FUNCTIONAL GENOMICS OF THE UNICELLULAR CYANOBACTERIUM *Synechococcus elongatus* PCC 7942

A Dissertation

by

YOU CHEN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Microbiology

FUNCTIONAL GENOMICS OF THE UNICELLULAR CYANOBACTERIUM *Synechococcus elongatus* PCC 7942

A Dissertation

by

YOU CHEN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Susan S. Golden
Committee Members,	Daniel J. Ebbole
	Jin Xiong
	Philip A. Youderian
Head of Department,	Vincent M. Cassone

August 2007

Major Subject: Microbiology

ABSTRACT

Functional Genomics of the Unicellular Cyanobacterium

Synechococcus elongatus PCC 7942. (August 2007)

You Chen, B.S., Nanjing University;

M.S., Chinese Academy of Sciences, Shanghai

Chair of Advisory Committee: Dr. Susan S. Golden

Unicellular freshwater cyanobacterium *Synechococcus elongatus* PCC 7942 is the model organism for studying the circadian clock in cyanobacteria. Despite tremendous work over the last decade in identification of clock-related loci and elucidation of molecular mechanisms of the central oscillator, many details of the basic steps in generating circadian rhythms of biological processes remain unsolved and many components are still missing. A transposon-mediated mutagenesis and sequencing strategy has been adopted to disrupt essentially every locus in the genome so as to identify all of the loci that are involved in clock function.

The complete genome sequence has been determined by a combination of shotgun sequences and transposon-mediated sequences. The *S. elongatus* PCC 7942 genome is 2,695,903 bp in length, and has a 55.5% GC content. Automated annotation identified 2,856 protein-coding genes and 51 RNA coding loci. A system for community refinement of the annotation was established. Organization and characteristic features of the genome are discussed in this dissertation.

More than 95% of the PCC 7942 genome has been mutagenized and mutants affected in approximately 30% of loci have been screened for defects in circadian function. Approximately 70 new clock loci that belong to different functional categories have been discovered through a team effort. Additionally, functional analysis of insertion mutants revealed that the Type-IV pilus assembly protein PilN and the RNA chaperon Hfq are involved in transformation competence of *S. elongatus* cells.

Functional analysis of an atypical short period *kaiA* insertional mutant showed that the short period phenotype is caused mainly by the truncation of KaiA by three amino acid residues. The interaction between KaiC and the truncated KaiA is weakened as shown by fluorescence anisotropy analysis.

Deletion analysis of pANL, the large endogenous plasmid, implies that two toxin-antitoxin cassettes were responsible for inability to cure cells of this plasmid.

In summary, the results indicate that this functional genomics project is very promising toward fulfilling our goal to assemble a comprehensive view of the cyanobacterial circadian clock. The mutagenesis reagents and dataset generated in this project will also benefit the greater scientific community.

DEDICATION

To My Grandmother

Zhou Zhen-Hua (1916-2002),

My Grandfather

Wang Hua-Long (1916-2003),

My Aunt

Jiang Rui-Yun (1948-1987).

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Susan S. Golden, for her generous support and advice. I would also like to thank other members of my committee, Dr. Daniel J. Ebbole, Dr. Jin Xiong, and Dr. Philip A. Youderian for their critical reviews and helpful suggestions.

I am grateful for the support from my wife Xi Chen and my son Gavin Yufeng Chen, my parents Jia-Hai Chen and Chuan-Di Wang, and my parents-in-law Cong-Miao Chen and Xue-Zhen Lu.

I thank my collaborators, Yong-Ick Kim, Dr. Andy LiWang, Rick L. Hammer, Dr. Andrew G. Tag, Dr. Terry L. Thomas, and Dr. Roy D. Magnuson, for their help and excellent work. I thank Dr. C. Kay Holtman, Guogang Dong, Dr. Shannon R. Mackey, Dr. Xiaofan Zhang, and all other colleagues in Golden labs for excellent work, beneficial discussions and critical comments.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Research Background

The unicellular freshwater cyanobacterium *Synechococcus elongatus* PCC 7942

The cyanobacteria are a diverse group of Gram-negative prokaryotes that perform plant-type oxygenic photosynthesis (1). The unicellular cyanobacterium *S. elongatus* PCC 7942, formerly called *Anacystis nidulans* R2, is a freshwater obligate photoautotroph. It is closely related to *S. elongatus* PCC 6301 (*Anacystis nidulans*), which was isolated in Austin, Texas, in 1952 and was the first strain deposited in Pasteur Culture Collection of Cyanobacteria (1). Despite of its major contributions in elucidation of phycobilisome makeup (2) and DNA photolyase structure (3), PCC ¯6301 was largely substituted by PCC 7942, which was isolated later from California in 1973 (Pasteur Culture Collection) and was demonstrated to be naturally transformable (4). The *S. elongatus* cells are rod-shape, around 1-2 μm in width and a few μm in length (Fig. 1-1). These cells form round colonies on agar plates and suspend homogeneously (free-floating) in liquid medium. They do not have the ability to move around in liquid, but sometimes aggregate into flake-like clumps or films on surfaces of glassware. Like many other unicellular cyanobacteria, they undergo binary fission for reproduction. Unlike many other cyanobacteria, they cannot fix nitrogen.

This dissertation follows the style of *Proceedings of the National Academy of Sciences*.

Many lines of research have studied the relationship between these two *S. elongatus* strains. DNA reassociation kinetics experiments show that there is very little genetic difference between them (5). A single inversion event in their genomes accounted for a few RFLP (Restriction Fragment Length Polymorphism) differences between them (6). The recently published complete genome sequences of *S. elongatus* PCC 7942 (GenBank Accession No. NC_007604) and PCC 6301 (GenBank Accession No. NC_006576) (7) revealed that the overall nucleotide sequence of these two strains shows 99.93% identity. In addition to the previously identified ~188.6 kb large inversion between the two genomes, there are two other small regions in *S. elongatus* PCC 6301 that are deleted in the *S. elongatus* PCC 7942 genome (7). The genome of *S. elongatus* PCC 7942 consists of a circular chromosome (2.7 Mb), an essential large plasmid (pANL, ~46.3 kb), and a non-essential small plasmid, pANS (~8.6 kb). The G+C content of the PCC 7942 genome is around 55.5%. Chapter II of this dissertation comprises a more complete discussion of the genome content.

The only noticeable physiological and genetic difference between two *S. elongatus* strains is the superior transformation properties of *S. elongatus* PCC 7942 (4, 8), which has made it a model organism for studies of photosynthesis and light regulation (9-13), signal transduction (14, 15), transcription and its regulation (16, 17), response to nutrient deprivation (18, 19), as well as iron, sulfur, nitrogen, and carbon metabolism (15, 17, 20-25) and many more biological questions for decades. In recent years, *S. elongatus* PCC 7942 has become the only developed model organism for exploring the mechanism of a prokaryotic circadian clock (26). A functional genomics

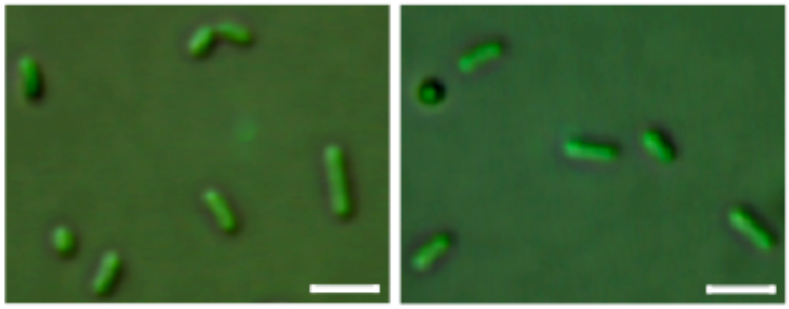


Fig. 1-1. Representative images of *S. elongatus* PCC 7942 cells under the light microscope. Bright field images were acquired with a Zeiss Axioplan2 microscope. Both images were processed with Adobe Photoshop. The scale bars represent 5 μm.

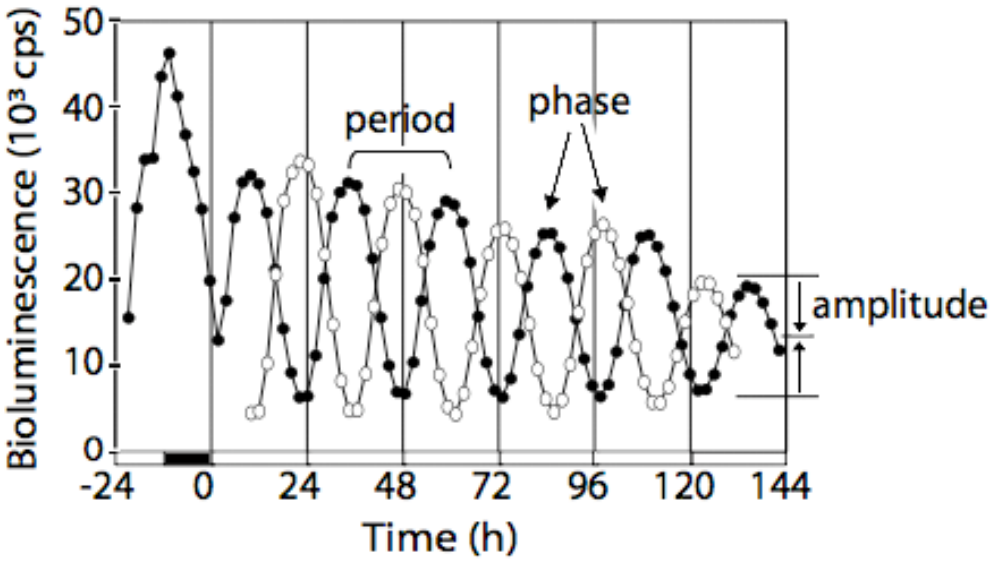


Fig. 1-2. Bioluminescence traces from reporter strains produces characteristic circadian patterns that depend on the promoter used to drive luciferase expression. Various mutants affect circadian period (length of cycle, *e.g.,* peak-to-peak), relative phasing (peak time relative to a reference point, such as lights on), or amplitude of the rhythm (deviation of peak and trough from the mean of the oscillation). X axis, circadian time in hours; cells were released to constant light (0-144 h) after one day's light/dark entrainment (-24-0 h). Y axis, bioluminescence plotted as 10^3 counts per second (10^3 cps).

project (http://www.bio.tamu.edu/synecho/index.htm) is on-going with aims to assay the function of each gene of the *S. elongatus* PCC 7942 genome through gene inactivation, with a focus on the phenotypes associated with the circadian rhythmicity of gene expression. The mutagenesis data will be available to the scientific community as soon as possible through publicly accessible databases. This project is described more fully in Chapter III of this dissertation.

Circadian rhythmicity is probably ubiquitous in cyanobacteria

The sun rises at dawn and sets at dusk with a very predictable daily pattern. In adaptation to daily changes in light, temperature, humidity, as well as other environmental factors, many living organisms have developed an internal timing system during the course of evolution. This intrinsic timing system, called a circadian clock, allows organisms to sense and even anticipate the environmental signals to optimize their daily physiological activities, as well as behavior, metabolism and gene expression (27). In constant conditions, circadian rhythms show a cycle time, or period, of about (circa) a day (dies). They can be entrained (phase reset) to appropriate environmental phases of light/dark, or temperature. In addition, their free-running period is temperature compensated, that is, nearly constant over a physiological range of ambient temperatures (27) (Fig. 1-2).

Most eukaryotic organisms, such as fungi (*Neurospora*), insects (*Drosophila*), plants (*Arabidopsis*), birds (chicken), and mammals (mice, humans etc.), have developed this endogenous circadian timer during evolution (28, 29). Cyanobacteria, a group of photosynthetic eubacteria, are the simplest organisms and the first prokaryotes known to

possess a circadian clock (26, 30). Studies on nitrogen fixation in some unicellular cyanobacteria revealed the earliest evidence for circadian rhythms in cyanobacteria. In 1985 Stal & Krumbein (31) showed a circadian rhythm of nitrogenase activity in a non-heterocystous filamentous cyanobacterial species, *Oscillatoria*. A similar rhythm of nitrogenase activity was also found in two unicellular marine cyanobacteria strains *Synechococcus* Miami BG 43511 and 43522 (32). However, the authors did not recognize that the phenomena they observed were controlled by an internal circadian clock. It was Huang et al in 1986 (33) who reported a *bona fide* circadian clock identified in the freshwater strain *Synechococcus* RF-1 based on the study on nitrogenase activity. Sweeney and Borgese in 1989 (34) found a temperature-compensated daily cycling of cell division in marine *Synechococcus* WH 7803 and also designated it as a genuine circadian rhythm. Since then, more cyanobacterial species have been established to have endogenous circadian clocks. A marine unicellular diazotroph, *Cyanothece* ATCC 51142, shows persistent circadian rhythms in photosynthesis and nitrogen fixation, as well as accumulation of stored carbohydrates (35). In the unicellular *Synechocystis* PCC 6803 (36), rhythmic expression of the bioluminescence from bacterial luciferase *luxAB* reporter genes has been demonstrated. Data from another *Cyanothece* species, BH68K, imply that the rhythmic expression of *ntcA* and *nifHDK* transcription may be under the control of a circadian clock (21). A circadian rhythm of nitrogenase gene expression was also confirmed in the diazotrophic filamentous non-heterocystous cyanobacterium *Trichodesmium* IMS101 (37-39).

