGGCACATTAACGCCGAAAGATTAAGAGAAAATGACAAGG-3'). Oligonucleotides were annealed to a primer (5'-GAAGGCCT-3'), extended with Klenow (exonuclease-) (NEB) to generate double-stranded DNA, and subsequently cloned into the XbaI and StuI restriction sites of pAM1580 (Min et al. 2004; Mackey et al. 2007). The resulting plasmid was transformed into an *S. elongatus* strain, AMC 395 (Min et al. 2004), expressing the

luxCDE substrate. Mutagenesis libraries were prepared using mutagenized oligonucleotides (Integrated DNA Technologies) with the previously described flanking sequences. Mutagenzied oligonucleotides were synthesized with a 15% substitution rate (5% chance that each of three non-endogenous nucleotides replaces the endogenous nucleotide at each position) in the promoter region. Mutagenized oligonucleotides were primer extended and cloned into pAM1580 as previously described. Over 1000 *Escherichia coli* colonies were combined and plasmid was extracted to generate a plasmid library with sufficient sequence diversity. The plasmid library was subsequently transformed into *S. elongatus* strain AMC 395. The promoter fragment in each resultant *S. elongatus* colony with bioluminescence above background was subjected to colony PCR (primers 5'-GACGGATGGCCTTTTTGCGTTTC-3' and 5'-TGGTGAGTTGTTCAAAATCA-3') and sequenced (sequencing primer 5'-GACGGATGGCCTTTTTGCGTTTC-3').

Quantitative PCR

RNA was extracted every four hours from 800 mL cultures grown in BG-11M supplemented with 10 mM HEPES-KOH pH 8.0 and no antibiotics. Cultures were entrained with two consecutive 12 hour dark-12 hour light periods prior to release into continuous light and manually maintained at an OD_{750} of ~0.3 during sampling. Cultures were bubbled at ~100 mL min^{-1} with ~1% CO_2 in air and were grown at 30° C under ~100 µE s m^{-2} cool white fluorescent lights. 60 mL of cells were collected every four hours by vacuum filtration onto nitrocellulose membranes and subsequently frozen in liquid nitrogen. RNA was extracted and reverse transcribed into cDNA as previously described (Vijayan et al. 2009). qPCR was performed using SYBR Green qPCR master

mix (Invitrogen) on an MX3000p (Stratagene) qPCR machine. The *hslO* (*synpcc7942_0559*) transcript was used for loading normalization of time-points since its expression is relatively constant over circadian time both by microarray and by RNA polymerase ChIP (Vijayan et al. 2009; Vijayan et al. 2011). Standards for each individual primer pair were created by qPCR of a dilution series of cDNA from an arbitrary time-point. As a result, only the relative level of expression of a single primer pair across a time-course can be compared and not the relative level of one primer pair versus another. The following primer pairs were used for qPCR analysis: *luxAB* primers (5'-GTATGAGTCGTACCAATGGC-3' and 5'-GCTACGATGTGACTAAGATT-3'), *hslO* primers (5'-CAGACCAACTGATTCGAGCG-3' and 5'-GGAGGCCAGGAGCAGTC-3'), *kaiBC* primers (5'-TACATTCTCAAGCTCTACG-3' and 5'-CGTCGCTAGGATTTTATCC-3'), and *purF* primers (5'-CTAAGAACCACGAGCTGAC-3' and 5'-CGATCGTCAGGCTAAAGG-3').

Acknowledgments

This work was funded by the Howard Hughes Medical Institute, National Defense Science and Engineering (V.V.) and National Science Foundation Graduate Research Fellowships (V.V.). We thank members of the O'Shea lab for comments and discussion.

References

Figueroa-Bossi, N., et al (1998) The supercoiling sensitivity of a bacterial tRNA promoter parallels its responsiveness to stringent control. *EMBO J* 17(8): 2359-2367.

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 U S A* 106(33): 14168-14173.

Kondo, T., et al (1993) Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria. *Proc Natl Acad Sci U S A* 90(12): 5672-5676.

Kutsuna, S., et al (2005) Transcriptional regulation of the circadian clock operon kaiBC by upstream regions in cyanobacteria. *Mol Microbiol* 57(5): 1474-1484.

Liu, Y., et al (1995) Bacterial luciferase as a reporter of circadian gene expression in cyanobacteria. *J Bacteriol* 177(8): 2080-2086.

Liu, Y., et al (1996) Circadian expression of genes involved in the purine biosynthetic pathway of the cyanobacterium Synechococcus sp. strain PCC 7942. *Mol Microbiol* 20(5): 1071-1081.

Liu, Y., et al (1995) Circadian orchestration of gene expression in cyanobacteria. *Genes Dev* 9(12): 1469-1478.

Mackey, S. R., J. L. Ditty, E. M. Clerico and S. S. Golden (2007) Detection of rhythmic bioluminescence from luciferase reporters in cyanobacteria. *Methods Mol Biol* 362: 115-129.

Menzel, R. and M. Gellert (1987) Modulation of transcription by DNA supercoiling: a deletion analysis of the Escherichia coli gyrA and gyrB promoters. *Proc Natl Acad Sci U S A* 84(12): 4185-4189.

Min, H., Y. Liu, C. H. Johnson and S. S. Golden (2004) Phase determination of circadian gene expression in Synechococcus elongatus PCC 7942. *J Biol Rhythms* 19(2): 103-112.

Nair, U., C. Thomas and S. S. Golden (2001) Functional elements of the strong psbAI promoter of Synechococcus elongatus PCC 7942. *J Bacteriol* 183(5): 1740-1747.

Niehus, E., E. Cheng and M. Tan (2008) DNA supercoiling-dependent gene regulation in Chlamydia. *J Bacteriol* 190(19): 6419-6427.

Peter, B. J., et al (2004) Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. *Genome Biol* 5(11): R87.

Pribnow, D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci U S A* 72(3): 784-788.

Schaller, H., C. Gray and K. Herrmann (1975) Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proc Natl Acad Sci U S A* 72(2): 737-741.

Smith, R. M. and S. B. Williams (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium Synechococcus elongatus. *Proc Natl Acad Sci U S A* 103(22): 8564-8569.

Straney, R., R. Krah and R. Menzel (1994) Mutations in the -10 TATAAT sequence of the gyrA promoter affect both promoter strength and sensitivity to DNA supercoiling. *J Bacteriol* 176(19): 5999-6006.

Vijayan, V., I. H. Jain and E. K. O'Shea (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5): R47.

Vijayan, V., R. Zuzow and E. K. O'Shea (2009) Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proc Natl Acad Sci U S A* 106(52): 22564-22568.

Woelfle, M. A., Y. Xu, X. Qin and C. H. Johnson (2007) Circadian rhythms of superhelical status of DNA in cyanobacteria. *Proc Natl Acad Sci U S A* 104(47): 18819-18824.

CHAPTER 5

Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria

*This chapter contains text and figures from:

Jain IH*, Vijayan V*, O'Shea EK (2012). Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria. *Proc Natl Acad Sci USA* 109(34):13638-13643.

*These authors contributed equally

Abstract

Many cyanobacteria have been shown to harbor multiple chromosome copies per cell, yet little is known about the organization, replication, and segregation of these chromosomes. Here, we visualize individual chromosomes in the cyanobacterium *Synechococcus elongatus* via time-lapse fluorescence microscopy. We find that chromosomes are equally spaced along the long axis of the cell and are interspersed with another regularly spaced subcellular compartment, the carboxysome. This remarkable organization of the cytoplasm along with accurate mid-cell septum placement allows for near-optimal segregation of chromosomes to daughter cells. Disruption of either chromosome ordering or mid-cell septum placement significantly increases the chromosome partitioning error. We find that chromosome replication is both asynchronous and independent of the position of the chromosome in the cell, and that spatial organization is preserved following replication. Our findings on chromosome organization, replication, and segregation in *Synechococcus elongatus* provide a basis for understanding chromosome dynamics in bacteria with multiple chromosomes. percent of non-coding regions are transcribed on either the plus or minus strand.

Introduction

Several cyanobacterial species contain multiple complete copies of their single chromosome (Mann et al., 1974; Binder et al., 1990; Binder et al., 1995; Mori et al., 1996; Griese et al., 2011; Watanabe et al., 2012). This is in contrast to more traditionally studied bacteria, such as *E. coli*, that typically contain only one or two complete chromosome copies (Sherratt et al., 2003. The existence of multiple chromosome

copies raises intriguing questions concerning chromosome positioning, replication, and segregation. For example, how are multiple copies of a chromosome organized within the bacterial cell? How are chromosomes segregated to daughter cells? How is replication regulated and coordinated between chromosome copies? Here, we investigate such questions by visually tracking chromosomes through multiple cell divisions in the cyanobacterium *Synechococcus elongatus* PCC 7942 (hereafter, *S. elongatus*).

S. elongatus is a rod-shaped bacterium with multiple and varying numbers of copies of a ~2.7 Mb chromosome. The existence of ~1 to 6 chromosomes in *S. elongatus* has been documented by distributions of DNA staining fluorescence (Binder et al., 1990; Binder et al., 1995; Mori et al., 1996) and PCR-based methods (Griese et al., 2011; Watanabe et al., 2012). Peaks in DNA staining fluorescence not restricted to 2^n chromosomes (Binder et al., 1990; Binder et al., 1995) and whole-genome sequencing of bromodeoxyuridine (BrdU)-labeled DNA (Watanabe et al., 2012) have suggested that chromosome replication is asynchronous – not all chromosomes replicate simultaneously. Individual replication loci have been visualized in single cells by comparing isolated BrdU-positive regions to more widespread 4',6-diamidino-2-phenylindole (DAPI) staining (Watanabe et al., 2012). While indirect and static methods have been used to investigate chromosome number and replication in *S. elongatus*, direct visualization of replication events in live cells has not been reported and may aide in our understanding of DNA replication in cyanobacteria.

DNA staining of another cyanobacterium with multiple chromosome copies, *Synechocystis* PCC 6803, has shown significant variance in DNA content between

daughter cells (Schneider et al., 2007), suggesting unequal partitioning of chromosomes upon cell division. Nucleoid separation in *Synechocystis* takes place very late in the cell cycle – immediately prior to completion of cell septum formation – consistent with the idea that the act of constriction partitions chromosomes to daughter cells. These observations led to the proposal that chromosome segregation is random and passive – that there is no specific machinery dedicated to partitioning chromosomes at cell division (Schneider et al., 2007). This random partitioning mechanism may be tolerated in cells harboring many copies of their chromosome because daughters are likely to get at least one chromosome by chance, and subsequent chromosome replication may compensate for variance created during cell division.

Bacteria are now known to have intricate subcellular architectures which play an important role in physiology (Gitai et al., 2005). DAPI staining in *S. elongatus* has shown DNA to exist throughout the cell volume (Watanabe et al., 2012; Smith et al., 2006), but little is known about how each individual chromosome copy is organized within the cell and its relation to other cytoplasmic components. The organization of chromosomes within the cytoplasm may play a pivotal role in partitioning during cell division or in the localization of other biomolecules and cytoplasmic components. By observing individual chromosomes within the cell, we investigate the organization of the cyanobacterial cytoplasm and its role in chromosome segregation.

Results

To investigate chromosome location, replication and segregation in *S. elongatus*, we visualized individual genomic loci in live cells using the Fluorescent Repressor-

Operator System (Robinett et al., 1996; Straight et al., 1996; Gordon et al., 1997). Individual chromosomal loci are visualized by binding of fluorescently-tagged repressor proteins to a tandem array of operator sites. Specifically, we inserted lactose (*lac*) operator arrays and tetracycline (*tet*) operator arrays into different chromosomal sites, approximately 200° apart along the circular chromosome (Figure 5.1A). The *tet* operator arrays are positioned 11° from the putative replication origin (*oriC*) (Watanabe et al., 2012; Liu et al., 1996) and the *lac* operators 33° from the putative replication terminus (*terC*) (Watanabe et al., 2012). Tandem arrays with 120 operators were interspersed with heterogeneous sequences to reduce instability (Lau et al., 2003). Expression of tagged repressor proteins, EYFP-LacI and TetR-ECFP, allowed for the simultaneous tracking of genomic loci near the origin and terminus, in live cells (Figure 5.1B). Fluorescent loci are dependent on binding of the repressor proteins, as *tet* or *lac* loci disappear upon addition of saturating anhydrotetracycline (aTC) or isopropyl β-D-thiogalactoside (IPTG), respectively.

DAPI staining has previously suggested that *S. elongatus* chromosomes occupy a large volume (Watanabe et al., 2012; Smith et al., 2006). Imaging of strains with both *lac* and *tet* operators revealed that identical genomic loci are generally further apart than different genomic loci (Figure 5.1C), suggesting that inter-chromosomal distances are greater than intra-chromosomal ones. The *tet* and *lac* genomic loci appear to exist in pairs within the cell (Figure 5.1C), implying that individual chromosomes occupy separate territories within a cell, rather than being randomly dispersed or being clustered by genomic position. The existence of pairs of *tet* and *lac* genomic loci also suggests that most chromosomes in *S. elongatus* are fully or close-to-fully replicated

since the *tet* operators are positioned close to the putative terminus. Occasional unpaired *tet* loci represent chromosomes undergoing replication (see Figure 5.3). On average there are 3.3 ± 2.5 chromosomes in exponentially growing cells (Figure S5.1). There is a linear relationship between chromosome number and mean cell length (r^2 = 0.96) (Figure 5.1D).

Since chromosomes form distinct territories within cells, subsequent experiments were performed with cells containing only *tet* operators and TetR-EYFP, and these foci were used as a proxy of chromosome location. We measured the position of chromosomes within cells, and found them to be equally spaced along the long axis of the cell (Figure 5.1E).

Figure 5.1: Chromosomes form domains and are ordered along the length of the cyanobacterial cell.

(A) Genomic loci are labeled using fluorescently tagged repressor proteins and corresponding operator arrays. The *tet* operator array is located at 11º and the *lac* operator array at 213º on the chromosome relative to the putative replication origin (*oriC*). The *lac* operator array is 33º from the putative replication terminus (*terC*) (Watanabe et al., 2012; Liu et al., 1996).

(B) Distinct genomic loci can be visualized when tagged repressors are bound to operator arrays (TetR-ECFP bound to *tet* operators and EYFP-LacI bound to *lac* operators). A single z-section is shown.

(C) Left: computational identification of cells and operator arrays (cyan and red dots), and calculation of cell width and length (black lines). Multiple z-sections are used to identify chromosomes within the cell (Materials and Methods). Right: The distances between a given *tet* operator (cyan dot) and (*i*) the nearest *tet* operator (cyan dot) and (*ii*) the nearest *lac* operator (red dot) were calculated. The ratio of these distances (see schematic) is plotted as a histogram. Over 75 % of ratios are greater than one implying that identical genomic loci are generally further apart than different genomic loci. A total of 28849 pairs were analyzed.

(D) The number of chromosome copies in a cell as a function of cell length. Red dots represent the average cell length for cells with a given number of chromosomes. The correlation between the average cell length and chromosome number is $r^2 = 0.96$. A total of 1341 cells were analyzed.

(E) The position of chromosomes along the major axis of the cell (normalized for cell length), for cells containing one, two, three, or four chromosomes.

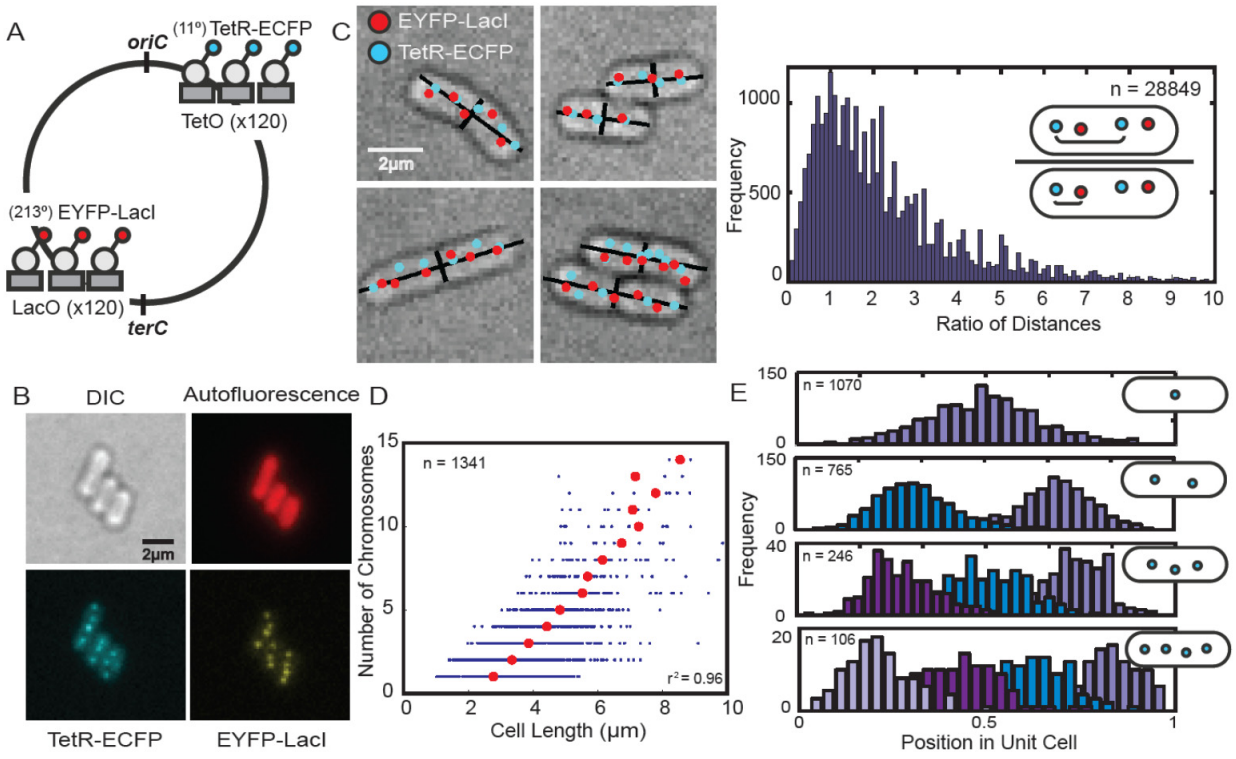


Figure 5.1 Continued

Chromosome ordering and mid-cell septum placement minimize chromosome partitioning error

We followed chromosomes through multiple generations of cell division (Video S1, Video S2 in associated publication) (Jain et al., 2012). Cells were grown on an agarose pad under continuous light, resulting in a doubling time of approximately 12 hours. In wild-type cells, the cell septum forms precisely at the mid-cell position (Figure 5.2A). The combination of spatial ordering of chromosomes and accurate mid-cell placement of the septum may allow for nearly-equal partitioning of chromosomes upon cell division. Indeed, quantification of chromosomes partitioned to each daughter cell shows nearly equal partitioning (Figure 5.2B). In addition, we find that cells have an average of 5.2 ± 1.5 chromosomes prior to cell division (Figure 5.2B), suggesting that cells do not commit to cell division immediately after reaching a particular number of chromosomes.

To investigate the role of septum placement in chromosome segregation, we followed segregation in a strain lacking *minD* (*Synpcc7942_0896*) (Video S3, Video S4 in associated publication) (Jain et al., 2012), which encodes a protein that plays a role in regulating cell septum position in bacteria (Lutkenhaus et al., 2007; Miyagishima et al., 2005). We find that *S. elongatus ΔminD* cells inaccurately position the cell septum along the cell length (Figure 5.2C). Despite inaccurate septum placement, chromosomes still remain ordered along the length of the cell in the *ΔminD* strain (Figure S5.2), suggesting that chromosome ordering and cell septum placement are independent events. Misplacement of the cell septum in *ΔminD* cells results in daughter cells receiving unequal numbers of chromosomes (Figure 5.2D), with the number of chromosomes

partitioned to each daughter cell determined by the position of the septum (Figure S5.3A, Figure S5.3B). Thus, no active mechanism exists to ensure that equal numbers of chromosomes are partitioned to daughter cells – if the septum is misplaced, chromosomes are simply partitioned based on their initial location in the cell and the position of the septum.

To quantify the partitioning error in cells, we defined the error as the statistical difference in chromosome number between daughters averaged across cell division events (Huh et al., 2011) (Figure 5.2E). An error of zero corresponds to perfectly equal segregation to daughter cells and one corresponds to a situation in which one daughter cell receives all chromosomes and the other receives none. To compare the segregation error of wild-type cells to cases of disordered chromosomes or random septum placement, we performed the following simulations: (*i*) perfectly ordered chromosomes and perfect mid-cell septum position; (*ii*) disordered chromosomes and mid-cell septum drawn from the wild-type distribution; and (*iii*) perfectly ordered chromosomes and random cell septum placement (Figure 5.2E, Materials and Methods). For each case, the result of 1000 independent simulations of cell populations is plotted as a histogram. The ideal partitioning error is not zero because cells may have an odd number of chromosomes at the time of cell division. The observed chromosome partitioning error in wild-type cells is slightly greater than the ideal scenario, possibly due to slight errors in ordering (Figure 5.1E), cell septum placement (Figure 5.2A), or miscounting of insufficiently separated chromosomal loci. However, the experimental partitioning error for wild-type cells is significantly lower than for simulations of disordered chromosomes or misplaced septa. The partitioning error for simulations of

cells with ordered chromosomes and random cell septum placement is in agreement with experimental observations of *ΔminD* cells. Together, these results imply that chromosome ordering combined with accurate mid-septum placement permits near-ideal accuracy of chromosome partitioning to daughter cells. Disruption of either process results in partitioning errors that may adversely affect the fitness of a cell population.