Overall, the accumulating evidence for circadian rhythms in various species of cyanobacteria greatly supports the assumption that circadian clocks are widespread among members of this diverse group. Sequences related to a gene that lies at the heart of the circadian oscillator in PCC 7942, *kaiC*, have been identified in up to 40 diverse strains of cyanobacteria by degenerate PCR survey and subsequent hybridization and sequence analysis (40). The circadian clock in cyanobacteria indeed provides adaptive significance. When strains with different circadian periods were co-cultured under different light/dark cycles, the strain whose endogenously-defined circadian period is closest to the imposed cycle length always outgrew other strains after over 20 days of incubation (41, 42). This adaptive advantage persisted in rhythmic environmental conditions, but disappears in constant conditions (43). The "fitness" advantage of the circadian clock has also been demonstrated in other model organisms, such as *Arabidopsis* (44, 45) and *Drosophila* (46).

Synechococcus elongatus PCC 7942 is the model organism for cyanobacterial clocks

The majority of the studies of cyanobacterial circadian rhythms has been done with *Synechococcus elongatus* PCC 7942 (47, 48), a unicellular fresh water obligate photoautotroph. The early studies of the circadian rhythms in *S. elongatus* demonstrated that expression of the *psbAI* gene, encoding the D1 subunit of photosystem II, is rhythmic in constant light, and this periodicity, which is temperature compensated, can be entrained by light/dark (L/D) cycles (30, 49). It takes at least several days of continuous monitoring to establish a circadian phenotype. For this purpose, a high

throughput, high precision, non-invasive screening method for unremitting recording of the circadian rhythms has been developed. This method is based on artificial bioluminescent reporters constructed by fusing the promoterless *luxAB* gene set, which encodes the luciferase enzyme from the marine bacterium *Vibrio harveyi*, to the promoter of desired cyanobacterial genes (30, 50-52). Alternatively, the firefly luciferase (*luc*) gene can also serve as a reporter.

Bioluminescence from cyanobacterial cells in 96-well microplates or individual colonies on agar plates can be counted automatically and continuously by a luminometer. The resulting bioluminescence records over time provide easily assayed circadian phenotypes for mutant analysis (Fig. 1-2). In addition, *S. elongatus* PCC 7942 offers many other advantages for elucidating clock function, which makes it the model system of circadian clock in cyanobacteria: ***1)*** it is easily cultured (generation time can be as fast as 6-8 hours); ***2)*** it is naturally transformable and favors homologous recombination with high efficiency; ***3)*** it has a small genome size (~2.7 Mb) and less redundancy than many other well-studied cyanobacteria (most genes are present in single copy).

The alternative model system for studying cyanobacterial clock function would be the unicellular facultative photoheterotroph *Synechocystis* sp. PCC 6803. The 3.6 Mb genome of *Synechocystis* PCC 6803, a distantly related species to *S. elongatus* PCC 7942, was sequenced several years ago (53), and it has been shown to exhibit circadian rhythms of gene expression (54, 55). However, the circadian rhythms from reporter genes are less robust than those in *S. elongatus* PCC 7942, and the repetitive

organization of the *Synechocystis* PCC 6803 genome is a major disadvantage in analysis of gene regulation and metabolism. *Synechococcus* PCC 7942 has much less functional redundancy. There is only one copy of *kaiC* in the genome of *S. elongatus* PCC 7942, whereas *Synechocystis* PCC 6803 has three paralogs of *kaiC*. The relative simplicity of the circadian clock in *Synechococcus* PCC 7942, together with the other advantages it offers, makes it a better system for elucidating the clock mechanism.

Circadian clock genes in *Synechococcus elongatus* PCC 7942

A clock system consists of three major parts: an input pathway (a mechanism for sensing environment cues, such as light and temperature, and setting the phase of the clock), a central oscillator (the timekeeper itself), and an output pathway (a means of relaying clock phasing to the various behaviors controlled by the clock) (56). Several genes necessary for clock function have been identified in *S. elongatus* PCC 7942, including *kaiABC* (57), a gene cluster that encodes the components of the central oscillator; *cikA* (58) and *ldpA* (59), input pathway components; key output pathway components, sensor kinase gene *sasA* (60) and its cognate response regulator gene *rpaA* (61); as well as other genes: *pex* (62), *cpmA* (63), *labA* (64), the *clpP2clpX* locus (65), and group 2 sigma factors (66, 67) (Fig. 1-3). Conservation of these clock genes among cyanobacterial genomes will be discussed in Chapter II. Clock functions of individual proteins are summarized below.

The *kaiABC* genes encode the circadian pacemaker of *S. elongatus* PCC 7942 (57). Both the monocistronic *kaiA* transcript and dicistronic *kaiBC* transcript display

circadian cycling in abundance (57). The KaiB and KaiC protein levels are also robustly rhythmic, whereas the KaiA protein level is not (68). Mutations in any one of the genes cause changes in circadian period (24-25 h in wild-type strains), ranging from 14 to 60 hours or complete arrhythmicity (49). Details of current studies on Kai proteins and the central oscillator will be discussed below.

Iwasaki *et al,* (60) identified a histidine kinase, SasA (*Synechococcus* Adaptive Sensor), which physically interacts with KaiC by using a yeast two-hybrid screen and informatics. SasA contains a sensory domain that is similar to KaiB and which interacts with KaiC. The similarity of the sensory domain to kaiB is mainly at the amino acid sequence level, but not evident at the 3-dimentional structure level (69). Inactivation of *sasA* alters circadian expression (period, amplitude, or phase angle) of all tested reporter genes and reduces the amplitude of many to the point of apparent arrhythmicity (60). Overexpression of *sasA* eliminates circadian rhythms of a P*kaiBC* reporter strain. SasA protein is important for the formation of the clock protein complex and assembles with Kai proteins in a circadian fashion (70). SasA must function very close to the central oscillator of *S. elongatus* PCC 7942. Its cognate response regulator, RpaA, has recently been identified to be an OmpR-like DNA binding protein, which mediates between the central oscillator and downstream clock-controlled global circadian gene expression (61).

Inactivation of *cikA*, a circadian input kinase, shortens the circadian period of *S. elongatus* PCC 7942 and abolishes the normal resetting of circadian phase by a 5-hour

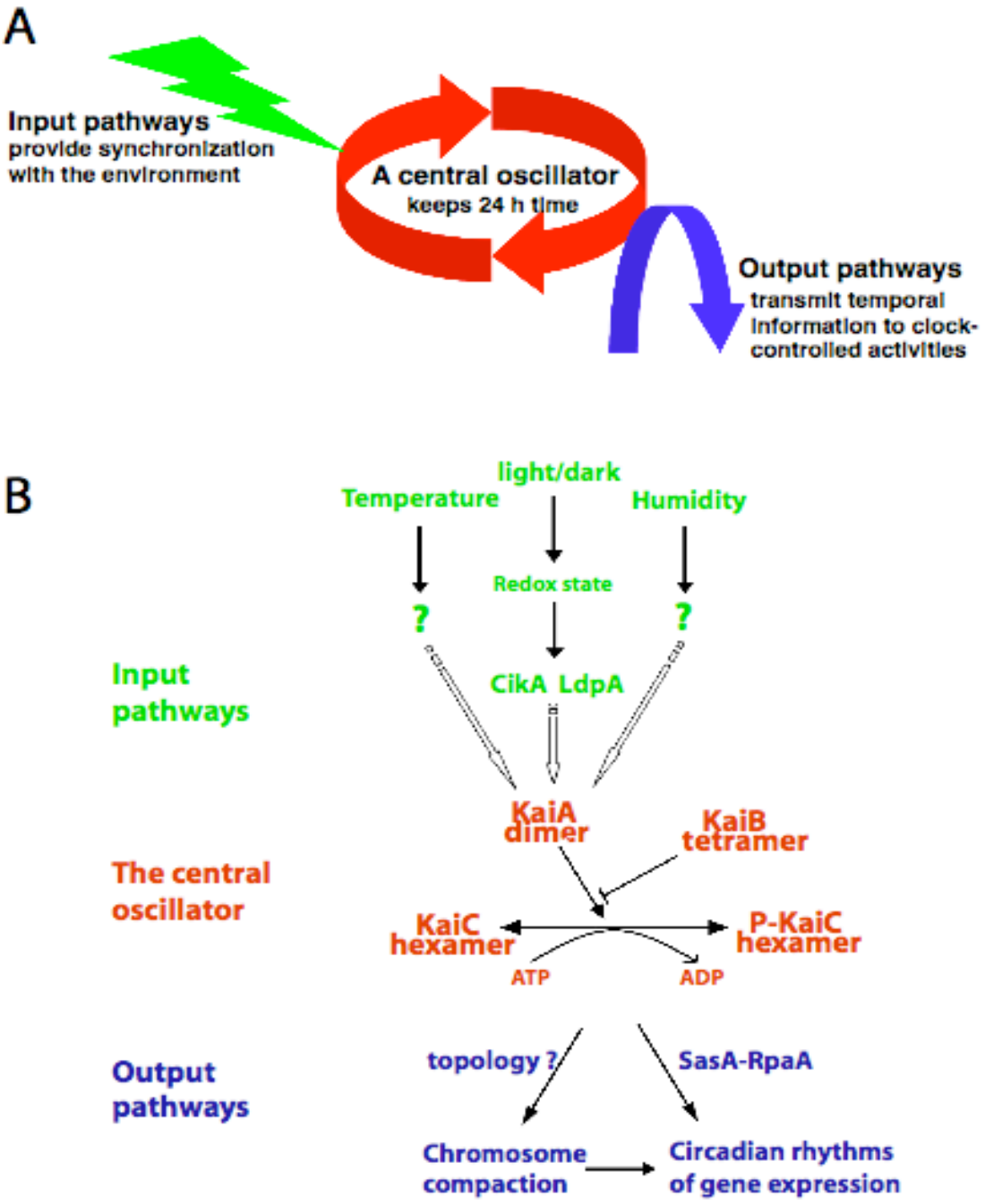


Fig. 1-3. Simplified depiction of the circadian clock in *S. elongatus* PCC 7942. ***A)*** The circadian clock consists of a central oscillator, input pathways, and output pathways. ***B)*** Key components of each of these divisions have been identified for *S. elongatus*, some of which are indicated here.

dark pulse (58). There are three conserved domains identified in CikA: a GAF domain, a histidine protein kinase (HPK) domain, and a *pseudo*-receiver domain (PsR). The autophosphorylation activity of the genuine HPK domain was shown to be enhanced by the GAF domain but attenuated by the PsR domain (71, 72). The PsR domain also docks CikA to the cell poles (72). Furthermore, the PsR domain directly binds a quinone analog 2,5-dibromo-3-methyl-6-isopropyl-p-benzoquinone (DBMIB, a photosynthesis inhibitor that prevents the electron transfer from plastoquinone pool to cytochrome complex b6/f), which results in destabilization of CikA (73). The data suggest that CikA responds to light density indirectly by sensing the redox status of the plastoquinone pool and hence the strength of photosynthesis activity. Studies on the NMR structure of the PsR domain revealed a likely quinone-binding surface and suggested a model for its interaction with the HPK domain (74). Thus, this bacteriophytochrome-like protein is likely a key component of an input pathway of the clock.

Another putative component of an input pathway, *ldpA* (Light-Dependent Period), encodes an iron-sulfur protein involved in light-dependent adjustment of the period of the *S. elongatus* PCC 7942 circadian clock (59). The LdpA protein was suggested to be a member of the ferredoxin superfamily (75), which carries two 4Fe-4S clusters and also senses the redox state of the cell (76). The co-purification of LdpA, KaiA, CikA, and SasA suggests that input pathway components are likely forming a large protein complex with the central oscillator and closely associated output pathway components, possibly in a circadian fashion (76).

The product of the period-extender gene, *pex*, probably functions as a modifier of the circadian clock. The expression of *kaiA* is greatly enhanced when *pex* is disrupted (62). Pex was also found to be dark-responsive and required for extending the circadian clock in light/dark cycles (77). A putative output pathway component, *cpmA* (Circadian Phase Modifier), displays low-amplitude bioluminescence and altered circadian phasing phenotypes in a subset of reporter strains (*psbAI*, *psbAII*, and *kaiA*) when disrupted (63).