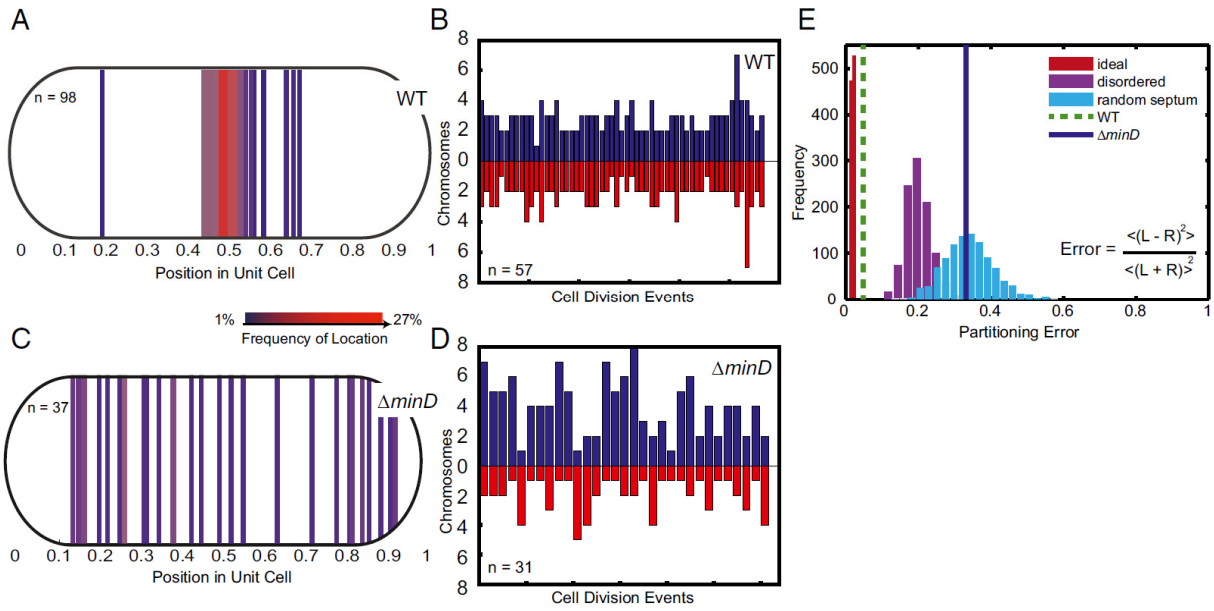


Figure 5.2: Chromosome ordering and mid-cell septum formation enhance the accuracy of chromosome partitioning.

(A) The site of septum formation in wild-type cells during cell division. Each line corresponds to a single cell division site and the color represents the frequency of the septum site.

(B) The number of chromosomes partitioned to pairs of daughter cells upon cell division. Each pair of bars represents a single cell division event, with the height corresponding to the number of chromosomes received by a daughter cell.

(C) The site of septum formation in the *ΔminD* strain.

(D) The number of chromosomes partitioned to pairs of daughter cells upon cell division in the *ΔminD* strain.

(E) The partitioning error for chromosome segregation in wild-type and the *ΔminD* strain. Simulations of partitioning error for: (*i*) ideal chromosome segregation (perfect ordering and perfect mid-cell septa) (red histogram); (*ii*) disordered chromosome segregation (disordered chromosomes and wild-type distribution of septa) (purple histogram); and (*iii*) random septum (perfect ordering and random septum) (cyan histogram). Inset gives equation for partitioning error with R and L representing the number of chromosomes segregated to each individual right and left daughter cell pair. Brackets represent averages of all cell division events of the population.

Asynchronous and position-independent chromosome replication

The repressor-operator system allows for visualization of replicating chromosomes. Since our *tet* operator arrays are positioned only 11° from the putative replication origin (Watanabe et al., 2012; Liu et al., 1996), we can visualize chromosomes undergoing replication. We visualized replicating chromosomes by imaging cells every fifteen minutes under constant light conditions and observe that nascent (newly-replicated or replicating) chromosomes become spatially ordered within a 45 minute window, and typically only one chromosome replicates at a given time (Figure 5.3, Figure S5.4). The 45 minutes it takes to order a nascent chromosome is much shorter than the average cell generation time of 12 hours, suggesting that ordering is a dynamic process, not necessarily coupled to cell growth. Upon following different cells, each containing three chromosomes and undergoing a single replication event, we found the first, second, and third chromosomes replicated 17, 18, and 12 times, respectively. Thus, chromosome replication is independent of the position of the chromosome in the cell.

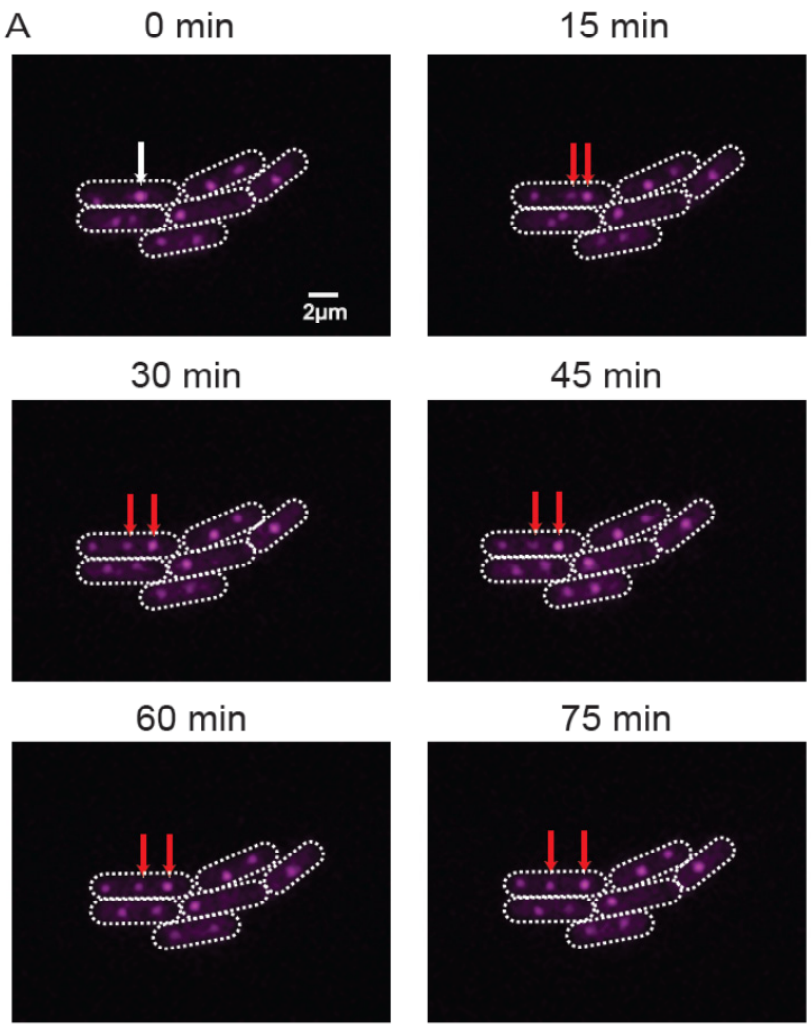


Figure 5.3: Chromosome replication is asynchronous within individual cells.

(A) A time course of wild-type cells. A single genomic locus proximal to the origin is labeled using *tet* operator arrays (pink dots). White arrows point to a replicating chromosome and red arrows point to the resulting, replicated chromosomes. A single z-section is shown.

Spatial organization of chromosomes and carboxysomes

Cyanobacteria have carbon fixation organelles known as carboxysomes, which were recently shown to be ordered along the long axis of the *S. elongatus* cell (Savage et al., 2010). Since we found chromosomes to display a similar pattern of ordering, we investigated the spatial relationship between these two components. To visualize carboxysomes, we expressed an ECFP fusion of the Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) large chain protein – RbcL (Savage et al., 2010), which localizes to the carboxysomes. When RbcL-ECFP is expressed in the repressor-operator strain, chromosomes and carboxysomes show a striking, alternating pattern (Figure 5.4A) with carboxysomes positioned closer to chromosomes than to other carboxysomes (Figure 5.4B).

To further investigate the relationship between chromosomes and carboxysomes, we deleted the *parA*-like gene *Synpcc7942_1833* (hereafter *parA*), encoding a member of a family of proteins known to play a role in chromosome and plasmid segregation in many bacteria (Toro et al., 2010). Although the deletion of *parA* in *S. elongatus* affects carboxysome alignment (Savage et al., 2010), it does not affect chromosome positioning (Figure 5.4A). Interestingly, the position of chromosomes and carboxysomes still remain mutually exclusive in the *ΔparA* strain. Thus, either independent mechanisms regulate the spacing of carboxysomes and chromosomes, or carboxysomes are organized by chromosomes and this organization is mediated via ParA (Thompson et al., 2006; Roberts et al., 2012; Thanbichler et al., 2008).

To identify a possible mechanism for spatial chromosome organization, we attempted to disrupt three additional cytoskeletal genes that may have involvement in chromosome organization (*ftsZ* (*Synpcc7942_2378*), *mreB* (*Synpcc7942_0300*), and a *parA*-like gene (*Synpcc7942_0220*) (Nakao et al., 2010)). Complete deletions of *mreB* and *ftsZ* could not be obtained, suggesting that these genes may be essential for viability in *S. elongatus*. Deletion of the *parA*-like gene had no effect on chromosome alignment and did not result in any detectable phenotypes. The *parA*-like genes (*Synpcc7942_1833* and *Synpcc7942_0220*) in *S. elongatus* differ from those with a role in plasmid or chromosome segregation in other bacteria (Toro et al., 2010) in that they do not have an associated *parB* and are located over 30° from the origin. A *parAB* gene (*Synpcc7942_B2637* and *Synpcc7942_B2626*) does exist on a 46 kb endogenous plasmid in *S. elongatus*, but its function was not investigated.

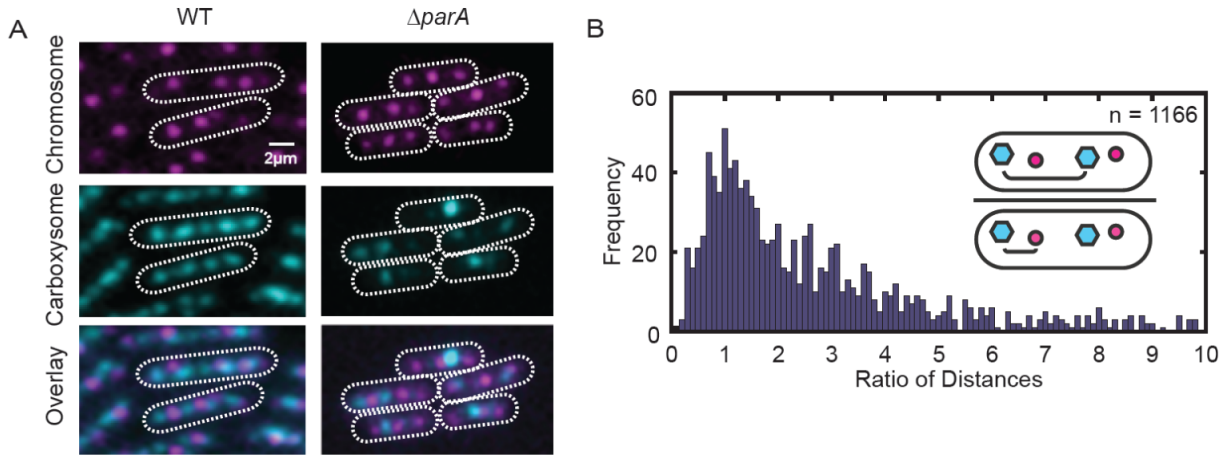


Figure 5.4: Chromosomes and carboxysomes are spatially mutually exclusive.

(A) Chromosome and carboxysome positioning in wild-type compared to *parA* mutants. A single z-section is shown.

(B) The distances between a given carboxysome (cyan dot) and (*i*) the nearest carboxysome (cyan dot) and (*ii*) the nearest chromosome (red dot) were calculated. The ratio of these distances (see schematic) is plotted as a histogram. Over 75 % of ratios are greater than one implying that carboxysomes are generally further apart than a carboxysome and the nearest chromosome. A total of 1166 pairs were analyzed.

Discussion

Chromosome organization and segregation have not been previously characterized in single *S. elongatus* cells. The principles of chromosome organization, segregation, and replication in *S. elongatus* may also be applicable to other bacteria with multiple chromosome copies. By tracking genomic loci through cell divisions, we find that chromosomes form separate domains that are linearly ordered along the cell length. The spatial organization of chromosomes combined with accurate mid-septum placement aids in near-optimal partitioning of chromosomes to daughter cells. In addition, we find that chromosome replication is asynchronous and position-independent, and that the bacterial cytoplasm is arranged with chromosomes and carboxysomes alternating along the long axis of the cell.

One potential benefit of chromosome ordering is that the concentration of molecules that localize near chromosomes may be uniform across the cell length (Montero Llopis et al., 2010). In addition, the observed correlation between chromosome number and cell length suggests a possible mechanism for gene dosage compensation. As the cell grows, a chromosome is replicated, maintaining a relatively constant ratio of genetic material to cell volume. The dilution of a molecule during cell growth may trigger chromosome replication. After a chromosome is replicated, the concentration of the molecule may be restored by the additional genetic material. In this manner, a coupling of chromosome replication to cell growth may provide the basis for gene dosage compensation in bacteria with multiple chromosome copies. Our observation of asynchronous chromosome replication is unique to *S. elongatus.* Well-studied bacteria such as *E. coli* undergo synchronous DNA replication through the

combined actions of SeqA and DnaA (Kaguni et al., 2006). The mechanism of asynchronous chromosome replication remains to be determined in *S. elongatus*.

When combined with accurate mid-septum placement, chromosome ordering also plays an important role in segregation of chromosomes to daughter cells. Previous studies in the spherical cyanobacterium *Synechocystis* showed random segregation of chromosomes to daughters (Schneider et al., 2007). The difference in segregation patterns between *Synechocystis* and *S. elongatus* may be related to their respective chromosome copy numbers – *Synechocystis* has ten times as many chromosome copies (Griese et al., 2011) as *S. elongatus* and therefore may be able to safely rely on random chromosome partitioning. Spatial ordering of chromosomes and accurate mid-septum placement may be a prominent feature of bacteria with an intermediate number of chromosomes where random segregation can lead to large variation between daughters and occasional anucleate cells. Inaccurate septum placement may be a tolerated feature when chromosomes are ordered since the number of chromosomes per unit of cell volume remains the same. However, inaccurate septum placement increases the likelihood of anucleate daughter cells, as observed in the *ΔminD* strain (Figure S5.3C), thereby decreasing the fitness of the population.

The natural question that arises is: what mechanism is responsible for chromosome ordering? In one model, chromosomes can be represented as self-avoiding polymers. When placed under spatial constraints, conformational entropy may drive chromosome segregation into ordered domains (Jun et al., 2006). Alternatively, there may be a molecular explanation for chromosome repulsion. In several bacterial species, the filamentous protein ParA drives chromosome and/or plasmid segregation

(Toto et al., 2010; Ebersbach et al., 2005). While deletion of *parA* and other candidate cytoskeletal genes did not perturb chromosome ordering in *S. elongatus*, we cannot eliminate the possibility of such a system.

We find that chromosomes and carboxysomes are ordered in the cytoplasm in a manner that is spatially mutually exclusive, suggesting that spatial constraints may influence ordering in the cell. Each *S. elongatus* cell is roughly 700 nm in width, but the majority of this space is occupied by thylakoid membranes (Schwarz et al., 2005). The ~100 nm diameter carboxysomes appear to occupy a significant fraction of the cytoplasm in transverse section (Savage et al., 2010; Schwarz et al., 2005), possibly occluding the chromosome. As a result, the simultaneous ordering of the two components may increase the accuracy of each individual component's ordering. The observation that *parA* deletion perturbs carboxysome ordering, but not chromosome ordering suggests two possible models. Either ParA is responsible for ordering carboxysomes using chromosome ordering as a template (Thompson et al., 2006; Roberts et al., 2012), or chromosomes and carboxysomes are ordered independently, but spatial constraints influence their ultimate positioning.

In this study, we characterize chromosome ordering, replication and segregation in *S. elongatus*. We observe an intricate organization of the cyanobacterial cytoplasm and demonstrate its role in chromosome segregation. However, the molecular mechanisms underlying the organization of chromosomes in the cytoplasm remain unknown. Further investigation of the mechanisms responsible for chromosome ordering in *S. elongatus* will improve our understanding of chromosome segregation and spatial organization in bacteria.

Materials and Methods

Strain construction and cloning

Relevant strains, plasmids, and primers are presented in Table S5.1, Table S5.2, and Table S5.3. Wild-type *Synechococcus elongatus* PCC 7942 was acquired from ATCC Catalog # 33912. Unless otherwise noted, *S. elongatus* were grown in modified BG-11 media (BG-11M) (Mackey et al., 2007) at 30° C with cool-white fluorescent illumination of 4000-6000 lux and all appropriate antibiotics. Antibiotic concentrations were 2.5 μg/ml each spectinomycin/streptomycin (Sp/Sm), 5 μg/ml chloramphenicol (Cm), 5 μg/ml kanamycin (Kan), and 2 μg/ml gentamycin (Gm). Transformations were performed following standard conditions (Mackey et al., 2007). If additional selection was necessary to reduce false-positive colonies, transformations were plated onto a sterile nitrocellulose membrane placed on top of a BG-11M agar plate and kept in low light (1500 lux) for two days prior to transfer to normal light conditions. On the third and fifth days the nitrocellulose membrane was moved to a fresh BG-11M agar plate to ensure continuous selection. After ten days, individual colonies were isolated and patched.

S. elongatus cells were always transformed in the following order: (*i*) *lac* and/or *tet* operator arrays, (*ii*) deletion vector (if necessary), (*iii*) carboxysome marker (if necessary), and (*iv*) TetR and/or LacI fusion proteins. TetR and/or LacI fusion proteins were always transformed last, and anhydrotetracycline (aTC) and/or isopropyl β-D-1-thiogalactopyranoside (IPTG) were required to mitigate growth defects during transformation and propagation of strains with both DNA-binding protein and operator

array. TetR/TetO strains required aTC at 5-7.5 µg/mL, and LacI/LacO strains required IPTG at 1 mM. Patched colonies were generally prepared for microscopy 3-5 days after patching, since cells stored for longer periods of time often lost operator arrays.

Two new *S. elongatus* integration sites were developed to insert *lac* operators into the genome. These sites, A and B, were strategically chosen to not affect transcription. They are situated in a region of negligible transcription between convergent transcripts as verified by RNA sequencing (Vijayan et al., 2011). Multiple-cloning-site (MCS) and chloramphenicol cassette for integration vectors were obtained from pAM1573 (Mackey et al., 2007), and cloned between 1 to 1.5 kilobase of upstream and downstream homologous sequence. Individual PCR products were first assembled using fusion PCR and subsequently integrated into pBR322 using GeneArt Seamless Cloning and Assembly Kit (Invitrogen). All PCR primers are provided in Table S4.3.

120 *tet* and 120 *lac* operator repeats with interspersed heterogeneous sequences were obtained from eBB110 (Marquis et al., 2008) and pLAU43 (Lau et al., 2003), respectively. Integration plasmids with operators were designed as shown in Table S5.1 and Table S5.3. Figure 5.1 shows cells with *lac* operators in site A and *tet* operators in neutral site 2.1 (NS 2.1) (Mackey et al., 2007). All other figures and analysis was performed on cells with only *tet* operators in NS 2.1.

C-terminal TetR and N-terminal LacI fluorescent protein fusions were used to visualize chromosomes. The tetramerization domain (last 12 amino-acids) of LacI was deleted. Expression of fusion proteins was driven by the *kaiBC* promoter. Levels of fusion proteins did not appear to have appreciable circadian oscillations (Chabot et al.,

2007). Sequences and annotation for TetR and LacI fusion protein constructs are provided in Table S5.4.

To visualize carboxysomes, RbcL-ECFP was expressed from neutral site 2.2 (NS 2.2) (Mackey et al., 2007) while the native copy of RbcL remained intact (Savage et al., 2010). The fusion protein was expressed using the *apcA* promoter. ECFP was obtained from pJRC23 (Chabot et al., 2007). Individual PCR products were assembled using fusion PCR and subsequently ligated between *SmaI* and *XhoI* of EB2065 (NS 2.2). All PCR primers are provided in Table S5.3.

Gene knockouts were generated by deletion of the gene with a gentamycin resistance cassette obtained from pAM2055 (Mackey et al., 2007). Constructs were created by flanking the gentamycin cassette with at least 700 bases of upstream and downstream homologous sequence by fusion PCR. The PCR product was subsequently ligated into pUC18. All PCR primers are provided in Table S5.3. Knockouts were created by transforming *S. elongatus* with the appropriate plasmid, patching single colonies, and verifying segregation by PCR.

Microscopy

Cells growing on plates were scraped from patched colonies within 3-5 days of patching and washed three times in 50 µl of BG-11M and re-suspended in 5 µl BG-11M. 1 µl of cells were placed in a Lab-Tek® II Chamber Coverglass Chamber (Electron Microscopy Sciences, Catalog # 70377-11) and overlaid with a 1.5 cm x 1.5 cm x 1 cm 2% ultrapure agarose (Invitrogen) pad with antibiotics. In general, to minimize stress, only the antibiotics selecting for operator arrays and repressor proteins were used

during imaging. Pads for TetR/TetO strains were supplemented with 0.005 - 0.01 μg/ml aTC to allow for growth. LacI/LacO strains did not require supplementation with IPTG for growth. Still images were taken on pads without aTC or IPTG. To prevent drying, sponges were soaked in water and placed along the inside of the chamber. Cells were allowed to equilibrate in chamber under 4000 lux illumination for at least 8 hours prior to imaging.

Samples were imaged on a Zeiss Inverted scope with a 100x 1.4 numerical aperture objective and equipped with a Photometrics Evolve 512 EMCCD camera. The microscope was controlled using Axiovision software (Zeiss). Photosynthetic lighting at approximately 4000 lux was constantly provided by a white LED ring light (Advanced Illumination, Part # RL1360-WHI-C2) and turned off only prior to exposures by Axiovision software. Cells were maintained in a 30° C, 0.5-1% CO_2 chamber throughout the experiment. Multiple positions (1 to 10) were imaged at regular time-intervals between 15 to 120 minutes. At each time-point, RFP (autofluorescence), CFP, YFP, and DIC were measured at five different z-stacks (300 nm intervals). For longer time courses, CFP exposure was limited to the middle z-stack to limit phototoxicity. Definite focus (Zeiss) was used to maintain focus over the duration of the experiment. Division times varied between 10 and 14 hours for wild-type cells with the operator-repressor system.