The sigma factor encoded by *rpoD2*, a member of the sigma70-like transcription factor gene family in *S. elongatus*, modifies the circadian expression of a subset of *S. elongatus* PCC 7942 genes. It is likely to be a component of an output pathway of the clock (66). Inactivation of other sigma factors (*rpoD3*, *rpoD4*, and *sigC*) also alters the circadian rhythms of a P*psbAI* reporter strain, but not that of P*kaiB*. Two of the double-mutation combinations significantly affect the expression of P*kaiB*, although all of them alter P*psbAI* expression. The data indicate that these sigma factors, although structurally similar and partially overlapping in function, are not entirely redundant in their clock function (67).

Other newly identified clock components include: LabA (Low-Amplitude and Bright), which affects negative feedback regulation of KaiC (64); and an ATP-dependent Clp protease complex, ClpP2 (the protease subunit) and ClpX (the ATPase subunit). The partial disruption of *clp* genes caused extended circadian period. They are the first genes identified to be involved in cyanobacterial circadian clock function that are essential for viability (65).

Current molecular mechanism for the cyanobacterial circadian oscillator

In addition to identification of these clock genes, tremendous progress has been achieved in elucidating the molecular mechanism of the central oscillator of the cyanobacterial circadian clock.

Transcription-translation feedback loops (TTFLs), consisting of transcriptional factors that regulate one another, sit at the center of eukaryotic circadian clock models (28). In any TTFL, clock proteins negatively regulate their own transcription such that rhythmic levels of mRNAs and hence rhythmic levels of proteins are produced. At least two interlocking feedback loops are required for robust oscillating of a circadian clock.

A TTFL also exists in the cyanobacterial circadian clock system. KaiC represses activities of the promoter of the dicistronic *kaiBC* operon, which is enhanced by KaiA (57). A negative element upstream of the *kaiBC* promoter has been identified in the C-terminal coding region of *kaiA* (78). It is probably relevant to the stimulation of *kaiBC* operon expression by KaiA. However, a heterologous promoter, such as a leaky *trc* promoter from *E. coli* or *purF* from *S. elongatus* that peaks 12 h out of phase from *kaiBC*, can drive *kaiBC* to sustain robust circadian rhythms of bioluminescence of reporter genes with normal period and phase (79-81). Furthermore, the rhythmic KaiC phosphorylation pattern is preserved even in constant darkness, during which *kaiBC* mRNA levels do not cycle at all (82). Finally, robust KaiC phosphorylation cycles were reconstituted in a test tube with the presence of only purified KaiA, KaiB, and KaiC proteins and ATP (83). Thus, the cyanobacterial circadian clock is likely to tick-tock using a post-translational oscillator rather than a TTFL.

It is well known that the whole genome of *S. elongatus* PCC 7942 is under the control of the circadian clock (84). The two-component signal transduction histidine protein kinase SasA and its cognate response regulator RpaA are necessary for global rhythmic gene expression (60, 61). Temporal information transmitted from the central oscillator is also involved in a chromosomal compaction rhythm, which is temperature-compensated, but SasA independent (85). It was demonstrated that autoregulatory feedback by KaiA and KaiC is exerted globally rather than specifically on the *kaiBC* promoter (78, 81, 86). Thus, TTFLs from Kai proteins still have their significance, not directly on the central oscillator, but probably on input and output pathways that help to sustain rhythmicity (87).

Structures of all three Kai proteins have been resolved: KaiA proteins form dimers (69, 88); KaiB proteins are tetrameric complexes (89, 90); KaiC forms a hexamer in the presence of ATP (91, 92). Phosphorylation and de-phosphorylation on key threonine and serine residues at the C-terminal domain of KaiC is likely to be the essential timing mechanism for cyanobacteria circadian clock (93, 94). KaiC has two duplicated domains, with each containing an ATP-binding Walker's motif (95). It has been shown that the N-terminal motifs are required for KaiC hexamerization, while the C-terminal motifs are involved in autokinase activity of KaiC (96).

KaiA stimulates KaiC autophosphorylation, while KaiB attenuates KaiA-enhancement of the KaiC auto-phosphorylation state (95, 97-99). The interaction between KaiA and the C-terminal domain of KaiC probably occurs at two interfaces: the C-terminal KaiC peptide (100) and the ATP-binding pocket on KaiC (101). The binding

interface of KaiB on KaiC is yet to be determined, but is has been suggested that KaiA and KaiB from *Anabaena* compete for a common binding site on KaiC (102).

The Kai proteins, plus SasA, interact with each other to form protein complexes in a circadian fashion *in vivo* (70, 103). An *in vitro* study also showed that KaiC form various complexes with KaiA or KaiB or both during the circadian time course (104). Moreover, KaiC hexamers of different phosphorylation status exchange monomers, which is important to sustain robust circadian rhythms (104, 105).

Although potential orthologs of KaiC have been found in almost all sequenced genomes of cyanobacteria (except *Gloeobacter*) and of many *Archaeal* and proteobacterial species (106), no apparent homologs of the KaiABC proteins are encoded in any eukaryotic genome or any chloroplast genome of higher plants. Conversely, homologs of known eukaryotic clock proteins, such as PER, TIM, CLOCK, BMAL and FRQ (107), have not been found in any cyanobacterial genome. Even among eukaryotes, key clock proteins in animals are not present in fungi or plants, nor are the clock proteins of these latter two groups present in the other two (28). Thus, there must be different origins and evolutionary paths for these different clock systems. Nonetheless, there are some aspects that are very similar between the cyanobacterial circadian clock and its eukaryotic counterparts: ***1)*** Phosphorylation of core clock proteins is essential in all clock systems; ***2)*** Transcription-translation feedback loops are present in all circadian clocks, though their role in ticking and sustaining circadian rhythms varies; ***3)*** Several conserved domains, such as PAS and GAF, are found in clock proteins from many clock model organisms. The recent progress in elucidating the

circadian oscillator of *S. elongatus* PCC 7942 not only deepens our understanding in molecular mechanisms of prokaryotic clocks, but also provides insights to corresponding eukaryotic systems.

Despite recent achievements in elucidating the molecular mechanism of the cyanobacteria circadian clock, many details of the basic steps (input, central oscillator, and output) in generating circadian rhythms of biological processes remain unsolved and many components are still missing. The input pathway components that connect CikA and KaiA have not yet been fully identified. How KaiA and KaiB function to regulate the phosphorylation status of KaiC is still not clear. The period length of cyanobacterial circadian clocks was suggested to be determined by phosphorylation status and degradation rate of KaiC (80, 83). However, the molecular mechanism of how interactions among Kai proteins affect the phosphorylation status of KaiC and hence the circadian period is currently obscure. Phase determination is another mystery. Apparently, no specific *cis* element is responsible for determination the timing of peak gene expression (108). The cyclic pattern of KaiC phosphorylation might not be relevant to phase (109). It is likely that DNA topology or chromosome compaction is involved in phase determination and phase resetting (85, 108). We also do not understand how the circadian clock is embedded in cellular physiology in cyanobacteria.

Functional genomics of *Synechococcus elongatus* PCC 7942

To fully understand how the circadian clock in cyanobacteria functions at the molecular level and how is it entrained by environmental signals, and to further understand the

physiological significance of circadian rhythms in cyanobacteria, it is necessary to identify all of the genes that are required for clock function. The most thorough approach to achieve this goal is to inactivate every single gene in the whole genome and screen every mutant for circadian phenotypes.

Transposon-based mutagenesis is widely used in functional genomics analysis of many genomes, including bacteria (110-112), fungi (113), plants (114, 115), worms (116), mammals (117, 118), and pathogens (119-121). Transposons are mobile DNA fragments ubiquitous in prokaryotic and eukaryotic genomes that have become a powerful tool for insertional mutagenesis. Compared to other *in vivo* mutagenesis methods, such as chemical mutagenesis (122) and UV mutagenesis (123), in which characterization of isolated point mutations and cloning of the mutated genes are often time-consuming, the main advantage of transposon-mediated mutagenesis is that transposons carry selectable markers, such as an antibiotic resistance gene that tag the affected locus and facilitate cloning (124). In addition, transposons can be engineered to contain sequencing primer binding sites, signature tags, or even promoters, which greatly facilitate the following molecular identification and genetic analysis (125). The feature that transposons can randomly insert into any locus in a genome, usually as a single insertion, makes it very convenient to construct pools of mutants *in vivo* or *in vitro* for global phenotypic analysis (126, 127).

The *in vitro* transposition systems offer many advantages over *in vivo* systems in functional genomics because of higher efficiency and less specificity for insertion site (110, 125). In addition, the *in vitro* reactions are generally more reliable, economical,

and convenient for large-scale mutagenesis. Usually, cosmid or plasmid libraries of genomic DNA are mutated *in vitro* and then reintroduced into a target organism that favors double recombination (allelic substitution), through transformation or conjugation to disrupt target genes. Highly efficient *in vitro* mutagenesis systems have been developed for many well-studied transposons, such as Tn*5* (128, 129), Tn*7* (130), and *Mu* (131), which are also commercially available with various antibiotic resistance markers: Tn*5* (EZ-Tn5™ Transposon Tools, Epicentre), Tn*7* (GPS™ Mutagenesis System, New England Biolabs) and *Mu* (GeneJumper™ Kit, Invitrogen).

Both *in vivo* and *in vitro* transposon mutagenesis, mainly based on Tn*5*, have been widely used in cyanobacteria. A Tn*5* derivative (Tn*5*-1063) was introduced into *Nostoc* sp. ATCC 29133 (*Nostoc punctiforme* PCC 73102) to generate mutants with defective phenotypes in nitrogen fixation (132). Several cell division genes were identified in *S. elongatus* PCC 7942 using another derivative of transposon Tn*5* (Tn*5*-692) (133, 134). Tn*5* derivatives have been used to isolate swimming motility mutants in a marine *Synechococcus* strain (135). All above are *in vivo* examples. *In vitro* Tn*5*-mediated mutagenesis has also been used in *Synechocystis* sp. PCC 6803 for mutants that were defective in optimal photoautotrophic growth (136) or in motility (137).

Traditional *in vivo* mutagenesis has contributed in identification of most known clock-related loci, which are non-redundant and non-essential, in *S. elongatus*, such as usage of chemical mutagenesis (EMS, ethyl methanesulfonate) to discover the *kaiABC* locus (49, 57) and application of Tn*5* transposon-mediated insertional mutagenesis in identification of *cikA* (52, 58). Error-prone PCR, an *in vitro* mutagenesis strategy, was

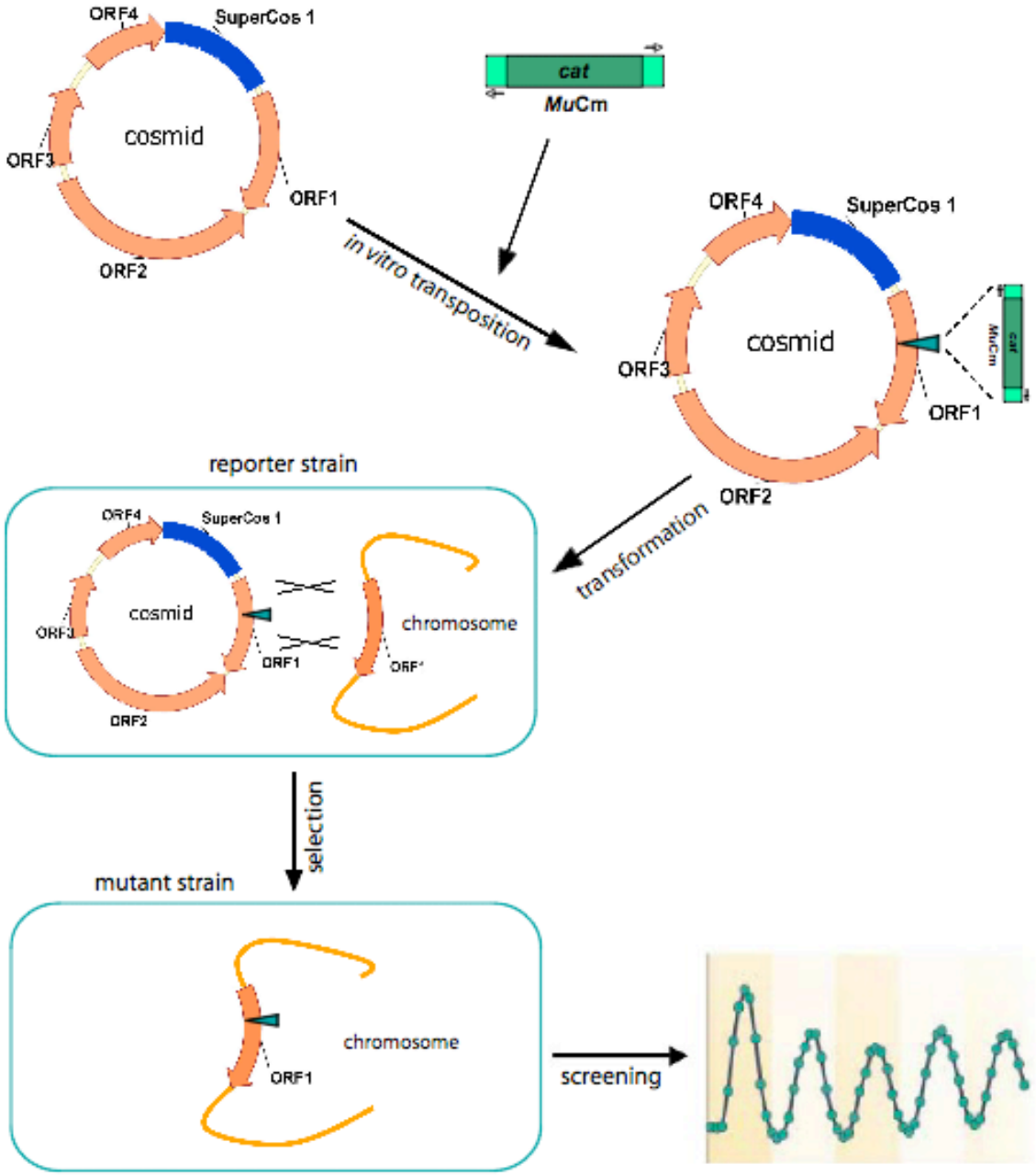


Fig. 1-4. Strategy for transposon *Mu*-mediated mutagenesis and sequencing in *S. elongatus* PCC 7942. Steps of *in vitro* transposition, transformation, selection, and circadian clock phenotype screening are shown. See text for details.

used to produce hundreds of point mutations in *kaiA* (138). As a strain famous for natural transformation and efficient double-recombination, *S. elongatus* PCC 7942 is amenable to large scale *in vitro* mutagenesis.

For global mutagenesis of *S. elongatus* PCC 7942 genome, two transposon systems were used as described in Chapter III: Tn*5* from Epicentre and *Mu* from Invitrogen. Bacteriophage *Mu* (139, 140) has the least target-specificity among known transposons (131, 141), with a broad consensus target site NY(G/C)RN or CY(G/C)RG (142, 143), which is relatively frequent in GC-rich *S. elongatus* (55.5% G+C). The base composition in open reading frames (ORFs) of *S. elongatus* averages about 60% in the third codon position, whereas intergenic regions tend to be more AT-rich. Thus, there are more chances for *Mu* to hop into coding regions. Another option, the hyperactive *in vitro* Tn*5* transposition system (128), modified with reduced site specificity, is also sufficient for inactivating essentially any gene (129).

As depicted in Fig. 1-4, *in vitro* transposition reactions, using *Mu* as an example, were performed on cosmids that carry ~30-40 kb genomic DNA. A collection of insertional cosmid alleles was then introduced into cyanobacterial reporter strains *via* transformation. Transposons were integrated into corresponding chromosome loci through double recombination. Insertional mutants were screened for circadian phenotypes by checking altered bioluminescence traces from the reporter genes. The primer binding sites at the ends of *Mu* transposon were used to amplify flanking genomic sequences for localization the insertion site.

Objectives of This Dissertation Project

The original aims of this project were: ***1)*** to inactivate each open reading frame (ORF), of the *S. elongatus* genome *via* transposon-mediated mutagenesis, ***2)*** to screen insertional mutants for altered circadian phenotypes, and ***3)*** to determine the complete genome sequence. In cooperation with the DOE Joint Genome Institute (JGI), the genome sequence has been finalized using a combination of JGI shotgun sequences and our transposon-mediated sequences, which greatly facilitated the functional analysis of the genome.

Chapter II describes sequence determination and functional annotation, as well as organization and characteristics of the genome of *S. elongatus* PCC 7942. Chapter III focuses on the strategy and progress of the functional genomics project. Also included is insertional analysis of *pilN*, encoding a type IV pilus assembly protein, and *hfq*, encoding a RNA chaperon, which are both involved in natural competence of *S. elongatus* cells. The comprehensive analysis of an atypical short-period *kaiA* insertional mutant is summarized in Chapter IV. The deletion analysis of the large endogenous plasmid of *S. elongatus*, pANL, is described in Chapter V.

CHAPTER II

THE COMPLETE GENOME SEQUENCE OF *Synechococcus elongatus* PCC 7942

Introduction

As a group of photoautotrophs, cyanobacteria show diversity not only in morphology and geography, but also in genome size and nucleotide composition (144). To date, dozens of cyanobacterial genomes have been completely sequenced and many more projects are under way.

Among approximately 50 cyanobacterial sequencing projects deposited in GenBank (NCBI), 24 are complete and 11 are in draft of contig assemblies, while others are still in the sequencing stage of progress (Table 2-1). These organisms are from genera *Acaryochloris*, *Anabaena*, *Crocosphaera*, *Cyanothece*, *Gloeobacter*, *Gloeothece*, *Lyngbya*, *Microcystis*, *Nodularia*, *Nostoc*, *Prochlorococcus*, *Prochloron*, *Synechococcus*, *Thermosynechococcus*, and *Trichodesmium*. The majority of these sequencing projects resulted from an increasing desire to understand the geographical, functional, and ecological roles of marine microbes. All 12 *Prochlorococcus* strains belong to the same *Prochlorococcus marinus* species, and 13 out of 17 *Synechococcus* strains live in coastal and/or open sea territories. Among the 4 fresh water *Synechococcus* strains, *Synechococcus* sp. JA-2-3B'a(2-13) and *Synechococcus* sp. JA-3-3Ab, both living at environmental temperatures over 50°C, were isolated from the

Table 2-1. Cyanobacterial sequencing projects listed in GenBank.

Organism	Size (Mb)*	Status	Accession Number	Reference	Sequencing center†
Acaryochloris marina MBIC11017	est. ~4.2	In progress	-	-	ASU
Acaryochloris sp. CCMEE 5410	est. ~4.2	In progress	-	-	GBMF
Anabaena variabilis ATCC 29413	~6.37	Complete	NC_007413	-	JGI
Crocosphaera watsonii WH 0002	est. ~5.0	In progress	-	-	GBMF
Crocosphaera watsonii WH 8501	est. ~6.24	Draft assembly	NZ_AADV00000000	-	JGI
Cyanothece sp. CCY0110	est. ~5.88	Draft assembly	NZ_AAXW00000000	-	GBMF
Dermocarpa sp. 0006	est. ~5.0	In progress	-	-	GBMF
Gloeobacter violaceus PCC 7421	~4.66	Complete	NC_005125	Nakamura et al. (2003)	Kazusa
Gloeothece sp. PCC 6909	est. ~2.5	In progress	-	-	GBMF
Lyngbya sp. PCC 8106	est. ~7.04	Draft assembly	NZ_AAVU00000000	-	GBMF
Microcystis aeruginosa PCC 7806	est. ~4.8	In progress	-	-	Pasteur
Nodularia spumigena CCY9414	est. ~5.32	Draft assembly	NZ_AAVW00000000		GBMF
Nostoc punctiforme PCC 73102	est. ~9.02	Draft assembly	NZ_AAAY00000000	Meeks et al. (2001)	JGI
Nostoc sp. PCC 7120	~6.41	Complete	NC_003272	Kaneko et al. (2001)	Kazusa
Prochlorococcus marinus str. AS9601	~1.67	Complete	NC_008816	-	GBMF
Prochlorococcus marinus str. MIT 9211	est. ~1.84	Draft assembly	NZ_AALP00000000	-	GBMF
Prochlorococcus marinus str. MIT 9215	est. ~1.74	In progress	-	-	JGI
Prochlorococcus marinus str. MIT 9301	~1.64	Complete	NC_009091	-	GBMF
Prochlorococcus marinus str. MIT 9303	~2.68	Complete	NC_008820	-	GBMF
Prochlorococcus marinus str. MIT 9312	~1.71	Complete	NC_007577	-	JGI
Prochlorococcus marinus str. MIT 9313	~2.41	Complete	NC_005071	Rocap et al. (2003)	JGI
Prochlorococcus marinus str. MIT 9515	~1.70	Complete	NC_008817	-	GBMF
Prochlorococcus marinus str. NATL1A	~1.86	Complete	NC_008819	-	GBMF
Prochlorococcus marinus str. NATL2A	~1.84	Complete	NC_007335	-	JGI
Prochlorococcus marinus subsp. marinus str. CCMP1375	~1.75	Complete	NC_005042	Dufresne et al. (2003)	CNRS
Prochlorococcus marinus subsp. pastoris str. CCMP1986	~1.66	Complete	NC_005072	Rocap et al. (2003)	JGI
Prochloron didemni	est. ~5.0	In progress	-	-	TIGR/Utah /UCSD
Synechococcus elongatus PCC 6301	~2.70	Complete	NC_006576	Sugita et al. (2007)	Nagoya
Synechococcus elongatus PCC 7942	~2.70	Complete	NC_007604	-	JGI

Table 2-1. Continued.

Organism	Size (Mb)*	Status	Accession Number	Reference	Sequencing center†
Synechococcus sp. BL107	est. ~2.28	Draft assembly	NZ_AATZ00000000	-	GBMF
Synechococcus sp. CC9311	~2.61	Complete	NC_008319	Palenik et al. (2006)	TIGR
Synechococcus sp. CC9605	~2.51	Complete	NC_007516	-	JGI
Synechococcus sp. CC9902	~2.23	Complete	NC_007513	-	JGI
Synechococcus sp. Eum14	est. ~2.5	In progress	-	-	GBMF
Synechococcus sp. JA-2-3B'a(2-13)	~3.05	Complete	NC_007776	Allewalt et al. (2006)	TIGR
Synechococcus sp. JA-3-3Ab	~2.93	Complete	NC_007775	Allewalt et al. (2006)	TIGR
Synechococcus sp. PCC 7002	est. ~3.2	In progress	-	-	PSU
Synechococcus sp. RCC307	est. ~2.37	In progress	-	-	Genoscope/ Roscoff/Pasteur
Synechococcus sp. RS9916	est. ~2.66	Draft assembly	NZ_AAUA00000000	-	GBMF
Synechococcus sp. RS9917	est. ~2.58	Draft assembly	NZ_AANP00000000	-	GBMF
Synechococcus sp. WH 5701	est. ~3.04	Draft assembly	NZ_AANO00000000	-	GBMF
Synechococcus sp. WH 7803	est. ~	In progress	-	-	Genoscope/ Roscoff/Pasteur
Synechococcus sp. WH 7805	est. ~2.62	Draft assembly	NZ_AAOK00000000	-	GBMF
Synechococcus sp. WH 8102	~2.43	Complete	NC_005070	Palenik et al. (2003)	JGI/NCBI
Synechocystis sp. PCC 6803	~3.57	Complete	NC_000911	Kaneko et al. (1996)	Kazusa
Thermosynechococcus elongatus BP-1	~2.59	Complete	NC_004113	Nakamura et al. (2002)	Kazusa
Trichodesmium erythraeum IMS101	~7.75	Complete	NC_008312	-	JGI
Trichodesmium thiebautii II-3	est. 8.0	In progress	-	-	GBMF

* est., estimated size for genomes not completely finished.

† ASU, Arizona State University; GBMF, Gordon and Betty Moore Foundation Marine Microbiology Initiative; JGI, DOE Joint Genome Institute; Kazusa, Kazusa DNA Research Institute; Pasteur, Institut Pasteur; CNRS, the Centre National de la Recherche Scientifique; TIGR, The Institute for Genomic Research; Utah, University of Utah; UCSD, University of California, San Diego; Nagoya, Nagoya University, Japan; PSU, Penn State University; Genoscope, Genoscope - the French National Sequencing Center; Roscoff, Roscoff Center for Oceanographic Studies; NCBI, the National Center for Biotechnology Information.

Octopus Spring cyanobacterial mat in Yellowstone National Park (145, 146). Thus, *Synechococcus elongatus* strains PCC 6301 and PCC 7942 are the only two mesophiles of fresh water *Synechococcus* species to have completely sequenced genomes. These two strains are closely related, with a single inversion accounting for the few detectable RFLPs between *S. elongatus* PCC 6301 and PCC 7942 (6). At the time we started the genome project of *S. elongatus* PCC 7942, little was known about the structure, genetic organization, and sequence of the genome, except the physical restriction map of *S. elongatus* PCC 6301 (147). Around 130 and 200 individual sequence submissions for *S. elongatus* PCC 6301 and PCC 7942, respectively, have been deposited in GenBank. Taken together, less than 15% of both genomes was sequenced through individual efforts. Even though these two *S. elongatus* strains are very closely related (~99.93% in their nucleotide sequence identity), there is at least one major difference in that *S. elongatus* PCC 7942 is naturally transformable, while *S. elongatus* PCC 6301 is not. The availability of the complete sequence of *S. elongatus* PCC 7942, together with the recently published complete genome of *S. elongatus* PCC 6301 (7), provides the possibility in determining the genetic loci that account for this major difference.