Image Analysis

Image analysis was performed using custom software written in Matlab (MathWorks). Images were corrected for x-y drift and segmented. DIC and CFP images

were used to distinguish cell boundaries and the autofluorescence at RFP emission wavelength was used to eliminate background noise. Identified objects were filtered by size and defined as cells. Inaccurately segmented cells were manually corrected. Custom software tracked the mother-daughter lineage by determining the maximum overlap in coordinates of cells through different time-points. In this manner, lineage was tracked through multiple generations.

Chromosomes and carboxysomes were assigned coordinates using custom Matlab software. Briefly, the YFP or CFP image was convolved with a Gaussian filter. The Laplace operator was applied to the filtered image (using an approximation kernel) to sharpen maxima and minima, defining boundaries between dots and the surrounding regions. Subsequently, a threshold was applied to create a binary image. Appropriate threshold values were determined by finding the number of foci identified in an image for a range of threshold values. The inflection point of this distribution was used as the final threshold value. Finally, the Matlab function *regionprops* was used to determine properties of closed regions (foci). To ensure that chromosomal foci on different focal planes were being identified, five 300 nm z-stacks were acquired at each time-point. Centroids were projected from 3D to 2D space. Foci that were within 320 nm of each other along the x-y plane were collapsed into the same foci so that chromosomes moving during z-stack acquisition were not identified as multiple chromosomes. After cell segmentation, *regionprops* was used to determine cell length and superimpose a rectangular coordinate system on each cell. The location of dots were determined in polar coordinates and transformed to a unit cell for analysis of relative spacing of dots within a cell.

Partitioning Error Simulations

The partitioning error was calculated to represent the difference in chromosomes partitioned to daughter cells, normalized for the total number of chromosomes in the mother cell (see inset Figure 5.2E) [18]. The partitioning error was calculated for wild-type and *ΔminD* cells using experimental data from Figure 5.1B and Figure 5.1D, respectively. Chromosome partitioning error was also simulated for the following conditions: *(i)* ordered chromosomes and perfect mid-cell septum placement, *(ii)* disordered chromosomes and a wild-type distribution of mid-cell septum placement and *(iii)* ordered chromosomes with random septum placement. For each scenario, 1000 independent experiments were simulated, with each simulated experiment consisting of the same number of cell divisions and chromosome numbers as the actual experiments (wild-type conditions for *(i)* and *(ii)* and *ΔminD* conditions for *(iii)*). Disordered chromosome partitioning was modeled using a binomial distribution with each chromosome representing an independent trial and the septum location determining the probability of segregation to a daughter cell (ex. perfectly mid-cell septum corresponds to probability of 0.5). Random cell septum placement was modeled using a uniform distribution from 0.05 to 0.95 cell lengths, representing the observed range of septum placements in *ΔminD* cells.

Acknowledgments

We thank members of the O'Shea lab for comments and discussion regarding the manuscript. We thank Shankar Mukherji for guidance with image processing. We thank the Harvard Center for Biological Imaging and Bernhard Goetze for microscope

access and assistance. This work was funded by the Howard Hughes Medical Institute, National Defense Science and Engineering (V.V.) and National Science Foundation Graduate Research Fellowships (V.V.).

References

Binder BJ, Chisholm SW (1990) Relationship between DNA cycle and growth rate in Synechococcus sp. strain PCC 6301. *J Bacteriol* 172(5):2313-2319.

Binder BJ, Chisholm SW (1995) Cell Cycle Regulation in Marine Synechococcus sp. Strains. *Appl Environ Microbiol* 61(2):708-717.

Chabot JR, Pedraza JM, Luitel P, van Oudenaarden A (2007) Stochastic gene expression out-of-steady-state in the cyanobacterial clock. *Nature* 450(7173):1249-1252.

Ebersbach G, Gerdes K (2005). Plasmid segregation mechanisms. *Annu Rev Genet* 39:453-479.

Gitai Z (2005) The new bacterial cell biology: moving parts and subcellular architecture. *Cell* 120(5):577-586.

Gordon GS et al (1997) Chromosome and low copy plasmid segregation in E. coli: visual evidence for distinct mechanisms. *Cell* 90(6):1113-1121.

Griese M, Lange C, Soppa J (2011) Ploidy in cyanobacteria. *FEMS Microbiol Lett* 323(2):124-131.

Huh D, Paulsson J (2011) Random partitioning of molecules at cell division. *Proc Natl Acad Sci USA* 108(3):15004-15009.

Jain IH, Vijayan V, O'Shea EK (2012) Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria. *Proc Natl Acad Sci USA XXX.*

Jun S, Mulder B (2006) Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. *Proc Natl Acad Sci USA* 103(22):12388-12393.

Kaguni J (2006) DnaA: Controlling the Initiation of Bacterial DNA Replication and More. *Annu Rev Microbiol* 60:351-371.

Lau IF et al (2003) Spatial and temporal organization of replicating *Escherichia coli* chromosomes. *Mol Microbiol* 49(3):731-743.

Liu Y, Tsinoremas NF (1996) An unusual gene arrangement for the putative chromosome replication origin and circadian expression of dnaN in Synechococcus sp. strain PCC 7942. *Gene* 172(1):105-109.

Lutkenhaus J (2007) Assembly Dynamics of the Bacterial MinCDE System and Spatial Regulation of the Z Ring. *Annu Rev Biochem* 76:539-562.

Mackey SR, Ditty JL, Clerico EM, Golden SS (2007) Detection of rhythmic bioluminescence from luciferase reporters in cyanobacteria. *Methods Mol Biol* 362: 115-129.

Mann N, Carr NG (1974) Control of Macromolecular Composition and Cell Division in the Blue-green Alga Anacystis nidulans. *J Gen Microbiol* 83:399-405.

Marquis KA et al (2008) SpoIIIE strips proteins off the DNA during chromosome translocation. *Genes Dev* 22(13):1786-1795.

Miyagishima SY, Wolk CP, Osteryoung KW (2005) Identification of cyanobacterial cell division genes by comparative and mutational analysis. *Mol Microbiol* 56(1):126-143.

Montero Llopis P et al (2010) Spatial organization of the flow of genetic information in bacteria. *Nature* 466(7302):77-81.

Mori T, Binder B, Johnson CH (1996) Circadian gating of cell division in cyanobacteria growing with average doubling time of less than 24 hours. *Proc Natl Acad Sci USA* 93(19):10183-10188.

Nakao M et al (2010) CyanoBase: the cyanobacteria genome database update 2010. *Nucleic Acids Res* 38:D379-381.

Roberts MA, Wadhams GH, Hadfield KA, Tickner S, Armitage JP (2012) ParA-like protein uses nonspecific chromosomal DNA binding to partition protein complexes. *Proc Natl Acad Sci USA* 109(17):6698-6703.

Robinett CC et al (1996) In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *J Cell Biol* 135:1685-1700.

Savage DF, Afonso B, Chen AH, Silver PA (2010) Spatially ordered dynamics of the bacterial carbon fixation machinery. *Science* 327(5970):1258-2161.

Schneider D, Fuhrmann E, Scholz I, Hess WR, Graumann PL (2007) Fluorescence staining of live cyanobacterial cells suggest non-stringent chromosome segregation and absence of a connection between cytoplasmic and thylakoid membranes. *BMC Cell Biol* 8:39.

Sherratt DJ (2003) Bacterial Chromosome Dynamics. *Science* 301(5634):780-785.

Smith RM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium Synechococcus elongatus. *Proc Natl Acad Sci USA* 103(22):8564-8569.

Straight AF, Belmont AS, Robinett CC, Murray AW (1996) GFP tagging of budding yeast chromosomes reveals that protein-protein interactions can mediate sister chromatic cohesion. *Curr Biol* 6(12):1599-1608.

Schwarz R, Forchhammer K (2005) Acclimation of unicellular cyanobacteria to macronutrient deficiency: emergence of a complex network of cellular responses. *Microbiology* 151(8):2503-2514.

Thanbichler M, Shapiro L (2008). Getting organized—how bacterial cells move proteins and DNA. *Nat Rev Microbiol* 6(1):28-40.

Thompson SR, Wadhams GH, Armitage JP (2006) The positioning of cytoplasmic protein clusters in bacteria. *Proc Natl Acad Sci USA* 103(21):8209-8214.

Toro E, Shapiro L (2010) Bacterial chromosome organization and segregation. *Cold Spring Harb Perspect Biol* 2(2):a000349.

Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47.

Watanabe S et al (2012) Light-dependent and asynchronous replication of cyanobacterial multi-copy chromosomes. *Mol Microbiol* 83(4):856-865.

CHAPTER 6

Concluding Remarks

Section 1: Chapters 2 and 4

Here we observed global circadian oscillations in circadian gene expression, provided evidence for supercoiling mediated control, and provided insight into the relationship between phase and sequence.

A balance between two antagonistic enzymes – gyrase and topoisomerase I – keeps the bacterial chromosome in its native negatively supercoiled state (Fisher et al., 1984). Gyrase introduces negative supercoils and topoisomerase I introduces positive supercoils. Negative supercoiling is critical for initiation of bacterial DNA replication and transcription, and its status is highly controlled by the cell. Not surprisingly, it is the target of many antibiotics including nalidixic acid, novobiocin, and ciprofloxacin. In addition, it aides in compacting the bacterial chromosome 10,000 fold to fit inside the cell (Holmes et al., 2000). The topological structure of the bacterial chromosome has been widely debated with estimates of independent domains ranging from 10 to 100kb (Postow et al., 2004; Sinden et al., 1981). Each chromosomal domain is restricted from others such that a nick in the chromosome will only unwind a single domain. The supercoiling density of each domain is expected to be similar since supercoiling repressed and activated genes maintain correlated expression patterns when placed in various positions in the *Escherichia coli* chromosome (Miller et al., 1993).

Several genes have been shown to have expression modulated by supercoiling. Two of the first such identified genes, *gyrA* and *gyrB*, are the two subunits of gyrase itself. Both transcripts are up-regulated in *E. coli* during chromosomal relaxation in an effort to maintain the supercoiling of the chromosome in homeostasis (Menzel et al.,

1983). Analysis of the *gyrA* and *gyrB* promoters has identified that the Pribnow box is particularly important in providing the supercoiling sensitivity of these transcripts (Menzel et al., 1987; Straney et al., 1994). Since then, several other genes whose expression is dependent on supercoiling have been identified, and the regions near and surrounding the promoter have been implicated in imparting the supercoiling sensitivity (See Chapter 4 for more details). More recently several genome wide studies have shown large scale gene expression changes in bacteria due to environmental stresses that change supercoiling levels (See Chapter 2 for more details). A study by Cozzarelli and colleagues showed that direct perturbations to supercoiling do in fact cause a large gene expression response (Peter et al., 2004). This study also identified a difference in AT content in the promoter and transcript of genes that are activated and repressed by supercoiling.

How does supercoiling cause differential transcriptional changes? We have shown that a difference in AT content in the promoter exists in genes activated and repressed by supercoiling, but what does this mean? One hypothesis is that the AT content we observe is a proxy for meltability (isomerization) of the promoter. When RNA polymerase binds the promoter, it first forms a closed complex. In order to initiate transcription, an open complex must be formed. Formation of the open complex requires unwinding (isomerization) of approximately 13 bp of DNA (from the -10 element to the transcription start site). Negative supercoiling as well as increased AT content aide in formation of the open complex. Therefore, when the chromosome is relaxed and the energy required for open promoter formation is greater, perhaps only high AT content promoters are able to form the open complex. When the chromosome is

supercoiled, either (1) polymerase redistributes to the high GC content genes – which can now be transcribed – or (2) the open complex in high AT content genes is over-stabilized and promoter clearance is rate-limiting. The first scenario is passive – doesn't require an active repression, while the second scenario requires that at a given level of supercoiling, an optimum AT content facilitates unwinding and subsequent closure of the transcription bubble.

Another hypothesis is that the AT content at the promoter simply decides which sigma factor binds to each promoter and the supercoiling sensitivity of each gene is dependent on the sigma factor pool that determines its transcription. In bacteria, the RNA polymerase holoenzyme consists of a core RNA polymerase and a sigma factor that helps recognize the site (-10 and -35 elements) to initiate transcription. It is possible that particular sigma factors – each with intrinsic supercoiling sensitivity – need to be available in order to provide the supercoiling sensitivity at a given promoter.

A third possibility relies on differential binding of DNA binding proteins or sigma factors at different superhelical densities and sequence contents. For example, the histone-like protein, H-NS, is known to preferentially bind AT rich curved sequences (Zuber et al., 1994) and represses transcription. Negative supercoiling is expected to enhance bending, therefore, H-NS proteins may bind and repress the transcription of AT rich promoters when the chromosome is supercoiled, imparting the supercoiling sensitivity that we observe. H-NS is present in a large enough copy number (20,000) in *E. coli* (Azam et al., 1999) to impart genome wide changes in transcription. It is not too hard to imagine that other DNA binding proteins (IHF, HU, Fis) or sigma factors may have preference to particular sequence content at particular superhelical densities.

The three hypothesis outlined above are simply conjecture, and it is possible that a combination of these hypothesis (or something different altogether) is used in bacterial cells. A few promoters have been studied *in vitro* to determine their mode of supercoiling dependent activation or repression, but no consensus mechanism has been identified. Some promoters are regulated at the transition from closed to open complex, others at the transition from open complex to transcription initiation (Parekh et al., 1996; Lim et al., 2003, Sheridan et al., 1998). To complicate matters, some promoters require the presence of other factors, like IHF or CRP, to accentuate or exhibit their supercoiling sensitivity.

Recent RNA polymerase ChIP-chip studies have identified promoter proximal stalling of RNA polymerase (See Chapter 3). These promoter proximal peaks could be caused by RNA polymerase molecules stuck in either the closed conformation or stuck in the open conformation performing abortive transcription. But, our ChIP sequencing study did not identify RNA polymerase at the promoters of transcripts, but rather within the transcript. Our initial hope with RNA polymerase ChIP was to correlate promoter "stalling" of RNA polymerase with supercoiling mediated transcription activation/repression. If we did in fact notice RNA polymerase stalling at promoters, RNA sequencing of small RNA fragments (~ 9 nucleotides) may identify abortive transcripts that are generated by RNA polymerase molecules unable to clear the promoter. Thus we may be able to distinguish those RNA polymerase molecules stuck in the closed complex from those unable to clear the promoter.

There is still much work to be done to understand the role of supercoiling in controlling gene expression, and several questions still remain concerning circadian gene expression in *S. elongatus.*

First, what controls supercoiling rhythms in *S. elongatus*? Previous experiments have shown that the supercoiling and chromosome compaction rhythms are dependent on a functioning KaiC (Woelfle et al., 2007, Smith et al., 2006). But how can KaiC control supercoiling rhythms? Three of the more likely scenarios are: (1) clock dependent transcriptional or post-transcriptional modulation of ATP/ADP ratio affects the activity of DNA gyrase; (2) clock dependent transcription factors (possibly through SasA/RpaA signal transduction pathway) affect the level of supercoiling modulating proteins; and (3) clock dependent changes in the activity of supercoiling modulated proteins. A clock dependent modulation of ATP/ADP ratio is an intriguing possibility since gyrase is known to be under control of the ATP/ADP ratio in (Westerhoff et al., 1988). In addition, we know the ATP/ADP ratio drops when the cyanobacteria are exposed to dark (Rust et al., 2011), and the chromosome quickly and dramatically relaxes within 20 minutes of dark application (Vijayan, unpublished results). If the circadian clock in *S. elongatus* modulates photosynthesis activity in continuous light, then maybe ATP/ADP levels fluctuate hence modulating gyrase activity.

Second, what else is required for supercoiling mediated gene expression changes? It is likely that several factors are required for supercoiling mediated gene expression changes in the circadian cycle. Global supercoiling changes must be interpreted at each promoter and therefore at the very least, RNA polymerase should be involved. As speculated earlier in this section, sigma factors, histone-like proteins, or

other DNA binding proteins may be required for supercoiling mediated gene expression changes.

Section 2: Chapter 3

Here we provided the first full-genome description of a cyanobacterium. We hope that our analysis will aide in future work in cyanobacteria including those related to circadian rhythms and bioenergy. We were able to identify transcription start and termination sites as well as operon structures and anti-sense/non-coding RNAs. Our initial goal with the project was to (1) identify the transcription start positions for further bioinformatics related to circadian phase; and (2) observe the dynamics of RNA polymerase in circadian genes. In particular, we were hoping to observe RNA polymerase stalling at promoters and use this to infer how supercoiling affects RNA polymerase (See previous section). Instead, we found that most RNA polymerase molecules do not stall at the promoter, but actually pause/stall within transcripts. The motivation and mechanism of RNA polymerase stalling within transcripts may be interesting for future studies. In addition, our observation of an equal number of non-coding/anti-sense and mRNA transcripts suggests a large and unappreciated role for non-coding and anti-sense transcripts (many of which are circadian in *S. elongatus*) in bacteria.

Section 3: Chapter 5

Here we provided the first visualization of chromosomes in live cyanobacterial cells. We found that *S. elongatus* contains multiple copies of each individual chromosome and these copies are aligned along the long-axis of the cell. We find that

this phenomena aides in providing equal segregation of chromosomes to daughter cells. This may be only one of many roles for chromosome ordering – and of course, for subcellular organization in general – in bacteria. For example, do the ordered chromosomes help organize other cytoplasmic components such as carboxysomes? In addition to physiological roles for cytoplasm organization, future studies incorporating circadian time with chromosome visualization may be able to tease out more physiological roles of the circadian clock. Previous studies have noticed a particular circadian time in which cells do not divide (Mori et al., 1996; Yang et al., 2010), but is there also a time at which the cells do not replicate chromosomes? In addition, although our particular method is probably not well-suited for live cell imaging of chromosome compaction, other related methods including histone tagging combined with super-resolution microscopy, may be able to visualize circadian changes in chromosome state over time.

Section 4: Final Remarks

Much of the work presented here is related to chromosomes, circadian gene expression, and the link between these subjects in cyanobacteria. When I started my Ph. D., very little was known about chromosomes or circadian gene expression in cyanobacteria. Over the past four years we have gone from buying our first cyanobacteria strains and fluorescent lights to a deeper understanding of these subjects. Although several questions remain unanswered, we hope to have provided starting points for future studies.

References

Azam TA, Iwata A, Nishimura A et al (1999) Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J Bacteriol* 181:6361–6370.

Fisher LM (1984) DNA supercoiling and gene expression. *Nature* 307(23):686-687.

Garg LC, Diangelo S, JJacob ST (1987) Role of DNA topoisomerase I in the transcription of supercoiled rRNA gene. *Proc Natl Acad Sci USA* 84:3185-3188.

Holmes VF, Cozarelli NR (2000) Closing the ring: Links between SMC proteins and chromosome partitioning, condensation, and supercoiling. *Proc Natl Acad Sci USA* 97:1322-1324.

Lim et al. (2003) Effect of Varying the Supercoiling of DNA on Transcription and Its Regulation. *Biochemistry* 42:10718-10725.

Menzel R, Gellert M (1983) Regulation of the genes for E. coli DNA gyrase: Homeostatic control of DNA supercoiling. *Cell* 35:105-113.

Menzel R, Gellert M (1987) Modulation of transcription by DNA supercoiling: A deletion analysis of the *Escherichia coli gyrA* and *gyrB* promoters. *Proc Natl Acad Sci USA* 84:4185-4189.

Miller WG, Simon RW (1993) Chromosomal supercoiling in Escherichia coli. *Mol Microbiol* 10(3):675-684.

Mori T, Binder B, Johnson CH (1996) Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours. *Proc Natl Acad Sci USA* 93:10183-10188.

Parekh BS, Hatfield GW (1996) Transcriptional activation by protein-induced DNA bending: Evidence for a DNA structural transmission model. *Proc Natl Acad Sci USA* 93:1173-1177.

Peter BJ et al. (2004) Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol* 5:R87.

Postow L, Hardy CD, Arsuaga J, Cozarelli NR (2004) Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev* 18(14):1766-1779.

Reddy KJ, Masamoto K, Sherman DA, Sherman LA (1989) DNA Sequence and Regulation of the Gene (*cpbA)* Encoding 42-Kilodalton Cytoplasmic Membrane Carotenoprotein of the Cyanobacterium *Synechococcus* sp. Strain PCC 7942. *J Bacteriol* 171(6):3486-3493.

Rust M, Golden SS, O'Shea EK (2011) Light-driven Changes in Energy Metabolism Directly Entrain the Cyanobacterial Circadian Oscillator. *Science* 331:220-223.

Sinden RR, Pettijohn DE (1981) Chromosomes in living Escherichia coli cells are segregated into domains of supercoiling. *Proc Natl Acad Sci USA* 78(1):224-228.

Smith RM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium Synechococcus elongatus. *Proc Natl Acad Sci USA* 103(22):8564-8569.

Sheridan SD, Benham CJ, Hatfield GW (1998) Activation of Gene Expression by a Novel DNA Structural Transmission Mechanism That Requires Supercoiling-induced DNA Duplex Stabilization in an Upstream Activating Sequence. *J Biol Chem* 273:21298-21308.

Straney R, Krah E, Menzel R (1994) Mutations in the -10 TATAAT Sequence of the *gyrA* Promoter Affect Both Promoter Strength and Sensitivity to DNA Supercoiling. *J Bacteriol* 176(19):6599-6006.

Westerhoff HV, O'Dea MH, Maxwell A, Gellert M (1988) DNA supercoiling by DNA gyrase. A static head analysis. *Cell Biophys* 12:157-181.

Woelfle MA, Xu Y, Qin X, Johnson CH (2007) Circadian rhythms of superhelical status of DNA in cyanobacteria. *Proc Natl Acad Sci USA* 104:18819-18824.

Yang Q, Pando B, Dong G, Golden SS, Oudenaarden A (2010) Circadian gating of the Cell Cycle Revealed in Single Cyanobacterial Cells. *Science* 327(5972):1522-1526.

Zuber F, Kotlarz D, Rimsky S (1994) Modulated expression of promoters containing upstream curved DNA sequences by the Escherichia coli nucleoid protein H-NS. *Mol Microbiol* 12:231–240.

APPENDIX A

Chapter 2 Supplemental Data

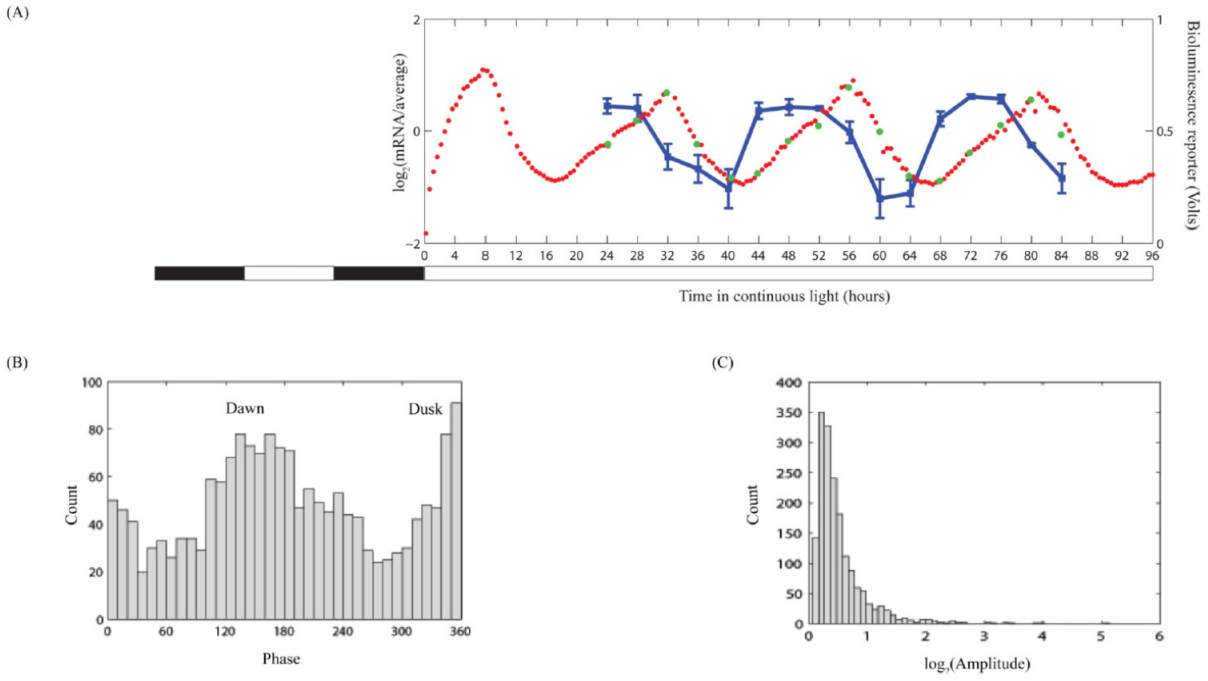


Figure S2.1: Circadian gene expression in *S. elongatus*.

(A) *S. elongatus* strain AMC 408 was subjected to two consecutive alternating 12 hour light-dark cycles for entrainment and subsequently released into continuous light at T = 0 hours. 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 hour intervals between T = 24 and T = 84 (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 1748 circadian genes have amplitude of 2-fold or greater.

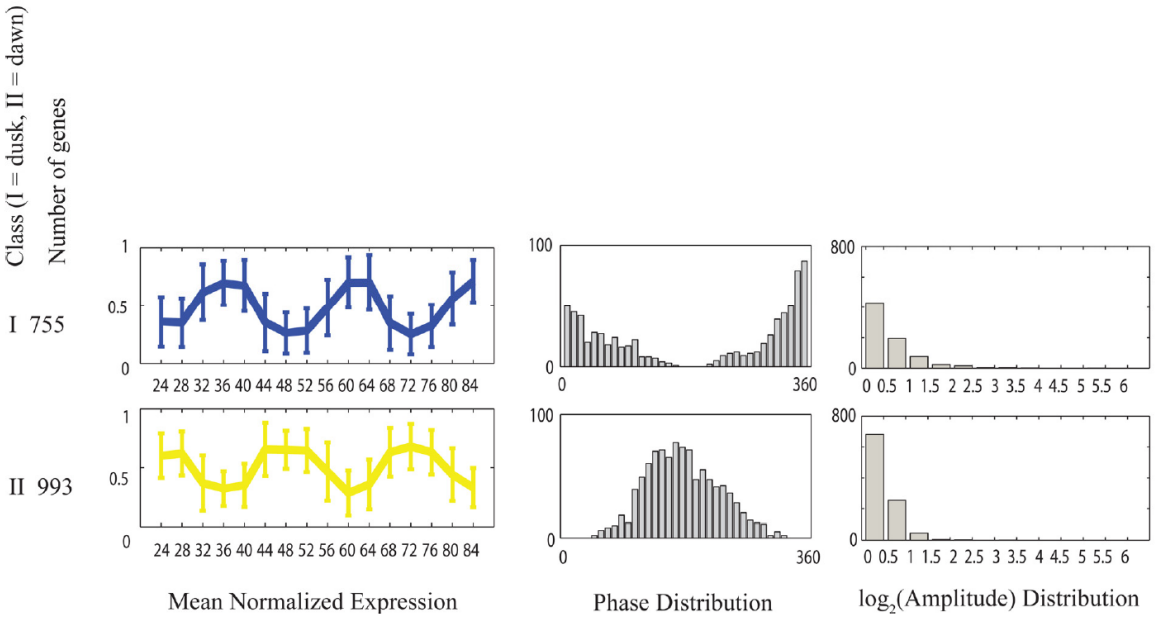


Figure S2.2: K-means clustering for analysis of circadian gene expression.

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 hours in the circadian cycle, and subjective dusk genes were defined as genes whose mRNA levels peaked at 8, 12, or 16 hours in the circadian cycle.

Figure S2.3: Spatial organization of gene expression phase.

(A) Classification (subjective dawn, subjective dusk, or non-circadian), 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 (Stothard et al., 2005). 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: (1) the distance between genes; (2) whether the genes are near each other in other genomes; (3) whether genes both belong to a narrow GO category; and (4) 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.

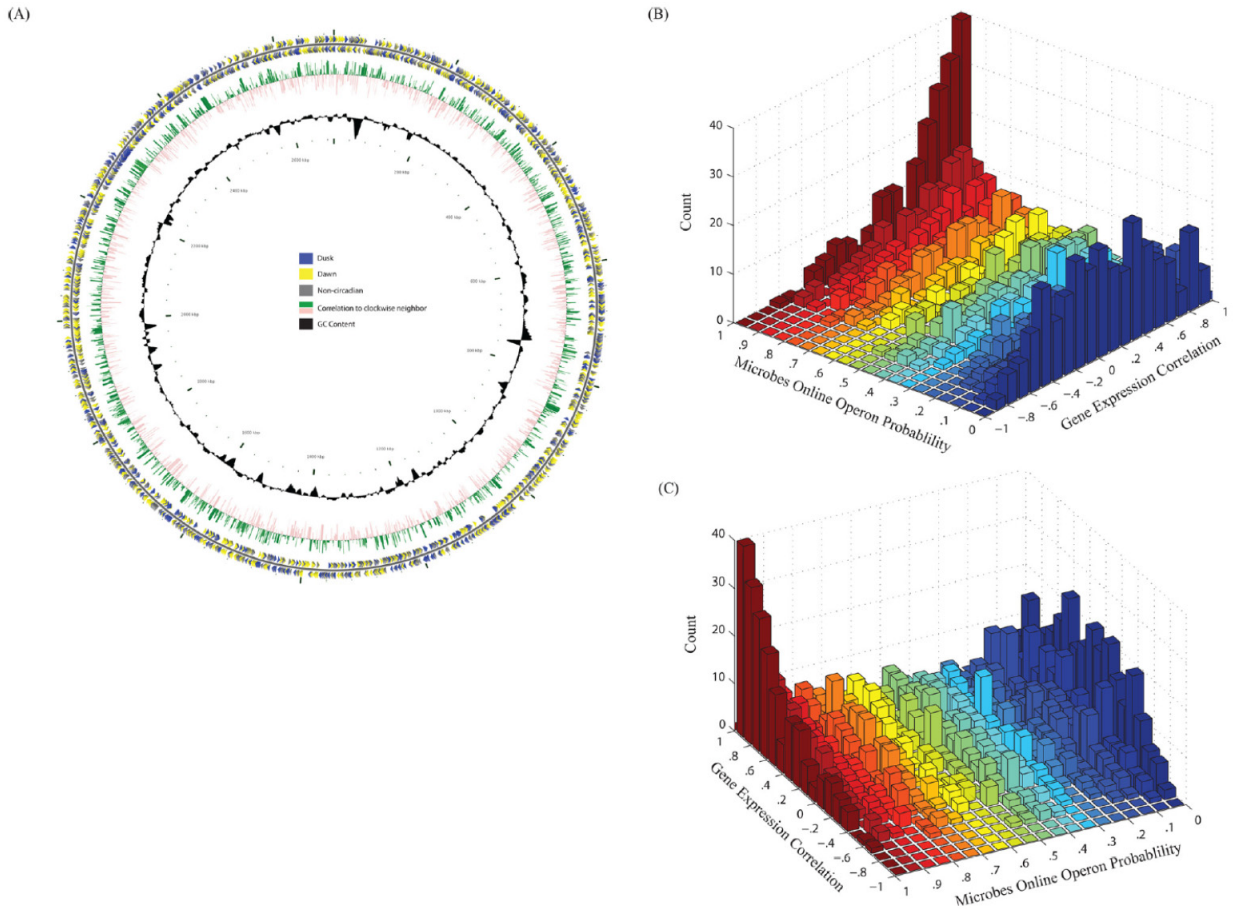


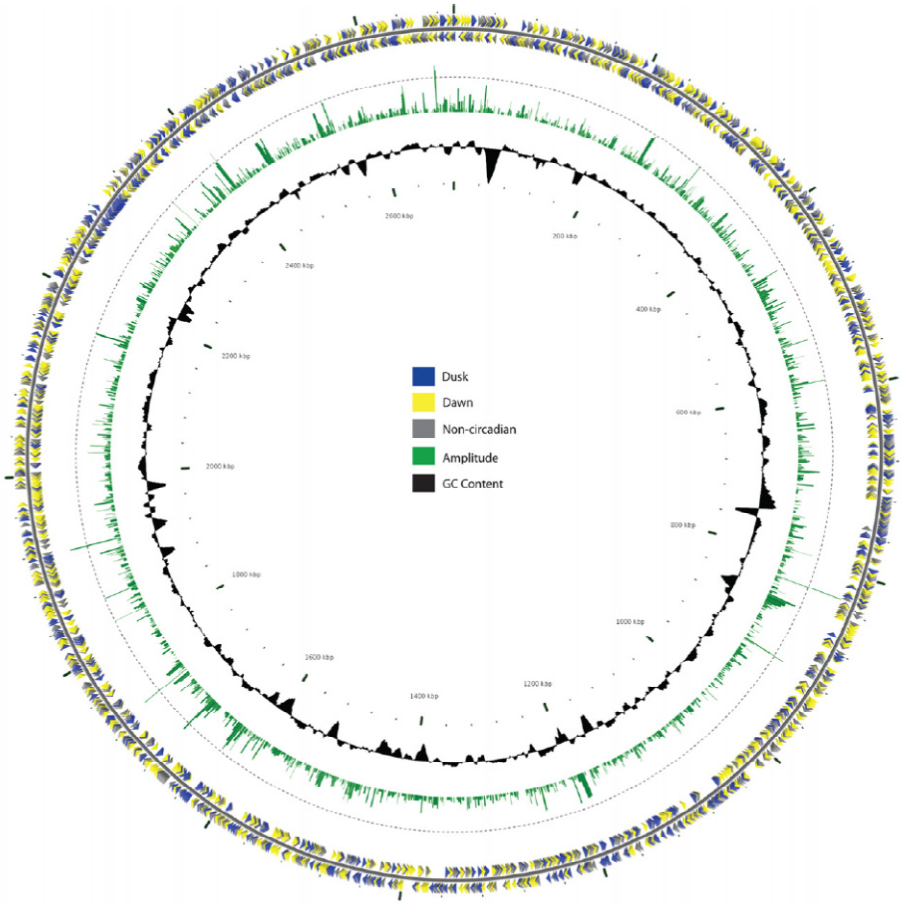
Figure S2.3 Continued

Figure S2.4: Spatial organization of gene expression amplitude.

(A) Classification (subjective dawn, subjective dusk, or non-circadian), amplitude, and GC content along the *S. elongatus* chromosome. Locations are based on Genbank file CP000100 and chromosome diagram was made with CGView (Stothard et al., 2005). 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.

(A)



(B)

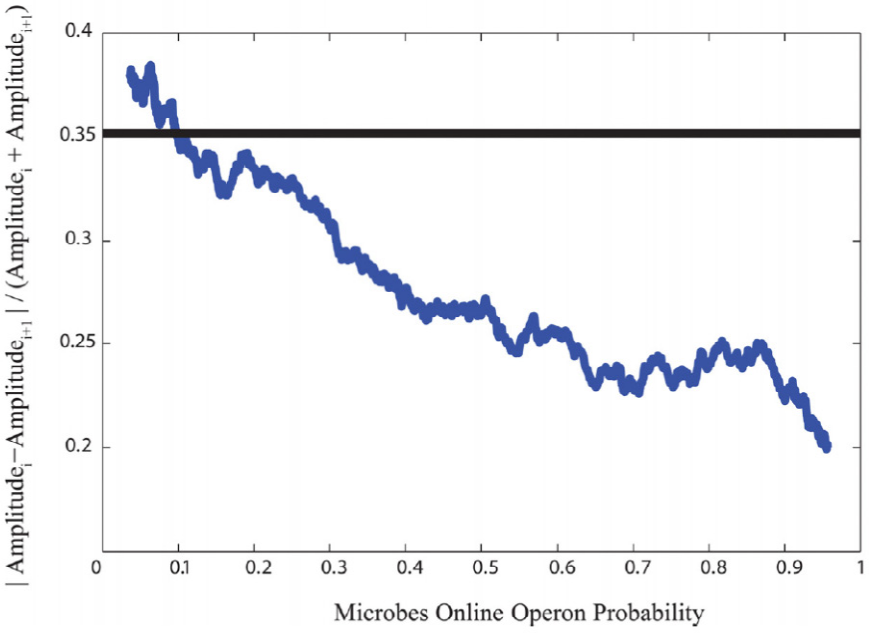


Figure S2.4 Continued

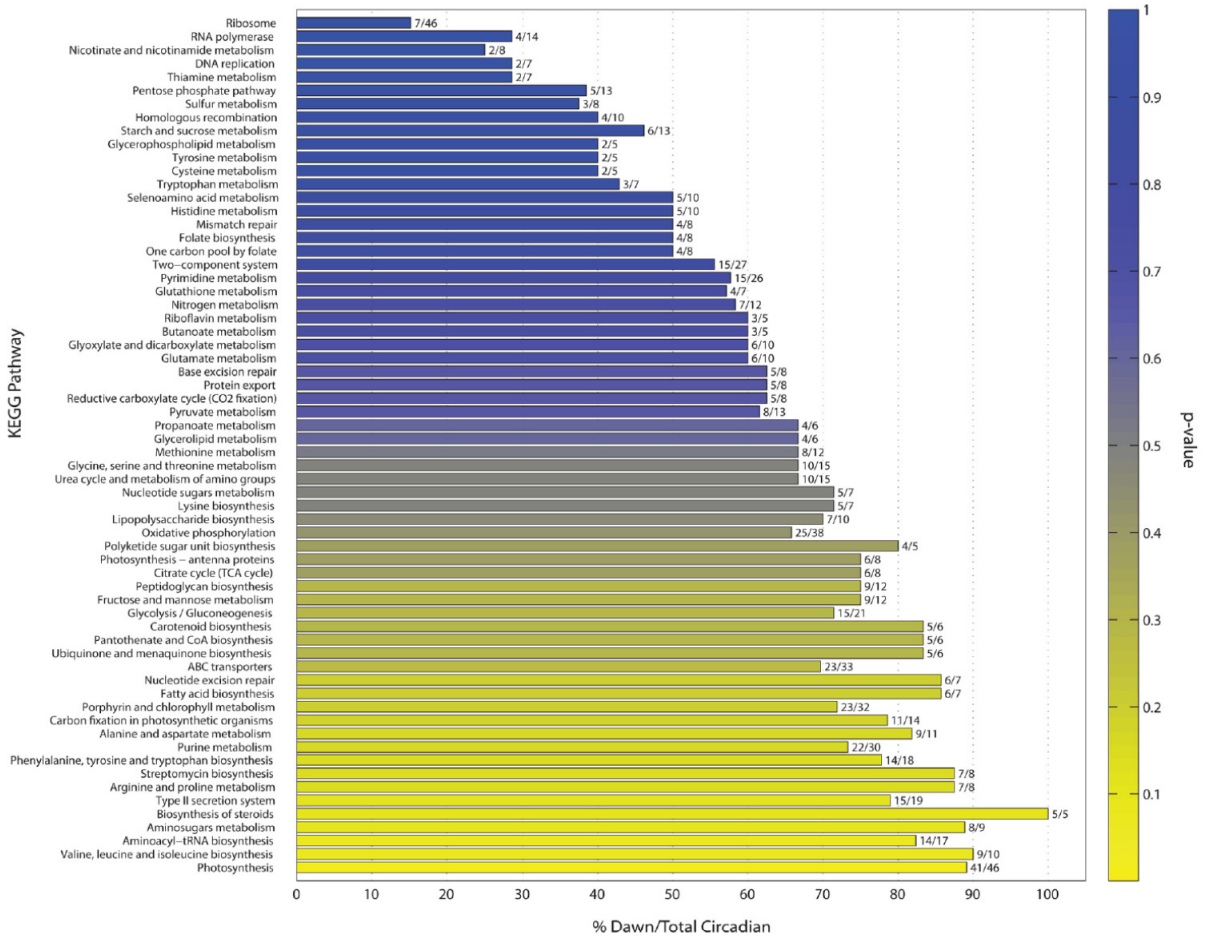


Figure S2.5: Functional significance of circadian gene expression.

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-4) and ribosomal protein genes (p = 3.4e-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. P-values were calculated using the cumulative hypergeometric distribution.

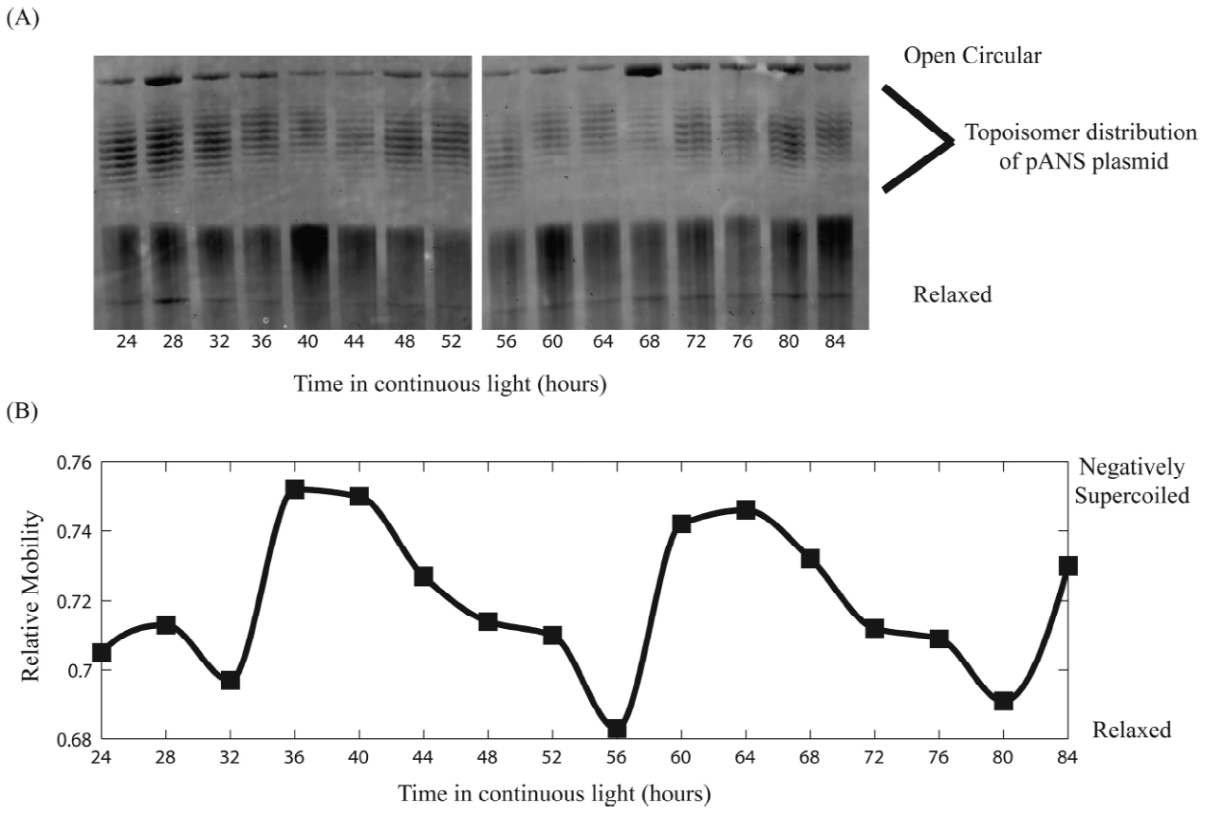
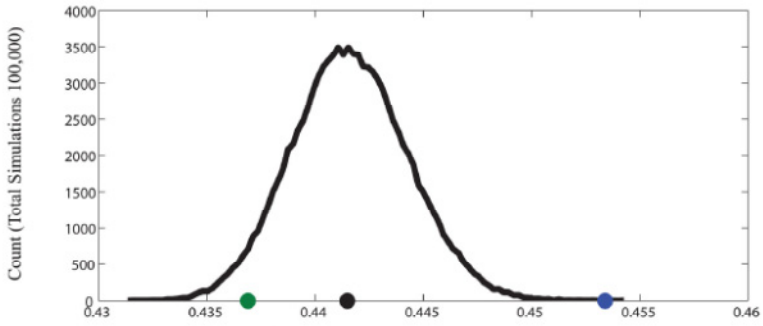


Figure S2.6: Chloroquine gel electrophoresis (CAGE) for determination of endogenous plasmid superhelicity during the circadian cycle.