Here, we reported the complete nucleotide sequence of *S. elongatus* PCC 7942 genome determined through a combination of transposon-mediated sequencing and shotgun sequencing by the DOE Joint Genome Institute (JGI). A portion of the automated annotation of the genome has been manually curated, and a system for community refinement of the annotation was established. Organization and characteristic features of the genome are described.

Results and Discussion

Whole genome sequence determination of *Synechococcus elongatus* PCC 7942

Because *S. elongatus* PCC 7942 is naturally transformable and mediates homologous recombination with high efficiency, we proposed a transposon-mediated mutagenesis and sequencing strategy to determine the sequences surrounding transposon insertions in essentially every *S. elongatus* PCC 7942 gene (65). Because the transposons we use insert almost randomly into the genome, the sequence surrounding the insertion sites, which can be determined by sequencing outward from transposon end primers, should provide widespread coverage of the whole genome. Transposon-mediated *in vitro* mutagenesis and sequencing was conducted, starting with a 960-cosmid genomic library. For each cosmid, hundreds of raw sequences were polished to remove transposon sequences and assembled into contigs, which were then orientated and connected using a "primer walking" method. In total, 9 cosmids were completely sequenced in this manner and manually annotated (see Materials and Methods), and 8 of them, including a cosmid that carries the large plasmid pANL, were deposited into GenBank (Table 2-2).

In November 2004, JGI completed a shotgun sequencing project of the *S. elongatus* PCC 7942 genome, for which our group is the collaborating laboratory. JGI sequences were created from 3-kb and 8-kb plasmid libraries and a fosmid library, providing 8-10 fold of coverage of the genome. The JGI assembly of sequences, including shotgun sequences and our transposon-mediated sequences that we provided to JGI, was sent to us as output files from the Phred/Phrap program (148, 149). The total

length of the *S. elongatus* PCC 7942 genome is 2,695,903 bp, slightly smaller than that of *S. elongatus* PCC 6301 (2,696,255 bp).

Annotation of the genome for protein- and RNA-encoding loci

Coding regions of the complete genome sequence were assigned by two independent computer predictions and currently are being subjected to manual refinement by a community oversight group.

Automated annotation of the genome was performed by Computational Biology at ORNL (DOE Oak Ridge National Laboratory) for JGI as the final step of the sequencing project. Because ORNL does not provide manual annotation support, we also submitted the genome sequence to the Annotation Engine of TIGR (The Institute for Genomic Research) for computer prediction. The automated annotation in the format of a MySQL database was then installed into a Linux-based manual annotation server (located at Laboratory for Functional Genomics, Department of Biology, TAMU) with web interface, Manatee, which was created by the bioinformatics department at TIGR.

ORNL combines three modeling programs for gene finding and function assignment: Generation, Glimmer, and CRITICA. Generation, courtesy of the Genome Informatics Corporation (Genomix, Oak Ridge, TN), predicts coding regions using predominantly 6-mer probabilities estimated from training sequences of the microbial organism. Glimmer (Gene Locator and Interpolated Markov ModelER) (150-152) uses a combination of Markov models from 1st- through 8th-order (Interpolated Markov Models/IMMs) to identify coding regions, which is more flexible and accurate than

Table 2-2. Sequenced and annotated cosmids of *S. elongatus* PCC 7942.

Cosmids	Size (bp)	Putative ORFs	Accession number	G+C (%)	Published genes or features
pANL (2A8)	46,366	58	AF441790	52.84	*srp* genes
7H1 & 2E8	60,090	58	U30252	54.87	*clpP2clpX*, NS1
3E9	42,558	48	X04616	55.76	*psbAI*
4G8	31,404	39	AY157498	54.48	*trxM*
6C3	38,188	32	AY120852	55.79	*rpsA*
7G3 & 8E10	74,355	88	AY120853	55.12	*kaiABC*
8D8I	707	2	-	56.01	-
8D8III	12,421	12	-	58.00	*psbD1psbC*

fixed-order Markov models. CRITICA (Coding Region Identification Tool Invoking Comparative Analysis) (153) is a microbial gene finder that combines comparative blastn alignments and dicodon (hexanucleotides) frequency statistics to recognize coding sequences. The completed computational annotation from ORNL has been submitted by JGI to GenBank with the accession number NC_007604.

Glimmer is also the primary microbial gene finder used at TIGR, where it was first developed. Each putative protein is then searched against an internal non-identical amino acid database (niaa) containing all available proteins. A modified BLAST search algorithm, BLAST-Extend-Repraze (BER), is employed for identifying potential frameshifts or point mutations in the sequence. This program combines a standard BLAST search and a modified Smith-Waterman alignment (154) on coding regions plus the sequences that extend 300 nucleotides upstream and downstream. All putative proteins are also searched against profile hidden Markov models (HMMs) (155), which are generally more sensitive and accurate than pairwise alignments. AutoAnnotate, a

computer program developed at TIGR, then analyzes the BER and HMM search results and assigns gene features, such as common name, gene symbol, Enzyme Commission (EC) number, TIGR role, and Gene Ontology (GO) terms, automatically when possible, annotating from the best piece of evidence available to it.

General results of automated annotation of *S. elongatus* PCC 7942 are summarized in Table 2-3, with *S. elongatus* PCC 6301 annotation as a reference. The difference in total number of coding regions between the ORNL (2662) and TIGR (2906) versions is mainly due to TIGR's effort in identifying small ORFs, usually less than 100 bp, which were skipped in ORNL annotation as a default setting. There is little difference between these two annotations in percentages of ORFs with or without function predictions. *S. elongatus* PCC 6301 has only 2578 coding regions annotated, including 51 RNA loci (7). One tRNA(Leu)(UAA) gene, *trnL,* which carries a 240-bp group I intron (156, 157), was missed in both *S. elongatus* PCC 7942 automated annotations. Please note that these three annotations each uses a different starting nucleotide, which complicates comparison.

Manual curation of the automated annotations on the Manatee server is mainly based on TIGR annotation. Once the MySQL database produced from the Annotation Engine is installed, annotators can get access to Manatee remotely to scrutinize stored data and modify annotation in a user-friendly, browser-based interface. For each gene, all evidences that were used in the automated annotation are shown on the "Gene Curation Page," which is also the main interface for manual annotation. The pipeline of manual curation of a gene is listed below: ***1)*** Look for the HMM section and the

Evidence Picture for the best HMM hit with a cutoff score above the trusted value for the whole protein or individual domains; ***2)*** Look at BER searches for the characterized match or highly similar homologues, as well as frameshifts or point mutations, if applicable; ***3)*** Look at other evidences (TmHMM, SignalP, Prosite, etc) for suggestions, such as *trans*-membrane domains; ***4)*** Look at gene context by examining the orientation and function of upstream and downstream genes in the "Genome Viewer" window; ***5)*** Suggest the protein name, gene symbol, EC#, and comments based on all the information listed above; ***6)*** Suggest Gene Ontology (GO) terms or make corrections on those already suggested by programs; ***7)*** Review TIGR roles that were automatically assigned; ***8)*** Check the translation start site in a separate window using alignment information of BER search results; ***9)*** Submit all curated information to the database. It is suggested to perform new BLAST searches for new protein homologues and conserved domains for each gene because the database has not been updated since July 2005. Another main caveat of operating a local Manatee server is that it cannot export any changes in annotation automatically.

We set up a small annotation consortium involving several experts in different fields of cyanobacterial research. So far, 129 genes (~5% of total genome) in several categories have been manually checked. Signal transduction genes, mainly two-component and one-component systems, have been identified and archived in the MiST (Microbial Signal Transduction) database (159, 160). There are 43 two-component proteins (21 systems) and 63 one-component proteins in the genome of *S. elongatus* PCC 7942, including one pair of two-component system proteins from pANL.

Table 2-3. Summary of *Synechococcus elongatus* genomes.

	PCC 7942 (TIGR)*		PCC 7942 (ORNL)†		PCC 6301†	
	Number	**% of Total**	**Number**	**% of Total**	**Number**	**% of Total**
Total DNA (bp)	2695903	100	2695903	100	2696255	100
Coding DNA (bp)	2416176	89.62	2408946	89.36	2377899	88.19
G+C content	-	55.47	-	55.47	-	55.48
Total genes	2906	100	2662	100	2578	100
Protein genes	2856	98.28	2612	98.12	2527	98.02
RNA genes	51	1.88	50	1.88	51	1.98
rRNA genes	6	0.23	6	0.23	6	0.23
tRNA genes	45	1.65	44	1.65	45	1.75
Genes with function prediction	1682	57.88	1549	58.19	1369	53.10
Genes without function prediction	1174	40.40	1063	39.93	1158	44.92

∗ Summarized from local Manatee server for *S. elongatus* PCC 7942 TIGR annotation. The numbers of genes are subject to change according to the progress of the manual curation. RNA genes have been curated.

† Modified from the Integrated Microbial Genomes (IMG) system of JGI (158).

General organization and characteristic features of the *Synechococcus elongatus* PCC 7942 genome

The genome of *S. elongatus* PCC 7942 consists of a circular chromosome (~2.7 Mb), an essential large plasmid (pANL, ~46.3 kb), and a non-essential small plasmid, pANS (~8.6 kb). Because the plasmids will be described in Chapter V, here we focus on the chromosome. The chromosome is 2,695,903 bp in length with a G+C content of ~55.47%. According to partially curated TIGR annotation, there are 2,856 protein-coding genes and 51 RNA coding loci, including 6 rRNA genes in 2 operons and 45 tRNA genes. The coding regions represent ~90% of the genome and the average length of protein coding genes is 846 bp. Around 60% of protein-coding genes have been assigned with predicted functions basing on their closest homologues or conserved domains. Others are either conserved hypothetical proteins or hypothetical proteins with no significant homologous hit in the database (Table 2-3). Unless otherwise specified, putative ORFs in this chapter are from the TIGR annotation and are temporarily named with ORF0XXXX. The first nucleotide of the genome sequence was set to be the first G of a *Bam*HI site (GGATCC) around 100 bp upstream of the coding region of *typA*, a predicted membrane GTPase involved in stress response. Thus, *typA* is the first gene in the genome and named ORF00001. Because *typA* is located in the region (~190 kb) that is inverted between *S. elongatus* PCC 7942 and PCC 6301, the majority of the two genomes outside the inversion could be exactly aligned with each other for convenience of genome comparison. In the ORNL annotation, however, the first nucleotide is assigned to the *dnaN* locus encoding the *beta* subunit of DNA polymerase III

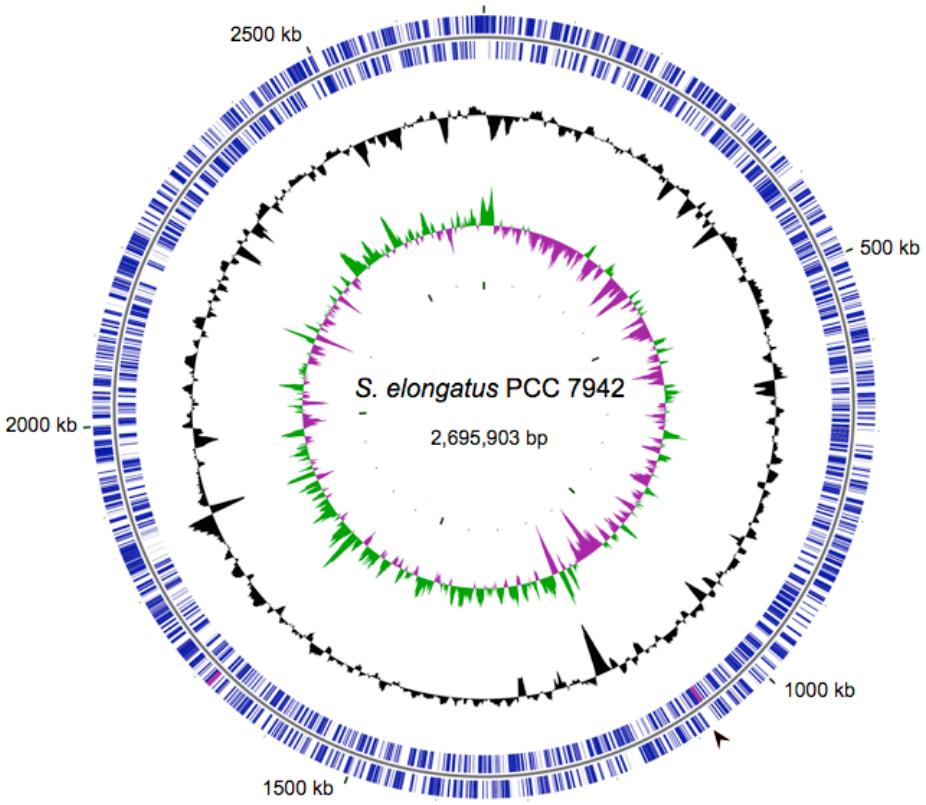


Fig. 2-1. Circular representation of the *S. elongatus* PCC 7942 genome. Genomic features are shown concentrically. Shown as circles from the outermost to the innermost: putative Open Reading Frames (ORFs) in blue, GC content in black, GC skew ([G-C]/[G+C]) in green and purple, and the scale in kb. Clockwise ORFs are shown outside the base line, while ORFs expressed in the counterclockwise direction are inside. The rRNA operons are shown in red. For the second circle, peaks outside the centerline correspond to regions that have a GC content of above the genome average (0.5547), while the peaks pointing inside refer to the regions with a below average GC content. The third circle shows GC skew values that are greater than the genome average (0.0010) in green, whereas GC skew values less than the average are in purple. The black arrow outside the ORF circle indicates the putative replication origin. The figure was generated using the CGVIEW program (161) with default settings. Plotting GC content and GC skew used a window size of 10000 and a step of 100.