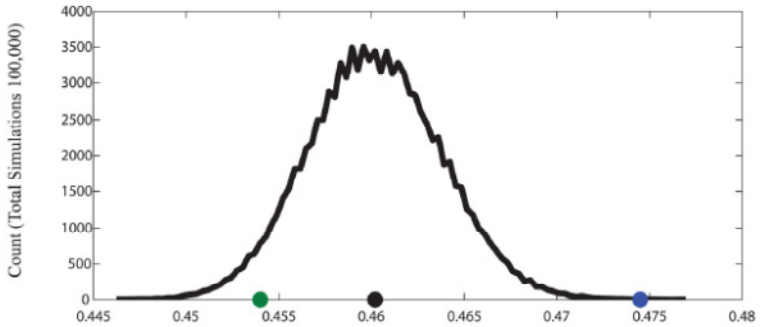
(A) SybrGold (Invitrogen) stained CAGE gel identifying distribution of topoisomers during the circadian cycle.

(B) Quantification of supercoiling from (A) as described in Materials and Methods of Chapter 2.

(A)



(B)



(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	

Figure S2.7: Statistical significance of increased and decreased AT content in monotonically relaxation activated and monotonically relaxation repressed genes, respectively.

(A) 250 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 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 Figure 2.1B.

(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.

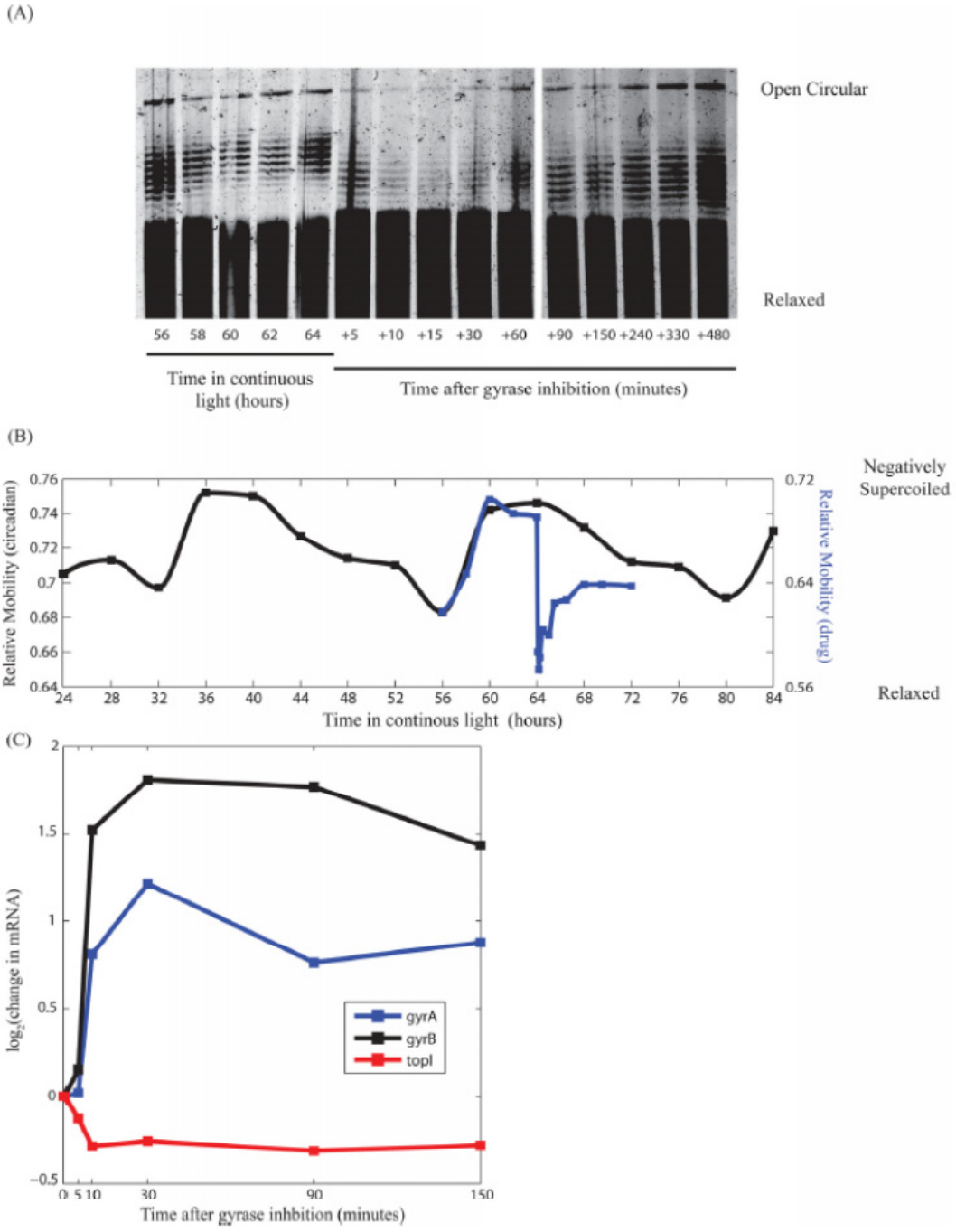


Figure S2.8: Chloroquine gel electrophoresis (CAGE) for determination of endogenous plasmid superhelicity after novobiocin induced relaxation.

(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.

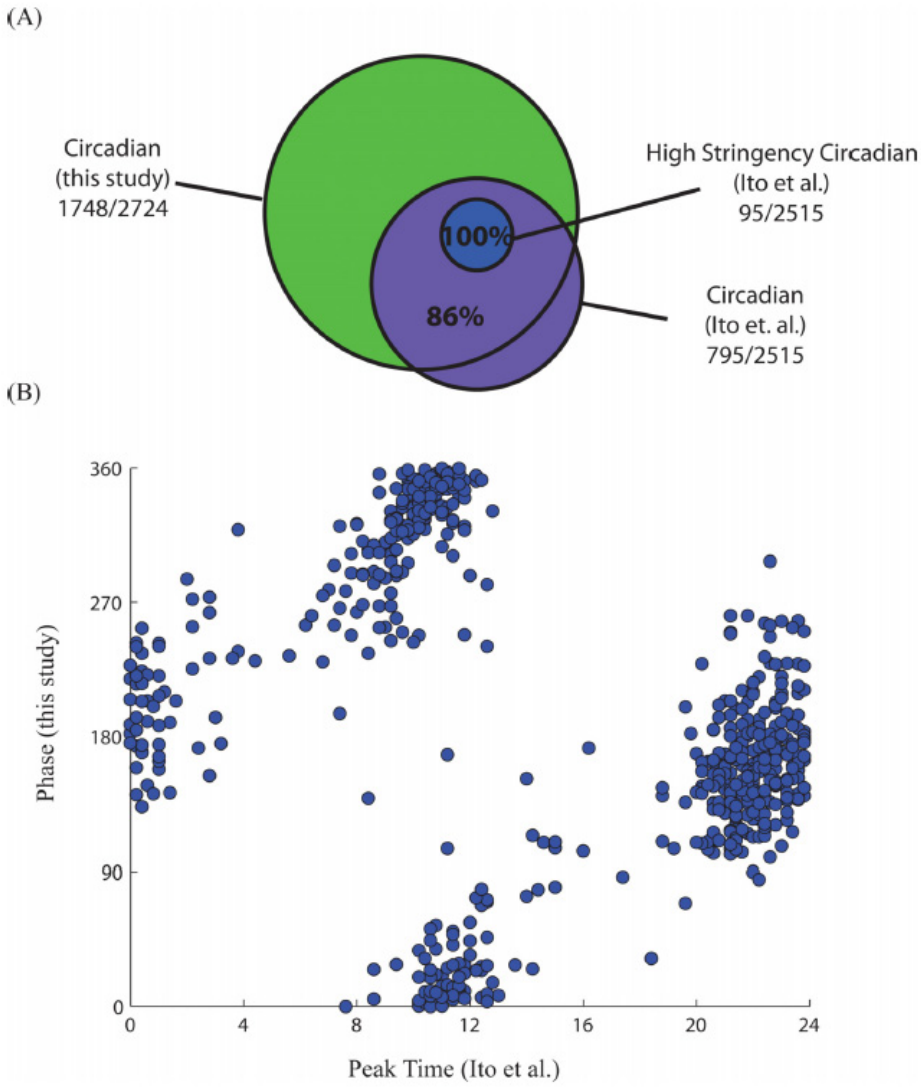


Figure S2.9: Comparison of circadian gene expression to Ito et al. 2009.

(A) Comparison of circadian genes found in this study versus Ito et al. 2009. 86% and 100% of the genes indentified by Ito et al. 2009 as low and high stringency circadian genes, respectively, were classified as circadian in this study. Although most circadian genes identified in Ito et al. 2009 are identified in this study, we additionally identify 955 circadian genes. Some of these genes are predicted ORFs not present in the Ito et al. 2009 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) versus peak time (Ito et al., 2009) in all genes identified as circadian in both studies. Genes oscillate with similar phase in both studies.

References

Alm EJ et al. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res* 17:1015-1022.

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.

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

APPENDIX B

Chapter 3 Supplemental Data

Figure S3.1: Examples of 5’ determination from RNA Sequencing.

(A) 5’ Determination of the *ntcA* transcript. A sharp drop in RNA sequencing reads is observed at the 5’ end of the mRNA. 5’ end determination by RNA sequencing and traditional methods (Luque et al., 1994) differ only by a single nucleotide.

(B) 5’ determination of the *purF* transcript. The RNA sequencing estimate is over 80 nucleotides different from that derived by traditional methods (Liu et al., 1996). Subsequent experiments (Min et al., 2004) have shown that the minimal promoter for the *purF* transcript contains the RNA sequencing 5’ end but not the literature 5’ end. We believe that RNA sequencing based 5’ end determination may be more robust and reliable than traditional methods. A more complete comparison of RNA sequencing and traditional transcription start determination is provided in Table S4 of the associated publication (Vijayan et al., 2011).

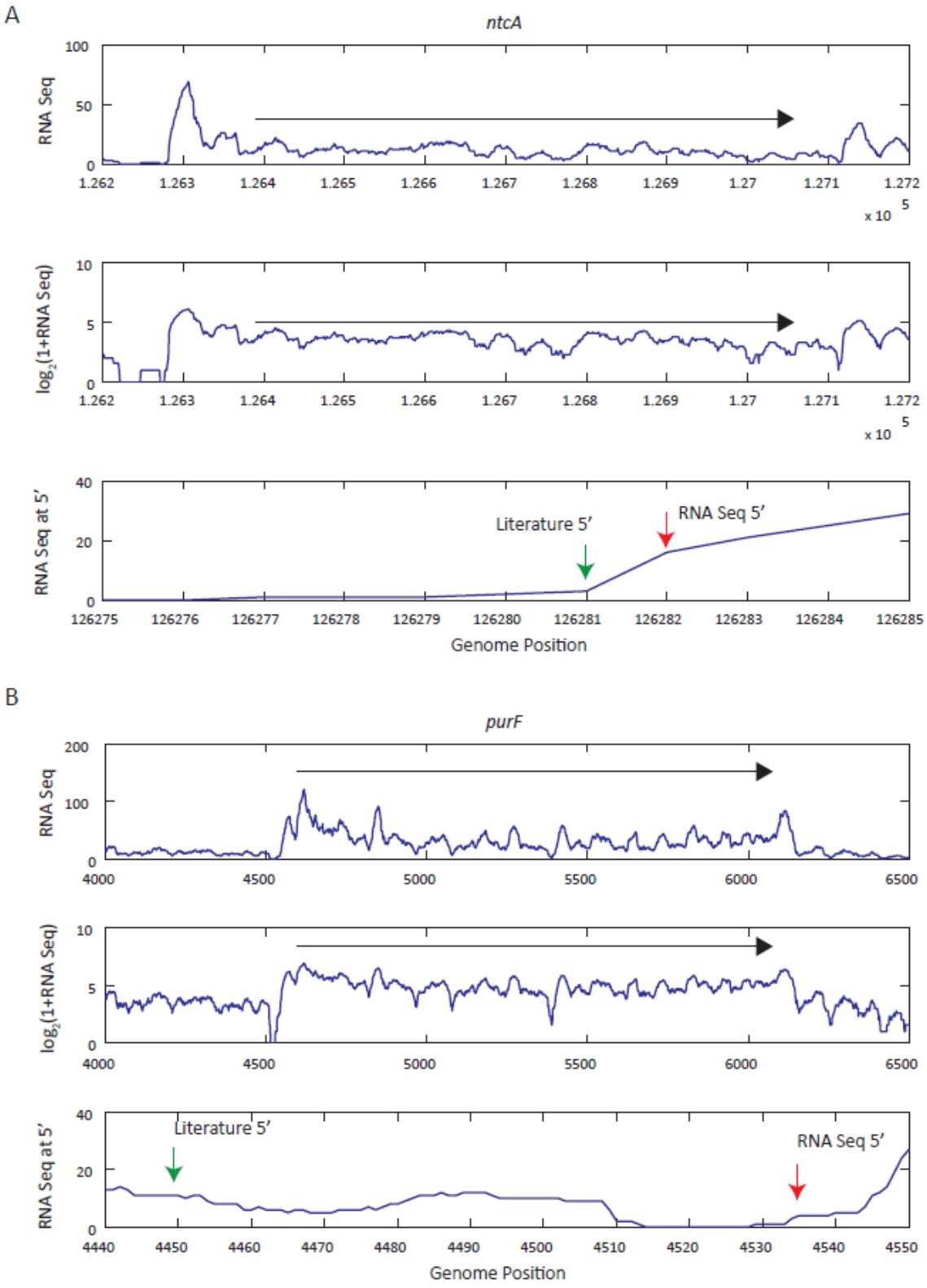


Figure S3.1 Continued

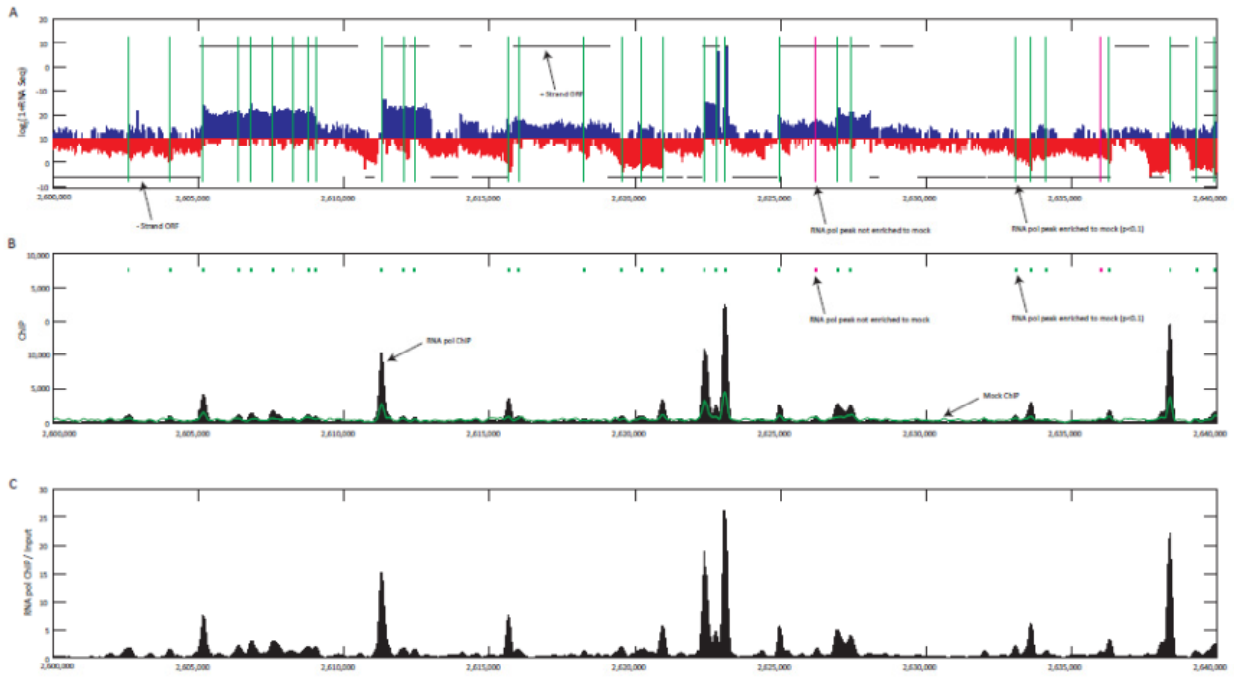


Figure S3.2: Representative RNA pol ChIP over a 40 kb region.

(A) RNA sequencing data. Positive strand transcription is shown in blue (positive y-axis), and negative strand transcription in red (negative y-axis). ORFs on the positive and negative strands are indicated by horizontal black lines. RNA polymerase (RNA pol) peaks significantly enriched over the mock IP ($p < 0.1$) are indicated with vertical green lines and those that are not ($p >= 0.1$) are indicated with vertical pink lines. Large RNA pol peaks tend to be located near the 5' end of transcripts although there are many peaks in the middle of transcripts potentially caused by RNA pol pausing.

(B) RNA pol ChIP and Mock. RNA pol ChIP (black) and mock IP (green) are normalized such that the genome average is 200 reads per nucleotide. Almost all RNA pol peaks are enriched over the mock IP. A complete listing of RNA pol peaks and their enrichment is provided in Table S3 of the associated publication (Vijayan et al., 2011).

(C) RNA pol ChIP normalized by input. Normalization of RNA pol ChIP by input does not qualitatively change the data (compare Figure S3.2B and Figure S3.2C).

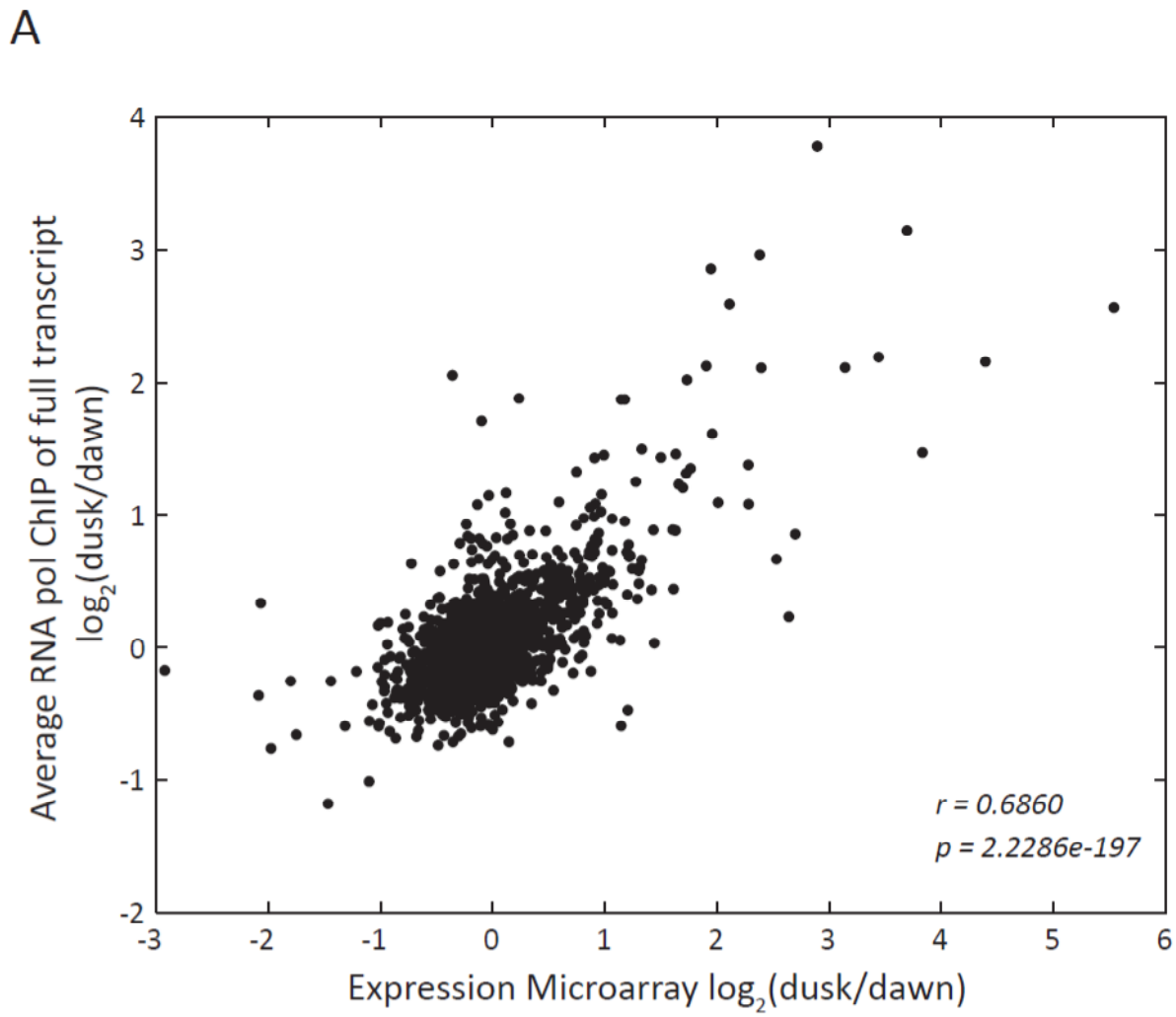


Figure S3.3: Comparison of changes in gene expression and RNA pol ChIP at two points in the circadian cycle.

(A) Changes in RNA pol occupancy at two separate times during the circadian cycle (dusk and dawn). Changes in RNA pol are reflective of changes in transcript level by microarray (Pearson correlation, r = 0.6860). Probability of getting a correlation as large by random chance (p-value) is 2.2286e-197.

Figure S3.4: Characteristics of transcription start.

(A) Melting temperature at transcription start. The melting temperature of 10 nucleotide fragments from -200 to +200 of all mRNAs was averaged (Materials and Methods of Chapter 3). A drop in the melting temperature is observed at the promoter.

(B) Nucleotide content at transcription start sites. Nucleotide content of all mRNAs aligned by transcription start. (C) Zoomed in nucleotide content at transcription start. Nucleotide content of all mRNAs aligned by transcription start. Preference for adenine at the +1 position and a -10 element can be observed.

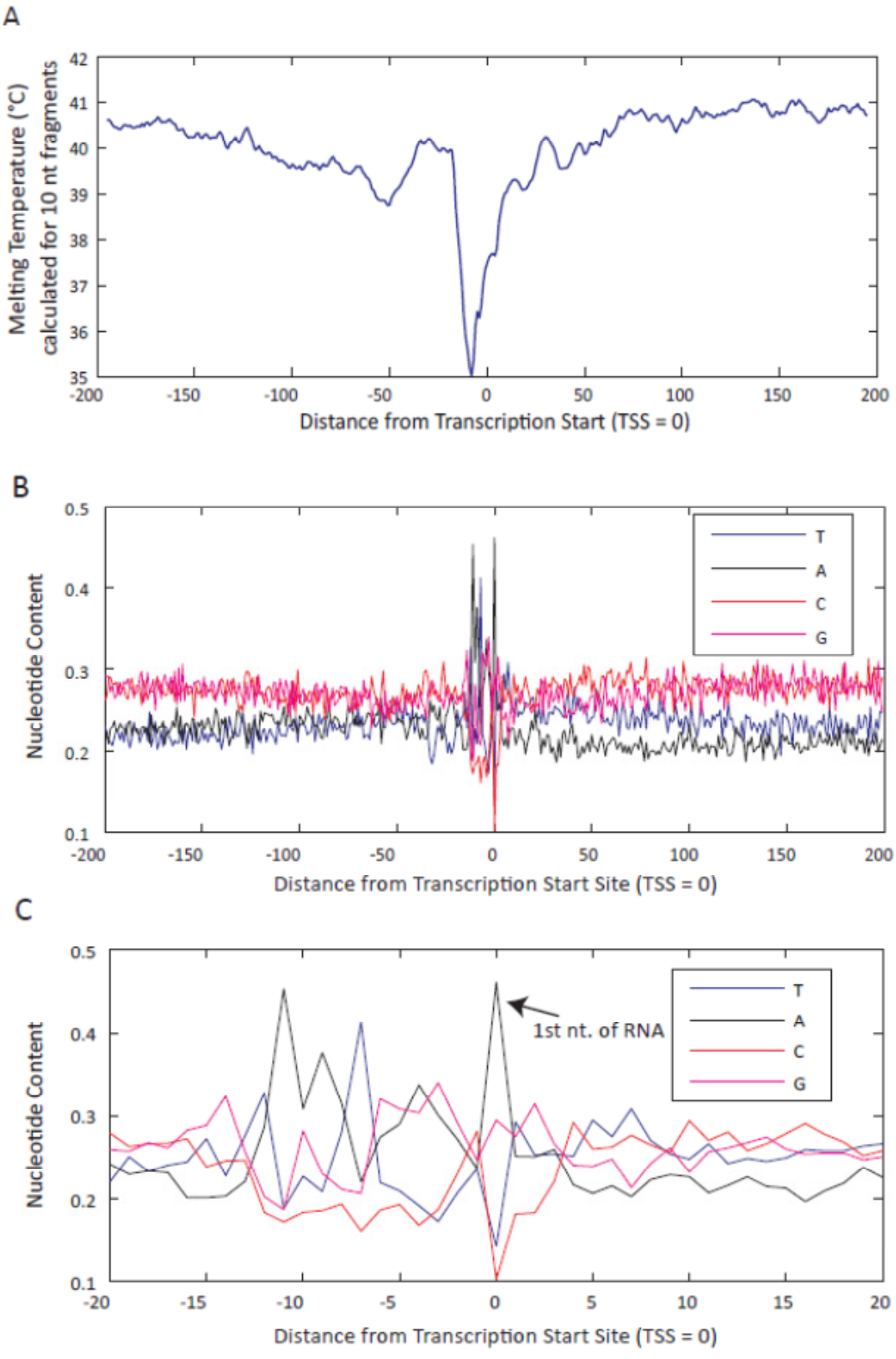


Figure S3.4 Continued

Figure S3.5: Comparison of minimum free energy changes with that of dinucleotide shuffled sequences.

(A) Minimum free energy change at RNA pol peaks. The minimum free energy of 60 nucleotide RNA fragments with 10 nucleotide spacing was calculated and averaged for all mRNAs (Materials and Methods of Chapter 3). A drop in minimum free energy slightly prior to the position of the RNA pol peak is observed. To prevent sequence features of the transcription terminus or promoters from interfering with this analysis, only a subset of 183 RNA pol peaks satisfying the following criteria were used: (1) RNA pol peak must be closer to a 5' end than 3' end; and (2) RNA pol peak must be +100 to +300 relative to the 5' end. Since RNA pol ChIP does not specify the strand being transcribed, the strand of transcription was inferred from RNA sequencing data. Dinucleotide shuffled sequences show a qualitatively similar trend to native sequences, suggesting that there is no specific secondary structure at this transition (Materials and Methods).

(B) Sequence changes near RNA pol peaks. A sequence content change from low to high GC content can be observed near the position of the RNA pol peaks. The same subset of RNA pol peaks are used here as in S5A. A smoothing window of 5 nucleotides has been applied to smooth nucleotide contents. These sequence changes may be responsible for the free energy changes we observe. It is also possible that these changes in sequence content may contribute to RNA pol pausing by an unknown mechanism.

(C) Minimum free energy change at transcription terminus. Minimum free energy was calculated as above after aligning all transcripts by transcription terminus. Dinucleotide shuffled sequences do not resemble native sequences, suggesting that a discrete hairpin-like structure exists at the terminus of transcripts (Materials and Methods of Chapter 3).

(D) Minimum free energy change at transcription start. Minimum free energy was calculated as above after aligning all transcripts by 5' transcription start. A drop in minimum free energy occurs globally within transcripts and may be related to our observation of global RNA pol pausing. Dinucleotide shuffled sequences show a qualitatively similar trend to native sequences (Materials and Methods of Chapter 3).

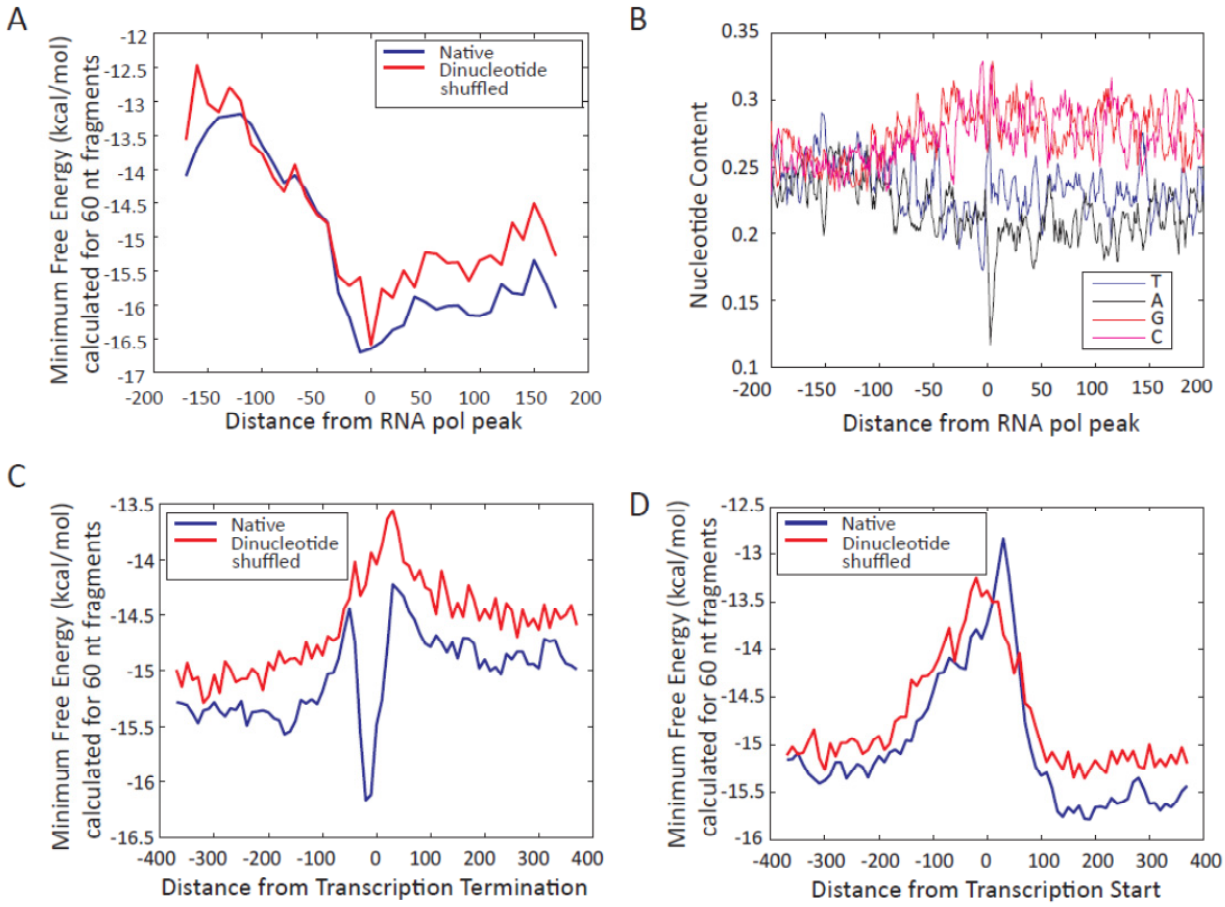


Figure S3.5 Continued

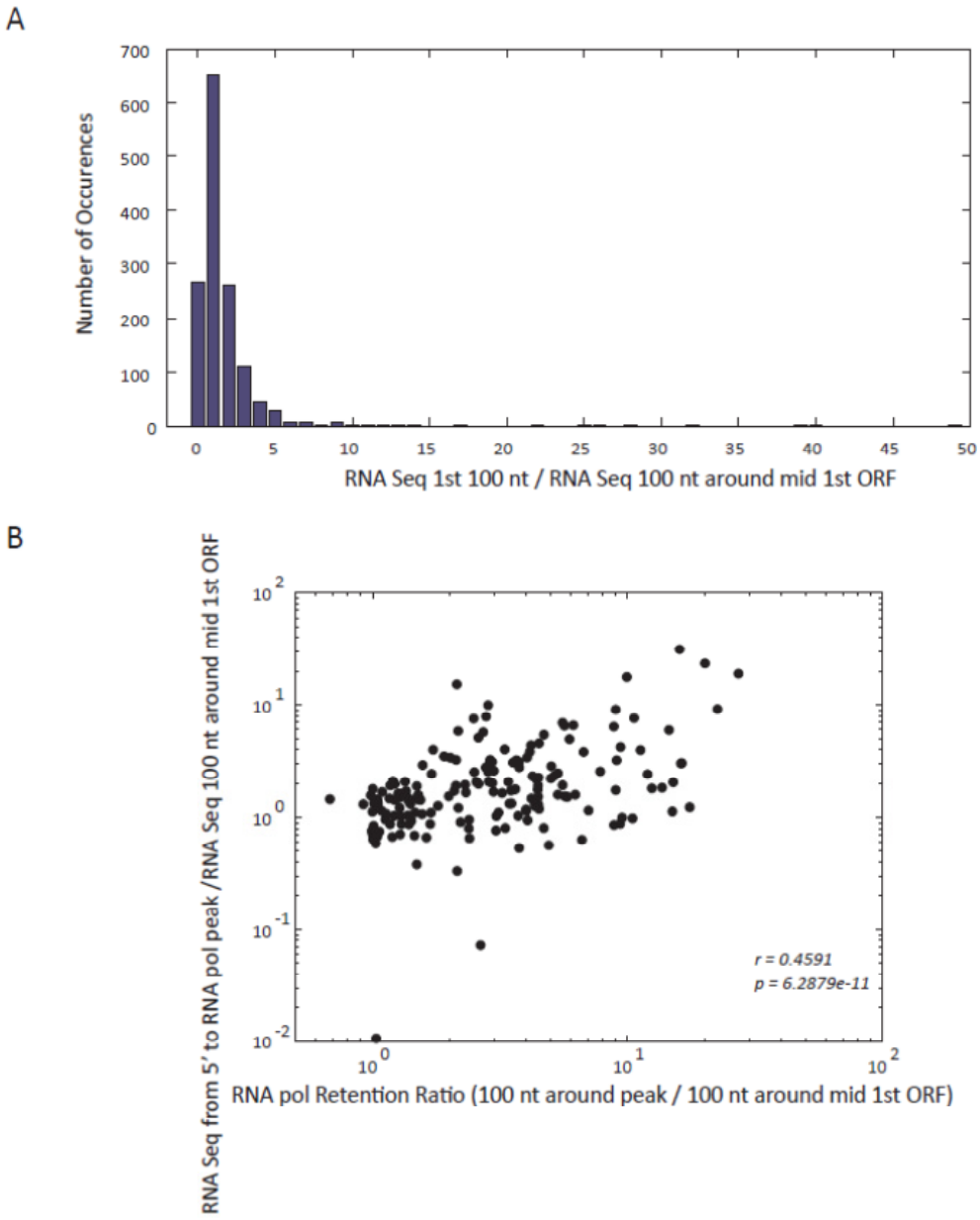


Figure S3.6: Enrichment in RNA sequencing at 5'.

(A) Increased RNA sequencing signal at 5' ends. An increase in RNA sequencing signal can be observed at the 5' end of mRNAs. Several biological phenomena may account for this enrichment, but one intriguing possibility is the existence of many partial or nascent transcripts caused by pausing of RNA pol near the 5' end of the transcript.

(B) RNA pol pausing at 5' ends may contribute to RNA sequencing enrichment at 5' ends. A slight but significant correlation exists between the retention ratio of RNA pol and the enrichment of RNA sequencing prior to the RNA pol peak. The same subset of RNA pol peaks were used as in Figure S3.5A. Pearson correlation is r = 0.4591, and probability of getting a correlation as large by random chance (p-value) is 6.2879e-11.

A

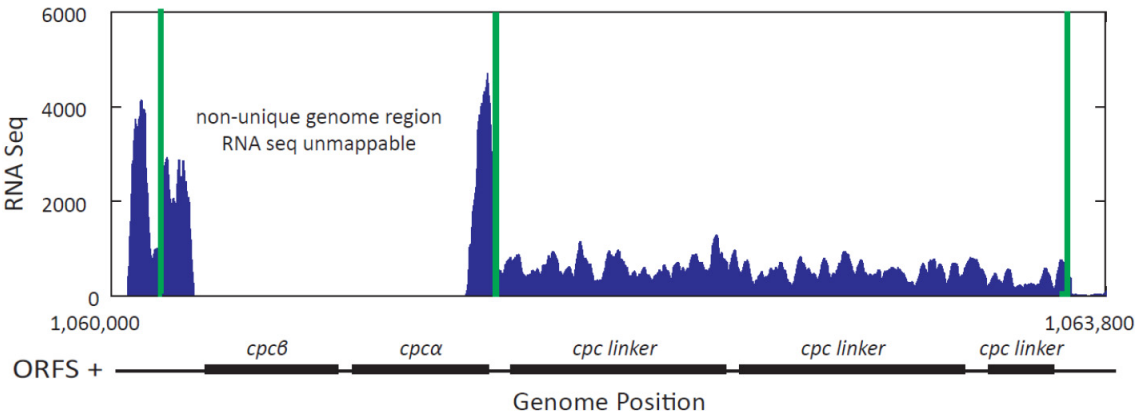


Figure S3.7: The phycocyanin operon – A functional case of partial transcription termination.

Partial transcription termination controls the stochiometry of *cpcβ* and *cpcα* to rod linker mRNA at ~6:1. This stochiometry reflects the organization of the phycobilisome – a hexameric α-β double disc with an associated linker (Belknap et al., 1987). RNA sequencing data cannot be mapped to the *cpcβ* and *cpcα* coding region because it is not unique in the genome (another copy of *cpcβ* and *cpcα,* corresponding to the core proximal phycobilisomes exists in the genome). Position of predicted terminators (from TransTermHP) is indicated in green, and the position of JGI predicted ORFs is indicated in black.

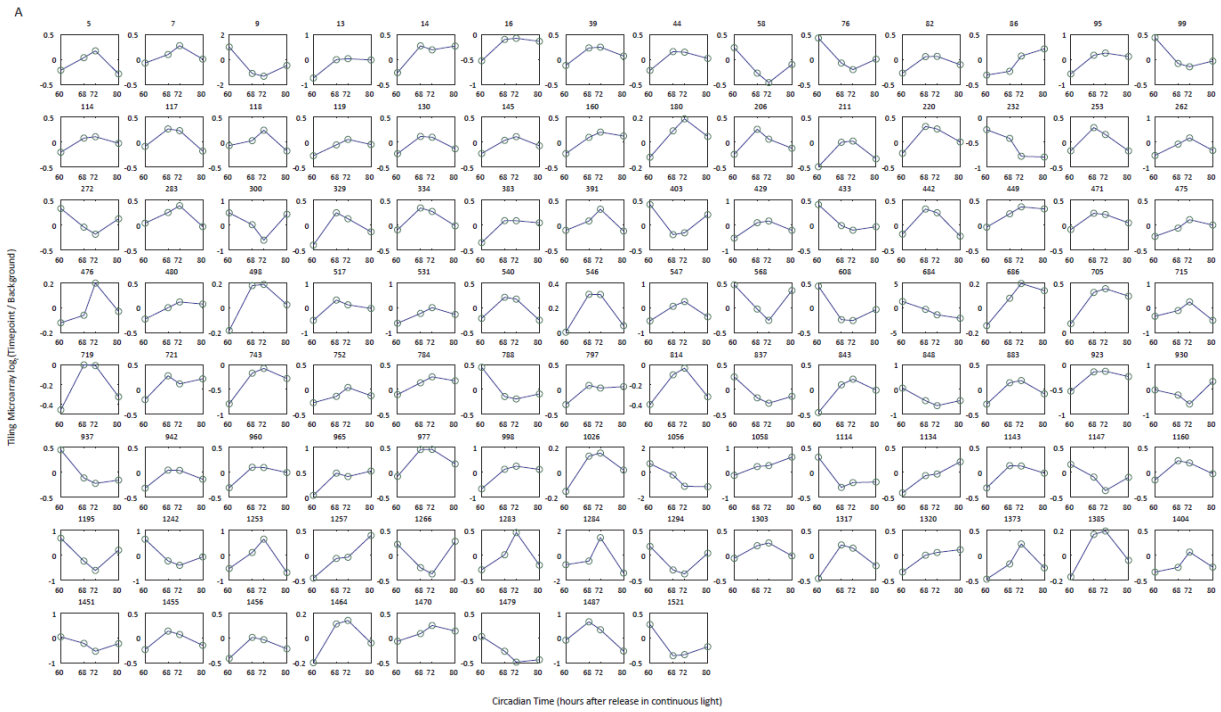


Figure S3.8: Circadian gene expression of putative non-coding RNAs.

(A) Gene expression by tiling microarray of high-confidence circadian non-coding RNAs. Gene expression of non-coding RNAs with potential for circadian gene expression are plotted by non-coding transcript ID in Table S2 of the associated publication (Vijayan et al., 2011). Gene expression ratios for non-coding RNAs are computed by averaging the gene expression ratios for all tiling probes internal to the non-coding transcript.

References

Belknap WR, Haselkorn R (1987) Cloning and light regulation of expression of the phycocyanin operon of the cyanobacterium *Anabaena. EMBO J* 6:871-884

Liu Y, Tsinoremas NF, Golden SS, Kondo T, Johnson CH (1996) Circadian expression of genes involved in the purine biosynthetic pathway of the cyanobacterium *Synechoccoccus* sp. Strain PCC 7942. *Mol Microbiol* 20:1071-1081.

Luque I, Flores E, Herrero A (1994) Molecular mechanism for the operation of nitrogen control in cyanobacteria. *EMBO J* 13: 2862-2869.

Min H, Liu Y, Johnson CH, Golden SS (2004) Phase determination of circadian gene expression in Synechococcus elongatus PCC 7942. *J Biol Rhythms* 19:103–112.

Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47.

APPENDIX C

Chapter 4 Supplemental Data

Figure S4.1: Temporal dynamics of mRNA measured by quantitative PCR.

(A) *luxAB* mRNA time-course when driven by the P1 fragment is in the class II phase. Endogenous *purF* and *kaiBC* mRNA time-courses from the same cells are shown as controls. mRNA dynamics from the P1 promoter are identical to the endogenous *purF* suggesting that the minimal P1 fragment does indeed capture the mRNA dynamics of its full length promoter. *kaiBC* mRNA time-course is shown as an example of the opposite (class I) phase of expression.

(B) *luxAB* mRNA abundance in M1-7 is altered from that of the P1 fragment.

(C) *luxAB* mRNA abundance in M1-8 is altered from that of the P1 fragment.

(D) *luxAB* mRNA abundance in M2-8 is altered from that of the P1 fragment.

(E) *luxAB* mRNA abundance in M2-9 is altered from that of the P1 fragment.