(ORF01187). Thus, ORF00001 (*typA*) from the TIGR annotation is designated synpcc7942_1525 in the ORNL annotation, which is adopted by both the IMG and GenBank databases. The replication origin of the chromosome has been suggested to be located at this *dnaN* locus (@~1,113 kb), which contains 11 DnaA boxes (TTTTCCACA) (162). The cumulative GC skew analysis also predicted a putative *oriC* close to this low G+C content region (Fig. 2-1).

Neutral Sites for gene transfer through homologous recombination

Neutral sites are recombination loci on the chromosome of *S. elongatus* PCC 7942 with no apparent growth defect when disrupted by insertion of exogenous DNA fragment (163). Two neutral sites are currently being used in our lab as cloning platforms: neutral site I (NS1 @ ~995 kb from TIGR start site) and neutral site II (NS2 @ ~1,195 kb) (Fig. 2-2). Foreign DNA sequence is first cloned into an NS vector and characterized in *E. coli* prior to transfer to the cyanobacterium. An NS1 vector contains ~1 kb upstream and downstream of an *Xho*I site at the C-terminus of ORF01066; an *omega*-cassette inserted in the *Xho*I site, carries an antibiotic-resistance gene as selection maker followed by transcription terminator sequences. After a gene of interest is inserted into a cloning site near the antibiotic marker gene, the final construct is then introduced into a wild-type strain of *S. elongatus* PCC 7942, which favors homologous double recombination. Because there is no *Synechococcus*-compatible replication origin present on the NS1 vector, only an integrated copy of an exogenous DNA sequence can survive, which ensures just one copy per chromosome.

NS1 is at the locus of ORF01066, which encodes a Band_7_flotillin domain (cd03399). The flotillin protein is an integral membrane protein probably involved in signal transduction and many other membrane-associated functions (164). A possible reason why there is no phenotype for disruption of ORF01066 is that there is another copy of a Band_7_flotillin domain gene localized just upstream of it, ORF01067. They encode proteins that are similar in size and share ~36% identity and ~61% similarity in their amino acid sequences, suggesting functional redundancy. Even though these two genes very likely form a dicistronic operon, the insertion in the *Xho*I site lies very close to the C-terminal coding region of ORF01066 and probably does not affect the expression of the upstream ORF01067. The insertion in *Xho*I site also is not likely to affect upstream genes, which are in the opposite direction. Downstream there are several genes in the same orientation as flotillin domain genes, including *ndk* and a five-gene cluster encoding subunits of a branched-chain amino acid ABC-type transporter (*livK*, *livH*, et al; Fig. 2-2A). Because *ndk* overlaps *livK*, it is likely that these six genes form an operon. Preliminary data, which was done by other people in the lab, showed that disruption of *ndk* did not have a circadian clock phenotype (personal communications).

The construction of NS2 vectors is similar to that of NS1 vectors. There are two different variants: NS2.1 and NS2.2, each of which uses a different restriction site in the NS2 locus. The insertion site for NS2.1 vectors is a *Bst*EII site in the middle of ORF01270, encoding a conserved hypothetic protein. There are no paralogues of this protein encoded by the *S. elongatus* PCC 7942 genome. However, the absence of an ORF01270 homologue from all published *Prochlorococcus* species and the genome of

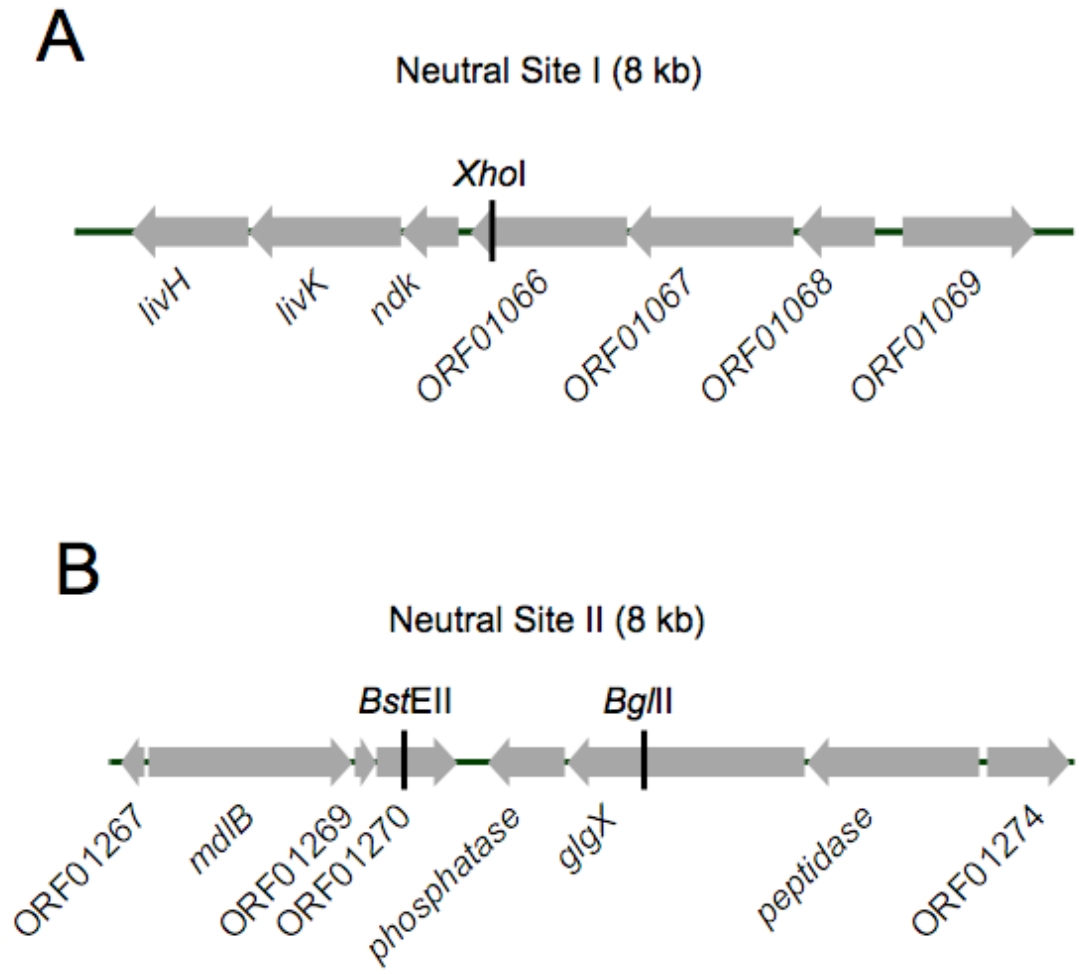


Fig. 2-2. Representation of the neutral site regions showing the relative positions of restriction sites for vector construction. ***A)*** Neutral Site I (NS1). ***B)*** Neutral Site II (NS2). Putative ORFs are shown as grey boxes with arrowheads; black vertical bars are restriction endonuclease sites.

T. elongatus BP-1 is consistent with a gene that is not essential. The upstream *mdlB* gene (ORF01268, ATP-binding protein of ABC transporter) has a paralogue in the genome. The other upstream gene, ORF01269, is a hypothetical small protein with no significant hit in the GenBank protein database. NS2.2 vectors are based on a *Bgl*II site located at the C-terminal coding region of ORF01272, encoding a putative glycogen debranching enzyme that is conserved in all published cyanobacterial genomes. There is a remote homologue of ORF01272 protein in *S. elongatus* PCC 7942, ORF02386, which encodes a 1,4-alpha-glucan branching enzyme. Both enzymes contain a glycogen branching enzyme N-terminal domain (cd02855) and an *alpha*-amylase catalytic domain (pfam00128). There might be some overlapping function between these two proteins. Both the upstream peptidase gene and downstream phosphatase gene have at least one remote but significant paralogue in the genome (data not shown).

Repetitive sequences and transposable elements

The absence of long repetitive sequences in the *Synechococcus* genus has been previously suggested based on the thermal kinetics analysis of DNA renaturation (144). There are dozens of repetitive sequences in the genome, identified by self-alignment of the genome sequence, with a size ranging from 200 bp to 5.5 kb. All of them are located in the coding regions, such as the RNA operons, *psbA* genes, or *clp* genes (data not shown). One small repetitive sequence, with a size of 100-150 bp, has at least 23 copies in the chromosome, usually in the intergenic regions between two head-on-head ORFs.

There is a potential small protein (<40 aa) encoded in the repeats that has no sequence similarity to any other proteins in the databases (data not shown).

An octameric highly iterated palindrome (HIP1), GCGATCGC, is over-represented in DNA sequences of many cyanobacterial strains, including *Synechococcus* species (165, 166). It has been suggested that HIP1 sites participate in homologous recombination and other sequence rearrangement events in cyanobacterial genomes (165, 167). There are 7,402 HIP1 site in the chromosome of *S. elongatus* PCC 7942, approximately 1 site per 364 bp. The distribution of HIP1 sites in the chromosome is roughly random, with a few hot spots and blank regions. Preliminary analysis has identified 14 relatively large regions, over 3 kb in size, with no HIP1 site. These include RNA operons, long essential genes, and several low G+C content segments (data not shown). It is obvious that there should be fewer HIP1 sites in low G+C content regions since the HIP1 site contains 75% G+C. However, the complete absence of HIP1 sites in those low G+C regions suggests that those sequences are likely the consequences of recent lateral gene transfer events.

No insertion sequence (IS) has been identified and there are only one putative transposase and two *pseudo*-transposases present in *S. elongatus* genomes (7). In comparison, *Thermosynechococcus elongatus* BP-1, with a similar genome size (~2.59 Mb), has 82 transposase genes and 70 IS loci (168). The strikingly "clean" feature of *S. elongatus* genomes may result from a lack of rearrangement of the chromosome mediated by insertion sequences or other transposons. It is interesting to note that there are only 3,681 HIP1 sites in the genome of *T. elongatus* BP-1, averaging one site per 705

bp, which is about one half of the HIP1 site density in *S. elongatus* strains; this correlation between low density of HIP1 sites in cyanobacterial species and higher numbers of transposase genes and insertion sequences has been reported previously (7). Thus, there is probably a balance between the number of HIP1 site and the number of mobile elements in these cyanobacterial genomes. *S. elongatus* strains may use HIP1 sites for rearrangement of the genome, while others use both HIP1 site and transposons. More genomes need to be checked for supporting evidence. The questions about the mechanism in *S. elongatus* strains for transposon elimination and whether the genomes of *S. elongatus* strains are more stable or less stable than other cyanobacterial strains also need to be further explored.