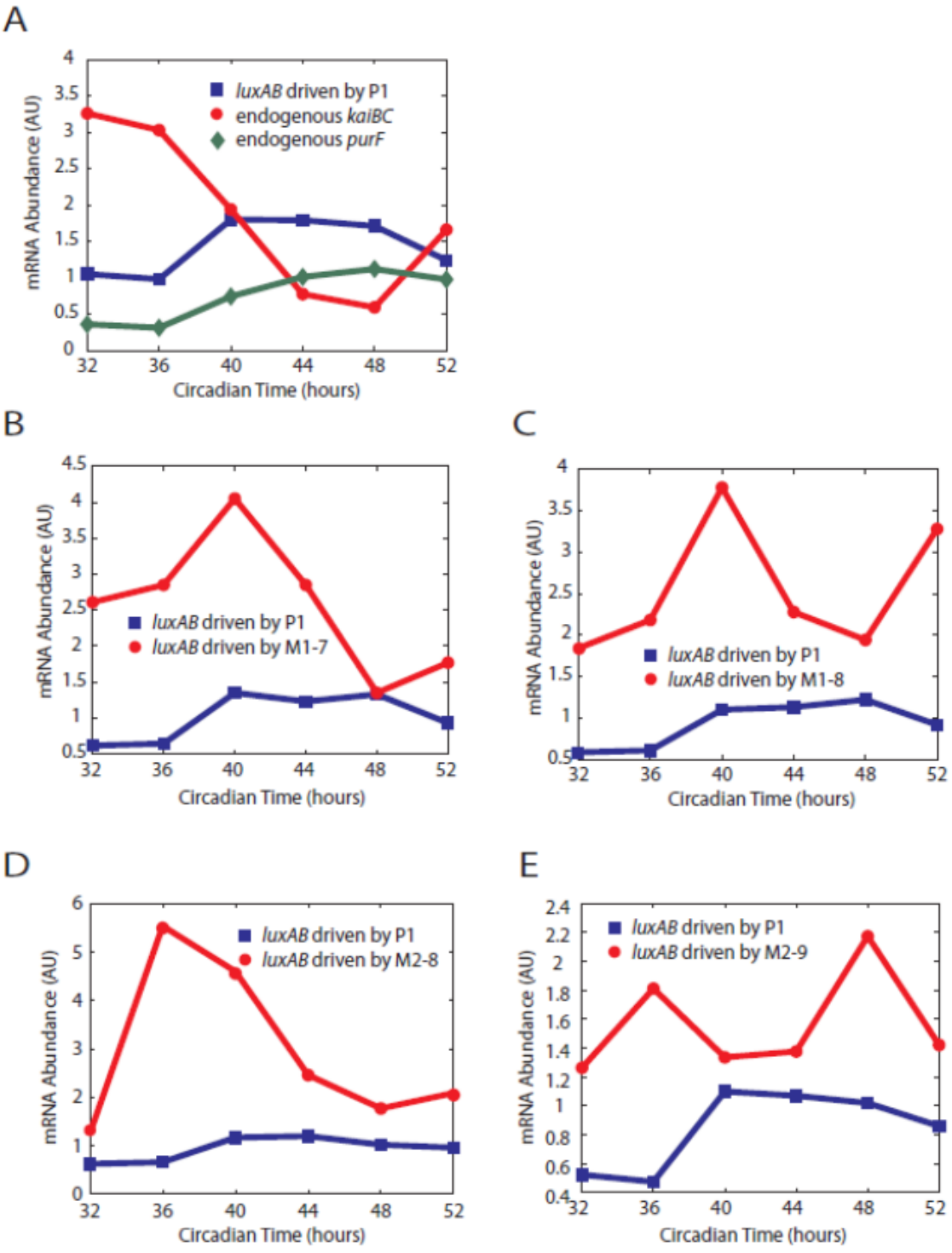


Figure S4.1 Continued

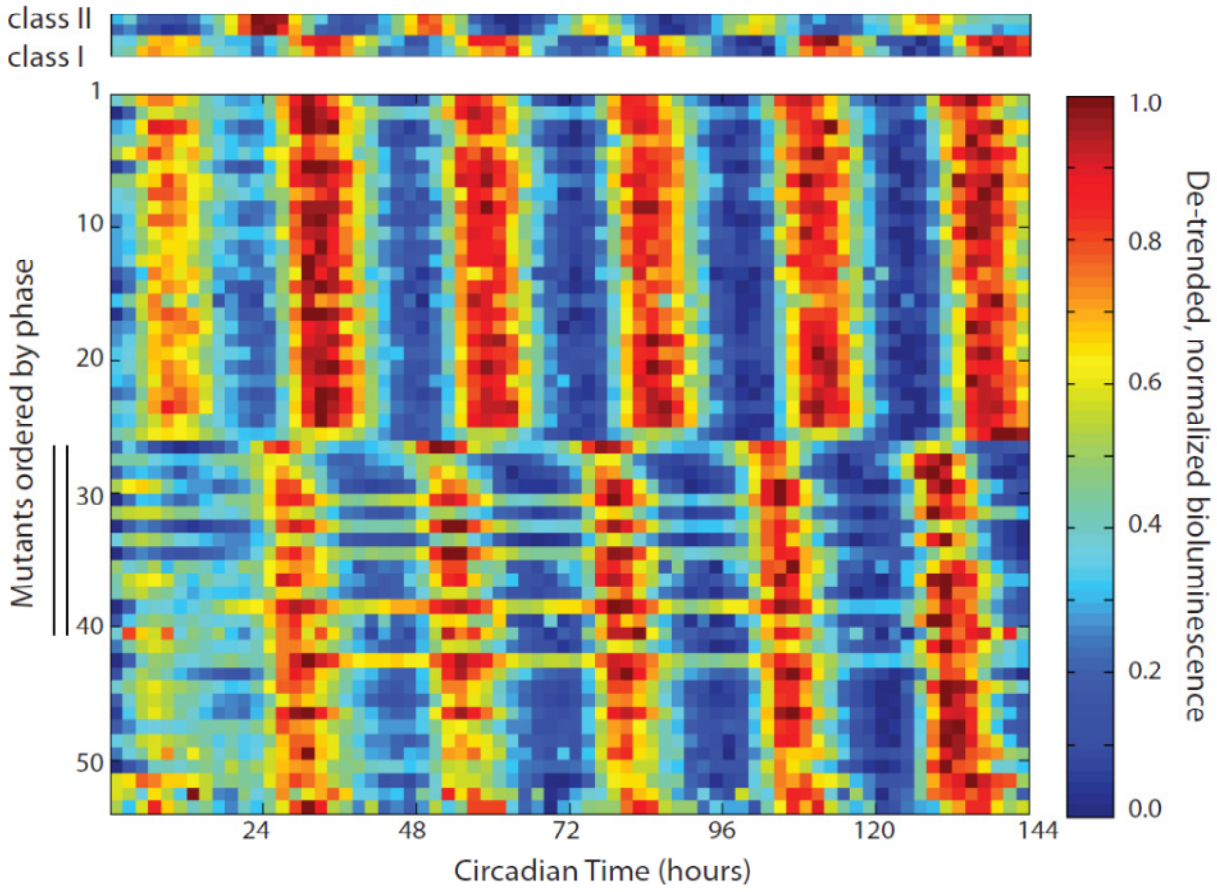


Figure S4.2: Random mutagenesis of the class I P3 promoter fragment.

Random mutagenesis of the promoter fragment P3 from the class I gene, *synpcc7942__2046*, yields class II mutants. Top panel: bioluminescence from two biological replicates of P1 and P3 are shown as class II and class I controls, respectively. Bottom panel: bioluminescence from all mutant clones with gene expression ordered by phase. Mutants with phase change are marked with the double lines on the y-axis. All bioluminescence traces have been linearly de-trended and normalized such that the minimum and maximum bioluminescence are 0 and 1 units, respectively.

APPENDIX D

Chapter 5 Supplemental Data

Table S5.1: Table of plasmids

Plasmid	Description	Resistance	Reference
pAM1303	Neutral site 1 integration vector	Sp/Sm	Mackey et al., 2007
pAM1579	Neutral site 2.1 integration vector	Kan (Amp)	Mackey et al., 2007
pAM1573	Neutral site 2.1 integration vector	Cm (Amp)	Mackey et al., 2007
EB2065	Neutral site 2.2 integration vector	Cm (Amp)	This work
EB2066	Site A integration vector (between chromosomal position 1595934 and 1595935 relative to Genbank CP000100)	Cm (Amp)	This work
EB2067	Site B integration vector (between chromosomal position 1758400 and 1758401 relative to Genbank CP000100)	Cm (Amp)	This work
eBB110	Source of 120 Tet operator repeats	Amp	Marquis et al., 2008
pLAU43	Source of 120 Lac operator repeats	Amp	Lau et al., 2003
EB2068	120 Tet operators PCRed using primers pBBSalI and pBBXbaI from eBB110 cloned between *SalI* and *XbaI* of pAM1579	Kan (Amp)	This work
EB2069	120 Lac operators from *XbaI, SmaI* fragment of pLAU43 cloned between *XbaI* and *SmaI* of pAM1573	Cm (Amp)	This work
EB2070	120 Lac operators from *SmaI*, *SalI* fragment of pLAU43 cloned between *SmaI* and *XhoI* of EB2066	Cm (Amp)	This work
EB2071	120 Lac operators from *SmaI*, *SalI* fragment of pLAU43 cloned between *SmaI* and *XhoI* of EB2067	Cm (Amp)	This work
pJRC23	Source of ECFP	Sp/Sm	Chabot et al., 2007
EB2072	*PkaiBC::tetR::ECFP PkaiBC::EYFP::lacI* cloned between *NotI* and *SacI* of pAM1303	Sp/Sm	This work
EB2073	*PkaiBC::tetR::EYFP* cloned between *NotI* and *SacI* of pAM1303	Sp/Sm	This work
EB2074	*PapcA::rbcL::ECFP* cloned between *SmaI* and *XhoI* of EB2065	Cm (Amp)	This work
pAM2055	Source of gentamycin cassette for deletion vectors	Gm (Amp)	Mackey et al .,2007
EB2075	*mreB (Synpcc7942_0300*) deletion vector with gentamycin cassette inserted between *BamHI* and *SacI* of pUC18	Gm (Amp)	This work
EB2076	*ftsZ (Synpcc7942_2378*) deletion vector with gentamycin cassette inserted between *HindIII* and *SacI* of pUC18	Gm (Amp)	This work
EB2077	*parA1 (Synpcc7942_0220*) deletion vector with gentamycin cassette inserted between *BamHI* and *SacI* of pUC18	Gm (Amp)	This work
EB2078	*minD (Synpcc7942_0896*) deletion vector with gentamycin cassette inserted between *BamHI* and *SacI* of pUC18	Gm (Amp)	This work
EB2079	*parA2 (Synpcc7942_1833*) deletion vector with gentamycin cassette inserted between *BamHI* and *SacI* of pUC18	Gm (Amp)	This work

*Markers shown in parenthesis are additional markers used for selection of plasmids in *E. coli.*

Table S5.2: Table of *S. elongatus* strains

Strain	Description	Plasmids used to create this strain from WT	Resistance	Reference
EOC200	TetO (NS 2.1)	EB2068	Kan	This work
EOC201	TetO (NS 2.1), LacO (A)	EB2068, EB2070	Kan, Cm	This work
EOC202	TetO (NS 2.1), LacO (B)	EB2068, EB2071	Kan, Cm	This work
EOC203	TetO (NS 2.1), *tetR::EYFP* (NS 1)	EB2068, EB2073	Kan, Sp/Sm	This work
EOC204	TetO (NS 2.1), LacO (A), *tetR::ECFP EYFP::lacI* (NS 1)	EB2068, EB2070, EB2072	Kan, Cm, Sp/Sm	This work
EOC205	TetO (NS 2.1), LacO (B) , *tetR::ECFP EYFP::lacI* (NS 1)	EB2068, EB2071, EB2072	Kan, Cm, Sp/Sm	This work
EOC206	TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcl::ECFP* (NS 2.2)	EB2068, EB2073, EB2074	Kan, Cm, Sp/Sm	This work
EOC207	*ΔmreB* incomplete segregation	EB2075	Gm	This work
EOC208	*ΔftsZ* incomplete segregation	EB2076	Gm	This work
EOC209	*ΔparA1*	EB2077	Gm	This work
EOC210	*ΔminD*	EB2078	Gm	This work
EOC211	*ΔparA2*	EB2079	Gm	This work
EOC212	*ΔmreB,* TetO (NS 2.1)	EB2075, EB2068	Gm, Kan	This work
EOC213	*ΔftsZ,* TetO (NS 2.1)	EB2076, EB2068	Gm, Kan	This work
EOC214	*ΔparA1,* TetO (NS 2.1)	EB2077, EB2068	Gm, Kan	This work
EOC215	*ΔminD,* TetO (NS 2.1)	EB2078, EB2068	Gm, Kan	This work
EOC216	*ΔparA2,* TetO (NS 2.1)	EB2079, EB2068	Gm, Kan	This work
EOC217	*ΔmreB* incomplete segregation, TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcL::ECFP* (NS 2.2)	EB2075, EB2068, EB2073, EB2074	Gm, Kan, Sp/Sm, Cm	This work
EOC218	*ΔftsZ* incomplete segregation, TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcL::ECFP* (NS 2.2)	EB2076, EB2068, EB2073, EB2074	Gm, Kan, Sp/Sm, Cm	This work
EOC219	*ΔparA1,* TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcL::ECFP* (NS 2.2)	EB2077, EB2068, EB2073, EB2074	Gm, Kan, Sp/Sm, Cm	This work
EOC220	*ΔminD*, TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcL::ECFP* (NS 2.2)	EB2078, EB2068, EB2073, EB2074	Gm, Kan, Sp/Sm, Cm	This work
EOC221	*ΔparA2,* TetO (NS 2.1), *tetR::EYFP* (NS 1), *rbcL::ECFP* (NS 2.2)	EB2079, EB2068, EB2073, EB2074	Gm, Kan, Sp/Sm, Cm	This work

Table S5.3: Table of primers

Primer	Sequence
Primers for site A (EB2066), B (EB2067), and NS 2.2 (EB2065) integration vectors	
MCS_F	5' GTGCTTCTGGCTATAGCTGACTCGTCACGGGTAACCGA 3'
MCS_R	5' CCGCGCTGGGGTTACCAG 3'
A-Up-F	5' GACCACACCCGTCCTGTGGATCCTGCAACTATCCCTCGATC 3'
A-Up-R	5' AGTCAGCTATAGCCAGAAGCACTGGGAGGCATTAGAAGC 3'
A-Dn-F	5' CTGGTAACCCCAGCGCGGATGCGGATTGAAATTGTTC 3'
A-Dn-R	5' TGCGTCCGGCGTAGAGGATCCGTTCTCGACCACCCATTCAG 3'
B-Up-F	5' GACCACACCCGTCCTGTGGATCCGTTTGTGTTAGGCCACGAT 3'
B-Up-R	5' AGTCAGCTATAGCCAGAAGCACCACGCTAACGGATGAGCG 3'
B-Dn-F	5' CTGGTAACCCCAGCGCGGGACTGTCATGAATTACCTTCAG 3'
B-Dn-R	5' TGCGTCCGGCGTAGAGGATCCAGTAAGCAACGTGGCC 3'
NS22-Up-F	5' GACCACACCCGTCCTGTGCAGGCTCAGTGTGGTTCG 3'
NS22-Up-R	5' AGTCAGCTATAGCCAGAAGCAC 3'
NS22-Dn-F	5' CTGGTAACCCCAGCGCGGAGATCTCGCAGCGTAAAGCCGT 3'
NS22-Dn-R	5' TGCGTCCGGCGTAGAGGATCCGATCGCCATCCACGCAC 3'
Primers for tet operator vector (EB2068)	
pBBSalI	5' TAAACTATGTCGACCTTTCTTATCTTGATAATAAGGGTAAC 3'
pBBXbaI	5'ATAGTTTATCTAGACCGTCCTTGAACATGACT 3'
Primers for PapcA::RbcL::ECFP vector (EB2074)	
SmaI_ApcA	5' CATGCCCGGGTACGAGCGCTATATCACCCC 3'
Apca_RBS	5' TAATGGATTCCTCCAAGACTAGATTTGAAACCAGACTGGCCTCCACC 3'
RBS_RbcL	5' TCTAGTCTTGGAGGAATCCATTAATGCCCAAGACGCAATCTG 3'
RbcL_linker	5' ACTAGAACCAGAACTACCACTAGAGAGCTTGTCCATCGTTTCGAATTC 3'
Linker_ECFP	5' TCTAGTGGTAGTTCTGGTTCTAGTATGGTGAGCAAGGGCGAGGAG 3'
ECFP_XhoI	5' CATGCTCGAGTTACTTGTACAGCTCGTCCATGC 3'
Primers for deletion vectors (EB2075 to EB2079)	
Gent_F	5' GACGCACACCGTGGA 3'
Gent_R	5' GCGGCGTTGTGACAA 3'
1833KO-Up-F	5' GATCGATCGGATCCACAAAAAGGGGGCTGGTTAG 3'
1833KO-Up-R	5' GTTTCCACGGTGTGCGTCACTAGT CGACAACCTCCCAAAGCG 3'
1833KO-Dn-F	5' AAATTGTCACAACGCCGCTGACTGACGCCTTTGACC 3'
1833KO-Dn-R_Sac	5' GATCGATCGAGCTCCAAACAAAATGCCCCAAAGT 3'
0300KO-Up-F	5' GATCGATCGGATCCGGTAGCCGAATCACTCCGA 3'
0300KO-Up-R	5' GTTTCCACGGTGTGCGTCACTAGTTCCGCCTTGATGACGTG 3'
0300KO-Dn-F	5' AAATTGTCACAACGCCGCAGCGGTGCGGAGCAG 3'
0300KO-Dn-R_Sac	5' GATCGATCGAGCTCAGAGTGTACTAGCGCTAGAAGTG 3'
0220KO-Up-F	5' GATCGATCGGATCCACCGAGATATGCTCGATTGC 3'
0220KO-Up-R	5' GTTTCCACGGTGTGCGTCACTAGTGGTGGATCTCAAGTTCAGGG 3'
0220KO-Dn-F	5' AAATTGTCACAACGCCGCCCATGACCCGCAAATCTG 3'
0220KO-Dn-R_Sac	5' GATCGATCGAGCTCGAACTGAGCTTTGGCTTGCT 3'
2378KO-Up-F_HindIII	5' GATCGATCAAGCTTAACGGCTGCAGCGCTT 3'
2378KO-Up-R	5' GTTTCCACGGTGTGCGTCACTAGTGGGGGTGAGTAGTAGCA 3'
2378KO-Dn-F	5' AAATTGTCACAACGCCGCTGGCTGTTCCGATCGCC 3'
2378KO-Dn-R_Sac	5' GATCGATCGAGCTCTCTGTCAATGGTCCCGCT 3'
0896KO-Up-F	5' GATCGATCGGATCCCGCCAGTAACCGTCACAGAT 3'
0896KO-Up-R	5' GTTTCCACGGTGTGCGTCACTAGTAGGGTCCGAAGAGCAGGAG 3'
0896KO-Dn-F	5' AAATTGTCACAACGCCGCGGGTCTTTGCAGCAATGCT 3'
0896KO-Dn-R_Sac	5' GATCGATCGAGCTCTGACCACCTAGCAAATGTTCC 3'

Table S5.4: Sequence of TetR and LacI fluorescent fusion protein constructs

EB2072: *PkaiBC*::*RBS*::*tetR*::*ECFP* *PkaiBC*::*RBS*::*EYFP*::*linker*::*lacI* cloned between *NotI* and *SacI* of pAM1303. Note that the beginning of *PkaiBC* contains the *kaiA* terminator which will serve to decouple TetR and LacI expression.

GCGGCCGCTCTAGACGCAAATTGCAATCTGCATTTGGGTGGAATCGACGGCAATTTTGCAGGATTGCCAGCGGGCGCTGTCGGCC
GATCGCTATCAACTCCAAGTCTGTGAGTCTGGCGAAATGCTCTTGGAGTATGCCCAAACCCATCGTGACCAAATCGACTGCCTGAT
TTTAGTGGCAGCCAATCCCAGCTTCAGGGCAGTTGTTCAGCAGCTCTGCTTTGAGGGAGTGGTGGTACCAGCGATTGTCGTAGGCG
ATCGCGACAGTGAGGATCCCGATGAACCAGCCAAAGAACAGCTCTATCACAGCGCTGAACTGCACCTCGGTATCCATCAGCTCGA
GCAATTGCCCTACCAAGTTGATGCTGCACTGGCTGAATTTCTGCGCTTAGCCCCGGTCGAGACCATGGCCGACCACATCATGCTGA
TGGGGGCCAACCACGATCCCGAGCTATCGAGCCAGCAGCGGGACCTCGCTCAGCGACTACAAGAGCGCCTAGGCTATCTCGGGG
TCTACTACAAGCGTGATCCCGATCGCTTTCTGCGCAACCTACCCGCCTACGAAAGCCAAAAGCTGCACCAAGCGATGCAGACTAG
CTATCGTGAAATCGTTTTGAGCTATTTTTCGCCGAATAGCAACCTCAACCAGAGCATTGACAACTTCGTCAACATGGCTTTCTTTGC
CGATGTTCCAGTCACCAAAGTGGTAGAAATTCACATGGAGCTGATGGACGAGTTTGCCAAGAAGCTCCGCGTAGAGGGACGTTCA
GAGGACATTTTGCTGGATTATCGGCTGACTTTAATTGATGTAATTGCACATCTTTGTGAGATGTATCGACGGTCTATCCCACGAGA
AACCTGAAAAGGTAAAGGAGGTCTTAAGCTCGGCTCAATTTCTCTCTTTATCCTGTTAGATGGTTTGATTGCTGTTGCTACCCCGTT
GATCTGCGTAGATCTTCTAAGGAGGAAAAAAAATGTCTAGATTAGATAAAAGTAAAGTGATTAACAGCGCATTAGAGCTGCTTAA
TGAGGTCGGAATCGAAGGTTTAACAACCCGTAAACTCGCCCAGAAGCTAGGTGTAGAGCAGCCTACATTGTATTGGCATGTAAAA
AATAAGCGGGCTTTGCTCGACGCCTTAGCCATTGAGATGTTAGATAGGCACCATACTCACTTTTGCCCTTTAGAAGGGGAAAGCTG
GCAAGATTTTTTACGTAATAACGCTAAAAGTTTTAGATGTGCTTTACTAAGTCATCGCGATGGAGCAAAAGTACATTTAGGTACAC
GGCCTACAGAAAAACAGTATGAAACTCTCGAAAATCAATTAGCCTTTTTATGCCAACAAGGTTTTTCACTAGAGAATGCATTATAT
GCACTCAGCGCTGTGGGGCATTTTACTTTAGGTTGCGTATTGGAAGATCAAGAGCATCAAGTCGCTAAAGAAGAAAGGGAAACAC
CTACTACTGATAGTATGCCGCCATTATTACGACAAGCTATCGAATTATTTGATCACCAAGGTGCAGAGCCAGCCTTCTTATTCGGC
CTTGAATTGATCATATGCGGATTAGAAAAACAACTTAAATGTGAAAGTGGGTCTATGGTGAGCAAGGGCGAGGAGCTGTTCACCG
GGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCA
CCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTG
GGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAG
GAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGC
ATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACATCAGCCACAACGTC
TATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTC
GCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCC
TGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACG
AGCTGTACAAGTAACGCAAATTGCAATCTGCATTTGGGTGGAATCGACGGCAATTTTGCAGGATTGCCAGCGGGCGCTGTCGGCC
GATCGCTATCAACTCCAAGTCTGTGAGTCTGGCGAAATGCTCTTGGAGTATGCCCAAACCCATCGTGACCAAATCGACTGCCTGAT
TTTAGTGGCAGCCAATCCCAGCTTCAGGGCAGTTGTTCAGCAGCTCTGCTTTGAGGGAGTGGTGGTACCAGCGATTGTCGTAGGCG
ATCGCGACAGTGAGGATCCCGATGAACCAGCCAAAGAACAGCTCTATCACAGCGCTGAACTGCACCTCGGTATCCATCAGCTCGA
GCAATTGCCCTACCAAGTTGATGCTGCACTGGCTGAATTTCTGCGCTTAGCCCCGGTCGAGACCATGGCCGACCACATCATGCTGA
TGGGGGCCAACCACGATCCCGAGCTATCGAGCCAGCAGCGGGACCTCGCTCAGCGACTACAAGAGCGCCTAGGCTATCTCGGGG
TCTACTACAAGCGTGATCCCGATCGCTTTCTGCGCAACCTACCCGCCTACGAAAGCCAAAAGCTGCACCAAGCGATGCAGACTAG
CTATCGTGAAATCGTTTTGAGCTATTTTTCGCCGAATAGCAACCTCAACCAGAGCATTGACAACTTCGTCAACATGGCTTTCTTTGC
CGATGTTCCAGTCACCAAAGTGGTAGAAATTCACATGGAGCTGATGGACGAGTTTGCCAAGAAGCTCCGCGTAGAGGGACGTTCA
GAGGACATTTTGCTGGATTATCGGCTGACTTTAATTGATGTAATTGCACATCTTTGTGAGATGTATCGACGGTCTATCCCACGAGA
AACCTGAAAAGGTAAAGGAGGTCTTAAGCTCGGCTCAATTTCTCTCTTTATCCTGTTAGATGGTTTGATTGCTGTTGCTACCCCGTT
GATCTGCGTAGATCTTCTAAGGAGGAAAAAAAATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCG
AGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGA
AGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTGGGGCGTGCAGTGCTTCAGCC
TACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGG
ACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACT
TCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGA
AGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACA
CCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAA
GCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGCTGCAGCCCGGG
GGATCCGTGGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGT
GAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGT
GGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTC
GCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAA
GCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGG
AAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACG
GTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGC
GCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCC
ATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGC
TGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAG
CTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCT
CTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAA
CCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGTGAGAGCTC

EB2073: *PkaiBC*::*RBS*::*tetR*::*linker*::*EYFP* cloned between *NotI* and *SacI* of pAM1303

GCGGCCGCTCTAGACGCAAATTGCAATCTGCATTTGGGTGGAATCGACGGCAATTTTGCAGGATTGCCAGCGGGCGCTGTCGGCC
GATCGCTATCAACTCCAAGTCTGTGAGTCTGGCGAAATGCTCTTGGAGTATGCCCAAACCCATCGTGACCAAATCGACTGCCTGAT
TTTAGTGGCAGCCAATCCCAGCTTCAGGGCAGTTGTTCAGCAGCTCTGCTTTGAGGGAGTGGTGGTACCAGCGATTGTCGTAGGCG
ATCGCGACAGTGAGGATCCCGATGAACCAGCCAAAGAACAGCTCTATCACAGCGCTGAACTGCACCTCGGTATCCATCAGCTCGA
GCAATTGCCCTACCAAGTTGATGCTGCACTGGCTGAATTTCTGCGCTTAGCCCCGGTCGAGACCATGGCCGACCACATCATGCTGA
TGGGGGCCAACCACGATCCCGAGCTATCGAGCCAGCAGCGGGACCTCGCTCAGCGACTACAAGAGCGCCTAGGCTATCTCGGGG
TCTACTACAAGCGTGATCCCGATCGCTTTCTGCGCAACCTACCCGCCTACGAAAGCCAAAAGCTGCACCAAGCGATGCAGACTAG
CTATCGTGAAATCGTTTTGAGCTATTTTTCGCCGAATAGCAACCTCAACCAGAGCATTGACAACTTCGTCAACATGGCTTTCTTTGC
CGATGTTCCAGTCACCAAAGTGGTAGAAATTCACATGGAGCTGATGGACGAGTTTGCCAAGAAGCTCCGCGTAGAGGGACGTTCA
GAGGACATTTTGCTGGATTATCGGCTGACTTTAATTGATGTAATTGCACATCTTTGTGAGATGTATCGACGGTCTATCCCACGAGA
AACCTGAAAAGGTAAAGGAGGTCTTAAGCTCGGCTCAATTTCTCTCTTTATCCTGTTAGATGGTTTGATTGCTGTTGCTACCCCGTT
GATCTGCGTAGATCTTCTAAGGAGGAAAAAAAATGTCTAGATTAGATAAAAGTAAAGTGATTAACAGCGCATTAGAGCTGCTTAA
TGAGGTCGGAATCGAAGGTTTAACAACCCGTAAACTCGCCCAGAAGCTAGGTGTAGAGCAGCCTACATTGTATTGGCATGTAAAA
AATAAGCGGGCTTTGCTCGACGCCTTAGCCATTGAGATGTTAGATAGGCACCATACTCACTTTTGCCCTTTAGAAGGGGAAAGCTG
GCAAGATTTTTTACGTAATAACGCTAAAAGTTTTAGATGTGCTTTACTAAGTCATCGCGATGGAGCAAAAGTACATTTAGGTACAC
GGCCTACAGAAAAACAGTATGAAACTCTCGAAAATCAATTAGCCTTTTTATGCCAACAAGGTTTTTCACTAGAGAATGCATTATAT
GCACTCAGCGCTGTGGGGCATTTTACTTTAGGTTGCGTATTGGAAGATCAAGAGCATCAAGTCGCTAAAGAAGAAAGGGAAACAC
CTACTACTGATAGTATGCCGCCATTATTACGACAAGCTATCGAATTATTTGATCACCAAGGTGCAGAGCCAGCCTTCTTATTCGGC
CTTGAATTGATCATATGCGGATTAGAAAAACAACTTAAATGTGAAAGTGGGTCTGACATCCTCGAGTTGGTGAGCAAGGGCGAGG
AGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCG
AGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGAC
CACCTTCGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAA
GGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACC
CTGGTGAACCGCATCGAGCTGAAGGGCATCAACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAAC
AGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGGCGGC
AGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCT
ACCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCT
CGGCATGGACGAGCTGTACAAGTAAGAGCTC

Table S5.4 Continued

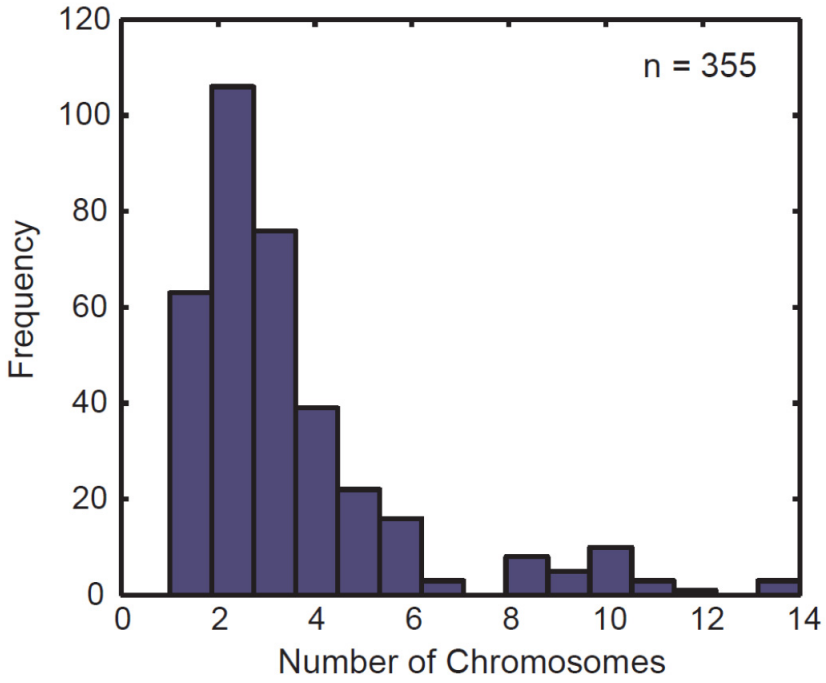


Figure S5.1: Histogram of the number of chromosomes per cell in an exponentially growing wild-type population.

Figure S5.2: Cells maintain chromosome ordering along the long axis in a *ΔminD* strain.

(A) Examples of representative *ΔminD* cells. A single z-section is shown.

(B) Cells maintain chromosome ordering along the long axis in a *ΔminD* strain. Sample size (n) refers to the number of cells analyzed with the given number of chromosomes.

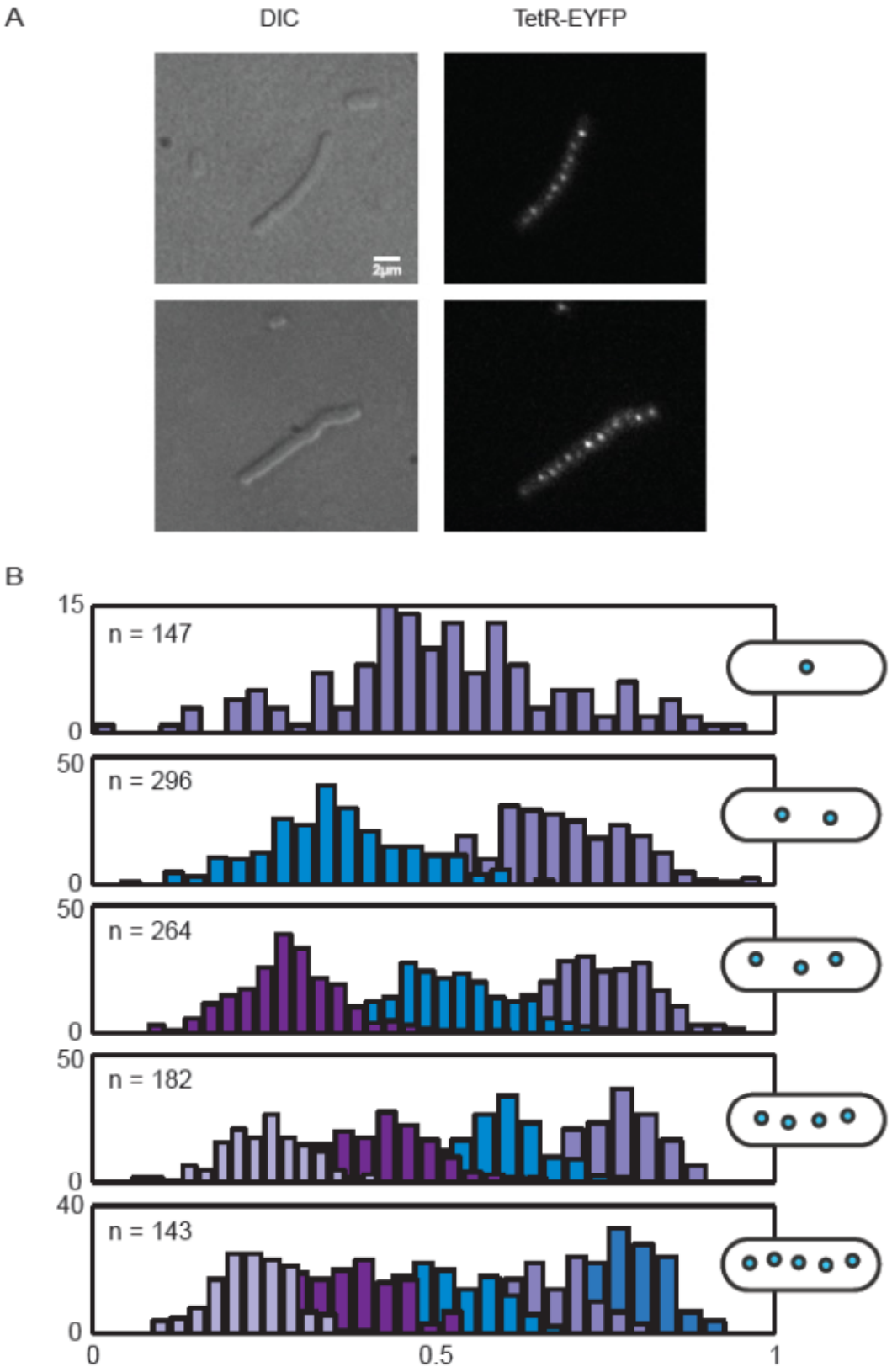


Figure S5.2 Continued

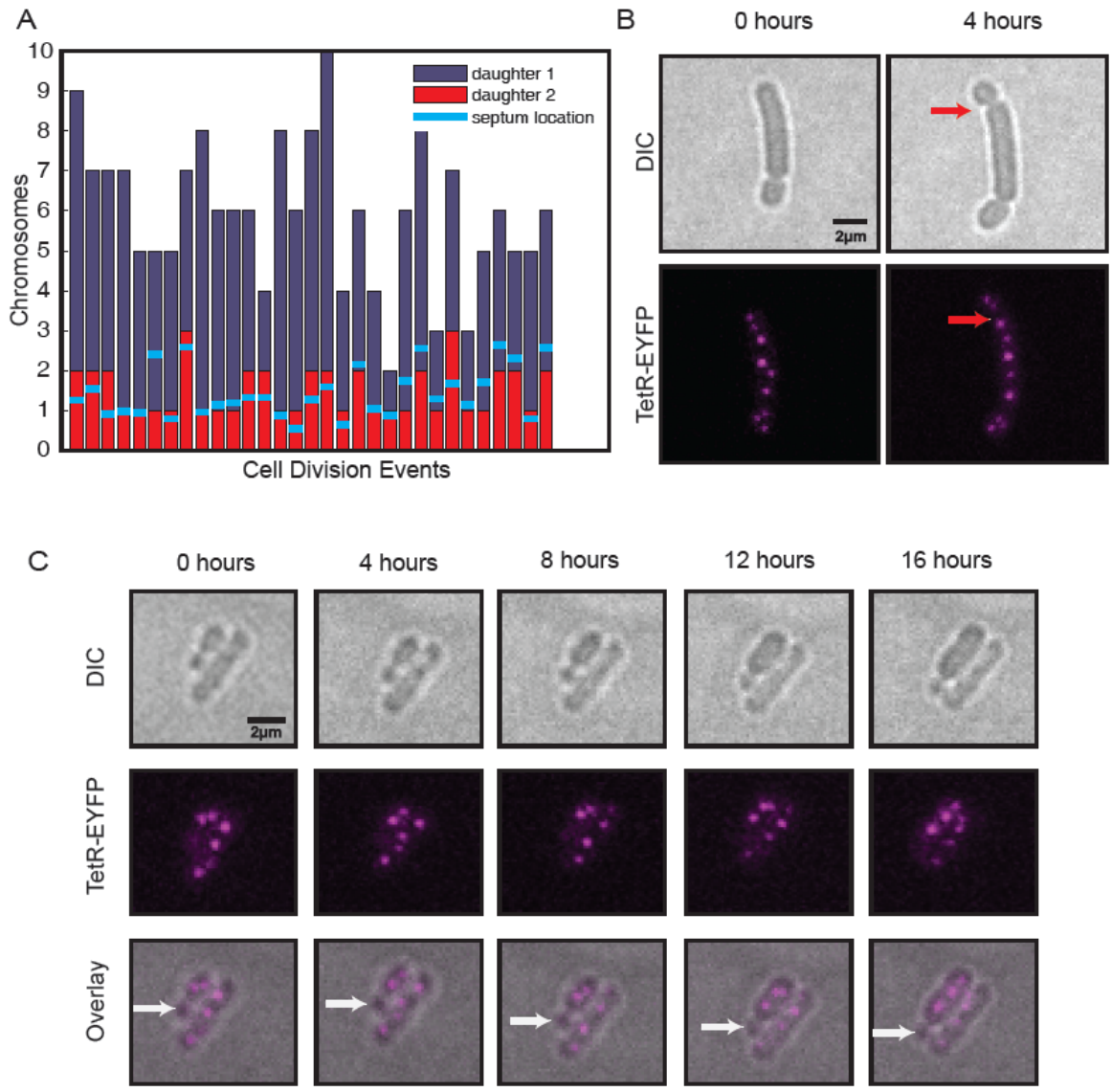


Figure S5.3: Cell division in *ΔminD* cells.

(A) The number of chromosomes partitioned to daughter cells in *ΔminD* cells. Each stacked bar represents a single cell division event with the height of red and dark blue bars representing chromosomes partitioned to each daughter cell. Light blue lines represent the actual septum location as a fraction of total chromosome number (bar height).

(B) An example of chromosome partitioning in *ΔminD* cells based on the position of the septum (red arrow). A single z-section is shown.

(C) Misplaced septum formation in *ΔminD* cells can result in anucleate daughter cells, shown by white arrows. A single z-section is shown.

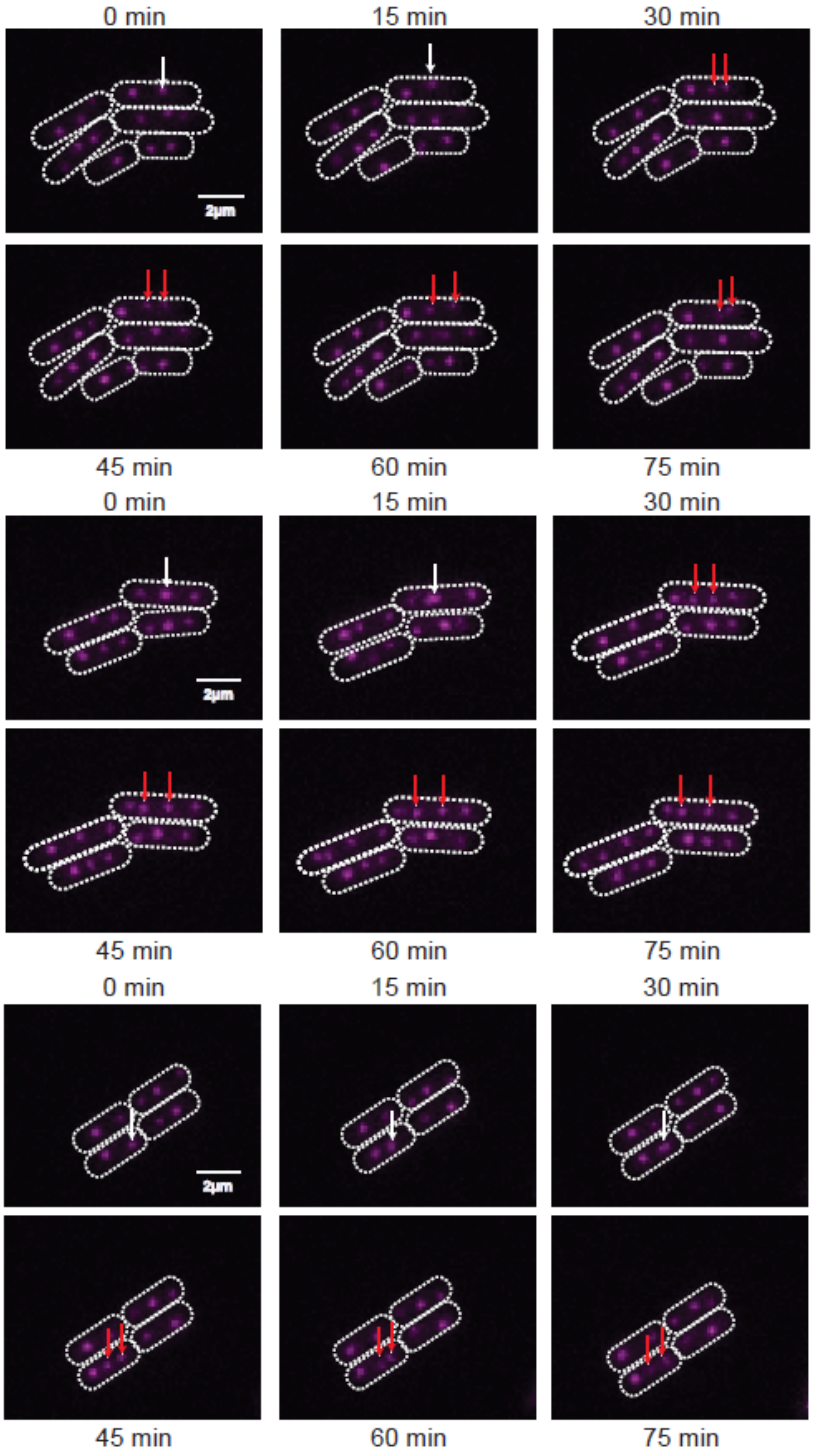


Figure S5.4: Three time courses of wild-type cells. A single genomic locus proximal to the origin is labeled using *tet* operator arrays (pink dots). White arrows point to a replicating chromosome and red arrows point to the resulting, replicated chromosomes. A single z-section is shown for each time course.

References

Chabot JR, Pedraza JM, Luitel P, van Oudenaarden A (2007) Stochastic gene expression out-of-steady-state in the cyanobacterial clock. *Nature* 450(7173):1249-1252.

Lau IF et al (2003) Spatial and temporal organization of replicating *Escherichia coli* chromosomes. *Mol Microbiol* 49(3):731-743.

Mackey SR, Ditty JL, Clerico EM, Golden SS (2007) Detection of rhythmic bioluminescence from luciferase reporters in cyanobacteria. *Methods Mol Biol* 362: 115-129.

Marquis KA et al (2008) SpoIIIE strips proteins off the DNA during chromosome translocation. *Genes Dev* 22(13):1786-1795.