Global comparison of *Synechococcus elongatus* PCC 7942 and PCC 6301 genomes

The genome of *S. elongatus* PCC 6301 is 2,696,255 bp in length, slightly larger (352 bp) than that of *S. elongatus* PCC 7942 (2,695,903 bp). The overall sequence of these two strains shows 99.93% identity. The previously identified large inversion between the two genomes is ~188.6 kb long, flanking by genes for two porin-like proteins and a pair of 20-bp inverted repeats (7). There are two other small regions in PCC 6301 that are missing in the PCC 7942 genome. The deletion of the larger one, a 243-bp *Pvu*I fragment, results in fusion of syc1650_d (*apt*, encoding adenine/guanine phosphoribosyltransferase) and syc1651_d (oligopeptides ABC transporter permease protein) in PCC 6301 into one gene, ORF01000, in PCC 7942. Both syc1650_d and syc1651_d are well conserved in other cyanobacteria as two separate ORFs: the fusion

protein is only present in PCC 7942. A pair of 12-bp direct repeats TGGCCG*CGATCG*/ TGGCCT*CGATCG*, each carrying a *Pvu*I site, is present at the ends of the fragment, whereas only one copy is present in PCC 7942. Thus, it is likely that a recombination event occurred between the direct repeats in an ancestor shared with PCC 6301, which resulted in the deletion of the 243-bp fragment in PCC 7942. The other deletion is a 56-bp fragment located at the N-terminus of syc0635_c, a hypothetic protein (152 aa) only found in *S. elongatus* species. The deletion in PCC 7942 causes only the C-terminal 70 aa to be encoded. The 56-bp fragment in PCC 6301 is adjacent to a ~1.6 kb region with ~44% G+C content, much lower than the overall 55.5% of the chromosome. The three putative ORFs encoded in this region are also *S. elongatus* specific, not found in any other sequences in the GenBank.

Chromosomal toxin-antitoxin systems

Toxin-antitoxin (TA) systems are widely distributed in plasmids and chromosomes of many free-living bacteria with functions including plasmid maintenance, programmed cell death, or stress response (169, 170). There are seven typical TA families: *ccdAB, mazEF (pemIK), vapBC, phd/doc, parDE, and higBA* (169, 171). Usually the first gene in the TA operons encodes the antitoxin, such as *ccdA* and *vapB*, while the downstream gene encodes the cognate toxin. One exception is the *higBA* operon, in which *higB* is the toxin and *higA* is the antidote.

Multiple copies of TA cassettes have been identified in cyanobacterial genomes: 27 in *Nostoc* PCC 7120 and 13 in *Synechocystis* PCC 6803 (171). There are also several

solitary TA loci in each genome. None of these paired or orphan TA genes belongs to either the *parDE* or *ccdAB* families. There are no TA systems identified in *Prochlorococcus marinus* species and *T. elongatus* BP-1. TA cassettes in *S. elongatus* PCC 7942 were identified based on conserved domains, gene context, and sequence similarities to known addiction genes of other cyanobacterial strains. As shown in Table 2-4, 7 pairs of TA cassettes and 2 solitary TA genes were found in the genome. They belong to *vapBC* (8 genes), *higBA* (2 genes), *phd*/*doc* (1 gene), *relBE* (1 gene), and *parDE* (1 gene) families.

In contrast to TA cassettes in *Nostoc* PCC 7120 and *Synechocystis* PCC 6803, no *mazEF* system was identified on the chromosome of *S. elongatus* PCC 7942 and one *parE*-family toxin is present. In addition, there is a copy of *pemIK* system and another copy of a *vapBC* cassette on the large plasmid, pANL. Among chromosomal systems, 2 of them are pure *vapBC* pairs (ORF00019/ORF00020 & ORF00872/ORF00873). Three other pairs (ORF01519/ORF01520, ORF02520/ORF02519, and ORF02523/ORF02524) have conserved toxin domains, but no recognizable antitoxin domains. These antidote genes were mainly determined based on their size (~80 aa) and their context (upstream of and overlap with or are very close to cognate toxin genes). The remaining 2 pairs are hybrid TA systems: ORF01779 encodes a *higBA* family antitoxin, while its cognate toxin encoded by ORF01778 belongs to the *vapBC* family. The antitoxin encoded by ORF02516 for a downstream PIN domain toxin, product of ORF02515, carries a Phd-like domain. Thus, a composite and modular mechanism, in which similar antitoxins can be assembled with different toxins, has been adopted for cyanobacterial toxin-antitoxin

systems as previously described for a toxin-antitoxin system on the *Bacillus thuringiensis* plasmid pG11 (172). The last three TA cassettes are neighboring genes, comprising six consecutive pairs of ORFs with different orientations in this 4-kb region. Both the size of encoded proteins and gene organization within the pairs indicate the possibility of TA cassettes. However, only three of them, listed in the Table 2-4, contain conserved domains characteristic of a toxin or antitoxin. It is interesting to note that these genes are just two ORFs away from the *kai* locus, which is in the complementary orientation. However, this region is not conserved in any other published cyanobacterial genome. There are many other gene pairs in the genome of *S. elongatus* PCC 7942 with the same structure feature and gene size as those TA cassettes mentioned above but with no obvious functional assignment. They are likely candidates for novel TA systems. Thus more TA cassettes may be present in cyanobacterial genomes than we know.

Proteins and conserved domains involved in signal transduction and photoreception

Signal transduction pathways are vital in sensing signals inside and outside cells and controlling cellular activities. Two-component systems, consisting of a sensor histidine protein kinase and its cognate response regulator, are widely distributed among living organisms (173). More signal transduction activities in prokaryotes, however, are transmitted through one-component proteins carrying both sensing and effector domains or a single domain with both sensing and effector functions (159). The numbers of signal transduction proteins vary widely among cyanobacterial genomes, which is probably

Table 2-4. Putative toxin-antitoxin cassettes in *S. elongatus* PCC 7942.

TIGR ID	JGI/GenBank ID	Size (aa)	Conserved domain	Function	TA family
ORF00013	Synpcc7942_1537	69	COG1487 (VapC)	Toxin	VapBC
ORF00019	Synpcc7942_1544	83	COG4691 (StbC)	Antitoxin	VapBC
ORF00020	Synpcc7942_1545	151	COG1487 (VapC)	Toxin	VapBC
ORF00872	Synpcc7942_2319	72	COG3093 (VapI)	Antitoxin	VapBC
ORF00873	Synpcc7942_2320	95	COG1848 (PIN)	Toxin	VapBC
ORF01519	Synpcc7942_0305	76	n.i.	Antitoxin	n.i.
ORF01520	n.i.*	96	COG3668 (ParE)	Toxin	ParDE
ORF01640.5	Synpcc7942_0409	94	pfam01402 (HTH_4)	Antitoxin	HigBA
ORF01778	Synpcc7942_0540	141	COG1569 (PIN)	Toxin	VapBC
ORF01779	Synpcc7942_0541	76	pfam01402 (HTH_4)	Antitoxin	HigBA
ORF02515	Synpcc7942_1203	137	COG4113 (PIN)	Toxin	VapBC
ORF02516	Synpcc7942_1204	104	COG4118 (Phd)	Antitoxin	Phd/Doc
ORF02519	Synpcc7942_1207	87	pfam06769	Toxin	n.i.
ORF02520	Synpcc7942_1208	84	COG2161 (StbD)	Antitoxin	RelBE
ORF02523	n.i.	74	n.i.	Antitoxin	n.i.
ORF02524	Synpcc7942_1213	138	COG1487 (VapC)	Toxin	VapBC

*n.i., not identified.

related to the genome size and living environment. According to the MiST database, there are 214 two-component proteins and 140 one-component proteins in *Anabaena* ATCC 29413, which contains the most signal transduction proteins in sequenced cyanobacteria, while the leanest example is *Prochlorococcus* MIT 9515, with only 9 two-component and 10 one-component proteins.

The MiST database displays 61 one-component signal transduction proteins encoded by the chromosome of *S. elongatus* strains. Many of these proteins are transcription regulators that belong to ArsR, Crp, GntR, LuxR, LysR, MerR, PadR, Rrf2, TetR, and XRE families. Some of them carry nucleotide-binding domains, such as cNMP (cyclic nucleotide-monophosphate binding domain), HD (metal dependent phosphohydrolase domain), CYCc (Adenylyl/Guanylyl cyclase, catalytic domain), CBS

(a domain shown to bind ligands with an adenosyl group such as AMP, ATP and S-AdoMet), as well as two well-known GGDEF (Gly-Gly-Asp-Glu-Phe, diguanylate-cyclase) and EAL (Glu-Ala-Leu, cyclic diguanylate-specific phosphodiesterases) domains, both participating in turnover of cyclic diguanosine monophosphate (C-di-GMP), a widespread intracellular second messenger in bacteria (174, 175). In *S. elongatus* PCC 7942, there are 15 GGDEF domains and 8 EAL domains. Both domains are always present as single copy in proteins. Among EAL domains, 7 of them are present in GGDEF-carrying proteins, suggesting coupled functions in cyclic diguanylate metabolism. Most of the 15 GGDEF-carrying proteins also have signal sensor domains, *e.g.*, MASE1 (integral membrane sensory domain), CHASE2 (extracellular sensory domain), 7TMR-HDED (extracellular domain of the 7TM Receptors with HD hydrolase), as well as famous PAS (Per-ARNT-Sim) and GAF (cyclic GMP, Adenylyl cyclase, FhlA) domains. GAF binds a chromophore in phytochromes and cGMP in some phosphodiesterases, while PAS domains bind small ligands and act as sensors for light and oxygen in many signaling proteins. There are 10 GAF and 11 PAS domains present in one-component signal transduction proteins. Only two of the PAS domain proteins also have GAF. Some proteins carry multiple GAF and/or PAS domains. There are also several proteins harboring domains involved in serine/threonine protein kinase or phosphatase activities, such as S_TKc and PP2C_SIG. Five serine/threonine protein kinases were identified in the genome of *S. elongatus* PCC 6301 (7) and PCC 7942 as well.

The annotation of PCC 6301 shows 37 two-component signal transduction proteins, which includes 13 histidine protein kinases (HPK), 21 response regulators (RR), and 3 hybrid sensory kinases (7). The MiST database (160), however, shows different numbers: 11 HPKs, 20 RRs, and 5 hybrid kinases in both *S. elongatus* strains. This difference is probably due to different sensitivity criteria applied for domain identification. In addition to HPKs and RRs, MiST lists 3 CheW single-domain proteins and 2 MCPsignal (methyl-accepting chemotaxis protein signaling) domain proteins that primarily function in chemotaxis signaling. Only five pairs of HPK and RR are clustered in likely operons. Two of the 20 RRs are essential for viability and another one, RpaA, has been found to affect global circadian transcription dramatically and function downstream of its cognate histidine kinase SasA (61). Many of the two-component proteins, HPK or RR or hybrid, also contain other signal transduction domains. GGDEF is present in two response regulators, and one of them carries an EAL domain and three PAS domains. The latter is also associated with three HPKs. GAF domains are in three HPK or hybrid kinases, including clock input kinase CikA (176). Other domains present in two-component signal transduction proteins include GerE (LuxR family regulatory domain), HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases domain), and HPT (histidine-containing phosphotransfer domain).

Many signal transduction components are involved in direct or indirect perception of light in photoautotrophic bacteria. There are three families of flavin-binding blue light receptor domains: BLUF (Blue Light Using FAD), PhrB-like (photolyase-like domain in cryptochromes; binds FAD non-covalently), and LOV (Light

Oxygen Voltage; binds FMN non-covalently) (177). In cyanobacteria only *Synechocystis* PCC 6803 and *T. elongatus* BP-1 are known to encode a BLUF domain protein. Cryptochromes are involved in circadian clock functions in plants and insects (178). Cryptochrome has also been found in some cyanobacteria genomes, such as *Synechocystis* PCC 6803, in which there are two proteins carrying a PhrB-like domain: one is a genuine photolyase and the other one is more similar to eukaryotic cryptochromes (179). There is only one copy of photolyase (ORF01300, 484 aa) encoded in *S. elongatus* genomes (180, 181), which has very low similarity to *Synechocystis* cryptochrome. However, there is another protein (ORF00379, 293 aa) in the genome that carries a copy of the FAD_binding_7 domain that is located in the C-terminal part of the DNA photolyase. This protein is conserved in many, but not all, cyanobacteria genomes as a single copy, but its function is unknown.

LOV domains are a subfamily of the PAS superfamily. As a blue-light sensor, LOV is coupled to many different signaling domains (182). About 10 residues, including the photoactive cysteine, directly interact with the FMN cofactor and those are well conserved among LOV domains. The photoactive function and FMN binding ability of some cyanobacterial LOV domains have been experimentally confirmed (183). There are at least two potential LOV domains in *S. elongatus* PCC 7942. They are located in ORF01382, which encodes a LOV/GGDEF/EAL protein (578 aa), and ORF02678, which encodes a REC/PAS/PAS/LOV/GGDEF/EAL protein (929 aa); downstream of ORF02678 are a putative phytochrome (ORF02680) and its cognate response regulator

(ORF02679). These two LOV domain proteins, ORF01382 and ORF02678, are very likely the blue light receptors in *S. elongatus*.

As mentioned above, many conserved domains are involved in photoreception and signal transduction: BLUF, LOV/PAS, and GAF for light signal perception, as well as coupled signal output domains, such as GGDEF, EAL, MCP, HPK, and serine/threonine protein kinase domains (184). Among conserved photosensory domains present in cyanobacterial proteins, GAF and LOV/PAS are well represented in *S. elongatus* PCC 7942, but no BLUF or cryptochrome are identified. All output domains are present in *S. elongatus* PCC 7942. Photosensory proteins in cyanobacteria, harboring these light signal input and output domains, are important for cellular functions like photosynthesis and light adaptation, as well as circadian clock. Among these signal input domains, GAF is of particular interest because of its presence in circadian input kinase CikA. However, the GAF domain in CikA lacks the conserved cysteine or histidine residue that serves for covalent binding of chromophore, such as bilin compound (176). It has been shown that CikA does not bind bilin chromophore *in vivo* and the GAF domain positively regulates kinase activity (71). Thus, it is possible that no cofactor is needed for CikA function. Protein-protein interaction may induce conformational change of the GAF domain, which in turn activates the histidine kinase domain (72). There are 22 GAF domains identified in 16 proteins encoded on the PCC 7942 chromosome. Only 6 GAF domains have the conserved cysteine residue for bilin binding. Five of them in a protein encoded by ORF02133 that also carries a C-terminal methyl-accepting chemotaxis domain, and the other one is in GafA (encoded by ORF01104) that

interacted CikA in yeast two-hybrid assay but with no circadian phenotypes (S.R. Mackey et al, PNAS, in press). The relationship of these genuine or *pseudo* GAF domains to circadian clock function will be tested in insertional mutants.

Genes involved in circadian clock function

S. elongatus PCC 7942 is a model organism for cyanobacterial circadian clocks, and several loci involved in clock function have been identified, including *kaiABC* (57), a gene cluster that encodes components of the central oscillator; input pathway components *cikA* (58) and *ldpA* (59) and *pex* (62); and key output pathway components, sensor kinase gene *sasA* (60) and its cognate response regulator gene *rpaA* (61); as well as *cpmA* (63), *labA* (64), *clpP2clpX* locus (65), and group 2 sigma factors (66, 67).

The KaiABC proteins comprise the circadian pacemaker of PCC 7942 (57). As a dicistronic operon, *kaiBC* is conserved in all but one sequenced cyanobacterial genome and many other prokaryotes, including many *Archaea* species (106). The cyanobacterial exception is *Gloeobacter violaceus* PCC 7421, which also has no thylakoid membranes. KaiB (ORF02528, 102 aa) defines a KaiB-like domain (cd02978), which is remotely related to bacterial thiol-disulfide isomerase or thioredoxin at the protein sequence level. KaiC (ORF02527, 519 aa) shows a duplicated structure, carrying a RecA-like ATPase domain (COG0467) in each half. As shown in Table 2-5, most sequenced cyanobacterial genomes encode only one copy each of KaiB and KaiC. However, four species have more: *Crocosphaera watsonii* WH 8501 has two *kaiBC* operons; *Cyanothece* sp. CCY0110 contains two *kaiBC* operons and an extra paralogue of *kaiB*; *Lyngbya* sp. PCC

Table 2-5. Distribution of KaiABC proteins in sequenced (complete or draft) cyanobacteria genomes.

Organism*	KaiA			KaiB			KaiC		
	GenBank ID	Size (aa)	Positives §	GenBank ID	Size (aa)	Positives	GenBank ID	Size (aa)	Positives
ATCC 29413	Ava_1020	89	65/84	Ava_1017	108	92/97	Ava_1016	519	462/504
WH 8501	CwatDRAFT_4942	309	187/290	CwatDRAFT_4943	104	94/98	CwatDRAFT_4944	519	465/503
	-	-	-	CwatDRAFT_5153	94	65/86	CwatDRAFT_5154	504	363/503
CCY0110	CY0110_21180	282	178/269	CY0110_21185	104	94/98	CY0110_21190	495	457/495
	-	-	-	CY0110_20660	94	64/86	CY0110_20655	504	359/498
	-	-	-	CY0110_06839	98	56/88	-	-	-
PCC 7421	n.i.†	-	-	n.i.	-	-	n.i.	-	-
PCC 8106	L8106_06499	278	184/279	L8106_06504	104	93/99	L8106_06509	522	479/516
	-	-	-	L8106_27062	103	65/91	L8106_27067	575	283/505
	-	-	-	L8106_27057	110	56/86	L8106_18916	485	228/493
CCY9414	N9414_02386	101	75/97	N9414_02381	104	90/95	N9414_02376	521	469/515
PCC 73102	Npun02006492	193	97/165	Npun02006491	104	92/97	Npun02006490	520	466/515
PCC 7120	alr2884	102	76/95	alr2885	108	92/97	alr2886	519	462/504
AS9601	n.i.	-	-	A9601_15441	105	92/100	A9601_15431	509	437/489
MIT 9211	n.i.	-	-	P9211_01747	114	92/100	P9211_01752	512	443/492
MIT 9301	n.i.	-	-	P9301_15291	105	92/100	P9301_15281	509	437/489
MIT 9303	n.i.	-	-	P9303_05431	119	94/98	P9303_05441	488	443/486
MIT 9312	n.i.	-	-	PMT9312_1441		92/100	PMT9312_1440	498	442/495
MIT 9313	PMT1419.5‡	54	27/50	PMT1419	119	93/98	PMT1418	499	448/497
MIT 9515	n.i.	-	-	P9515_15041	108	92/100	P9515_15031	509	445/504
NATL1A	n.i.	-	-	NATL1_17701	107	94/102	NATL1_17691	500	441/485
NATL2A	n.i.	-	-	PMN2A_0914	107	94/102	PMN2A_0913	500	441/485
CCMP1375	n.i.	-	-	Pro1424	117	95/102	Pro1423	501	441/484
CCMP1986	n.i.	-	-	PMM1343	107	91/100	PMM1342	509	446/497
PCC 6301	syc0332_d	284	283/284	syc0333_d	102	102/102	syc0334_d	519	519/519
PCC 7942	Synpcc7942_1218	284	284/284	Synpcc7942_1217	102	102/102	Synpcc7942_1216	519	519/519
BL107	BL107_16390	292	165/279	BL107_16385	120	95/99	BL107_16380	501	455/495
CC9311	sync_2222	328	170/284	sync_2221	119	96/100	sync_2220	511	458/505
CC9605	Syncc9605_2126	294	173/281	Syncc9605_2125	121	95/99	Syncc9605_2124	512	459/504
CC9902	Syncc9902_0547	292	167/280	Syncc9902_0548	120	95/99	Syncc9902_0549	512	457/504
JA-2-3B'a (2-13)	CYB_0490	334	163/305	CYB_0489	100	91/98	CYB_0488	541	458/514

Table 2-5. Continued.

Organism*	KaiA			KaiB			KaiC		
	GenBank ID	Size (aa)	Positives §	GenBank ID	Size (aa)	Positives	GenBank ID	Size (aa)	Positives
RS9916	RS9916_34297	294	170/285	RS9916_34292	119	95/99	RS9916_34287	512	462/505
RS9917	RS9917_08621	302	170/282	RS9917_08626	119	94/99	RS9917_08631	519	463/505
WH 5701	WH5701_14956	295	166/281	WH5701_14951	92	73/78	WH5701_14946	514	458/502
WH 7805	WH7805_12833	296	179/284	WH7805_12838	119	95/99	WH7805_12843	512	461/505
WH 8102	SYNW0548	296	176/285	SYNW0549	104	95/99	SYNW0550	512	459/507
PCC 6803	slr0756	299	177/294	slr0757	105	95/101	slr0758	519	467/503
	-	-	-	sll1596	108	60/84	sll1595	568	281/485
	-	-	-	sll0486	102	65/90	slr1942	505	366/497
BP-1	tlr0481	283	175/285	tlr0482	108	93/99	tlr0483	518	470/508
IMS101	Tery_3803	325	187/304	Tery_3804	104	94/101	Tery_3805	519	471/498

* ATCC 29413, *A. variabilis* ATCC 29413; WH 8501, *C. watsonii* WH 8501; CCY0110, *Cyanothece* sp. CCY0110; PCC 7421, *G. violaceus* PCC 7421; PCC 8106, *Lyngbya* sp. PCC 8106; CCY9414, N. *spumigena* CCY9414; PCC 73102, *N. punctiforme* PCC 73102; PCC 7120, *Nostoc* sp. PCC 7120; AS9601, *P. marinus* str. AS9601; MIT 9211, *P. marinus* str. MIT 9211; MIT 9301, *P. marinus* str. MIT 9301; MIT 9303, *P. marinus* str. MIT 9303; MIT 9312, *P. marinus* str. MIT 9312; MIT 9313, *P. marinus* str. MIT 9313; MIT 9515, *P. marinus* str. MIT 9515; NATL1A, *P. marinus* str. NATL1A; NATL2A, *P. marinus* str. NATL2A; CCMP1375, *P. marinus* subsp. Marinus str. CCMP1375; CCMP1986, *P. marinus* subsp. Pastoris str. CCMP1986; PCC 6301, *S. elongatus* PCC 6301; PCC 7942, *S. elongatus* PCC 7942; BL107, *Synechococcus* sp. BL107; CC9311, *Synechococcus* sp. CC9311; CC9605, *Synechococcus* sp. CC9605; CC9902, *Synechococcus* sp. CC9902; JA-2-3B'a(2-13), *Synechococcus* sp. JA-2-3B'a(2-13); JA-3-3Ab, *Synechococcus* sp. JA-3-3Ab; RS9916, *Synechococcus* sp. RS9916; RS9917, *Synechococcus* sp. RS9917; WH 5701, *Synechococcus* sp. WH 5701; WH 7805, *Synechococcus* sp. WH 7805; WH 8102, *Synechococcus* sp. WH 8102; PCC 6803, *Synechocystis* sp. PCC 6803; BP-1, *T. elongatus* BP-1; IMS101, *T. erythraeum* IMS101.

† n.i., not identified.

‡ PMT1419.5 is not in the original annotation.

§ Positives, Positives data of a pairwise alignment between a query protein from *S. elongatus* PCC 7942 and the corresponding subject protein in another cyanobacterium = Total positive amino acid residues/Total amino acid aligned.

8106 and *Synechocystis* sp. PCC 6803 also have two *kaiBC* operons and one solitary *kaiB* and one solitary *kaiC*. Furthermore, a KaiB-related protein with a size of ~250-285 aa is present in nine cyanobacteria strains, including *Anabaena variabilis* ATCC 29413 (Ava_3661, 254 aa), *Nostoc* sp. PCC 7120 (all3328, 254 aa), *Nodularia spumigena* CCY9414 (N9414_04275, 261 aa), *Nostoc punctiforme* PCC 73102 (Npun02007407, 285 aa), *Lyngbya* sp. PCC 8106 (L8106_03979, 268 aa), *Trichodesmium erythraeum* IMS101 (Tery_4671, 254 aa), *Cyanothece* sp. CCY0110 (CY0110_21130, 253 aa), *Crocosphaera watsonii* WH 8501 (CwatDRAFT_2485, 253 aa), and *Thermosynechococcus elongatus* BP-1 (tll0553, 267 aa). This protein carries a KaiB-like domain at its C-terminus. But its N-terminal fragment, which is conserved only in these strains, doesn't have any recognizable domain or motif. It is interesting that no *Prochlorococcus* or *Synechococcus* strain has this KaiB-related protein or extra copies of *kaiB* or *kaiC*, whereas all other strains carry either this protein or more than one copy of KaiB and KaiC. For the four strains that contain at least two copies of the *kaiBC* operon, *Synechocystis* PCC 6803 is the only strain that does not have this KaiB-related protein. The extra copies of KaiB and KaiC probably resulted from genome duplication and rearrangement, or lateral gene transfer (106). The origination and function of the KaiB-related protein is unknown.

In contrast to KaiB and KaiC, KaiA (ORF02529, 284 aa) is present in only some cyanobacterial strains, and not found in any other groups of organisms. As seen in Table 2-5, KaiA is well conserved in all freshwater or marine *Synechococcus* strains, as well as in *Lyngbya*, *Crocosphaera*, *Trichodesmium*, *Cyanothece*, *Thermosynechococcus*, and

Synechocystis species. For strains of the genera *Anabaena* and *Nostoc*, both belonging to the *Nostocaceae* family (including KaiA in strains not completely sequenced, data not shown), only the C-terminal ~100 aa of KaiA, which functions in stimulating autophosphorylation of KaiC (98), is conserved. There is no evidence for the presence of the N-terminal part of KaiA encoded elsewhere in the genome in any of these strains. However, a ~90 aa hypothetical protein appears at the place of the missing N-terminal part of KaiA, that is, immediately upstream of the C-terminal version of KaiA, in *Nostoc* PCC 7120 (GenBank accession No. BAB85866, located between all2883 and all2884 in whole genome annotation). This KaiA-associated protein is also conserved in all *Anabaena* and *Nostoc* species for both protein sequence and genomic context, and is not found in any other organisms. In *Anabaena* ATCC 29413, the *kai* locus is like that in *Nostoc* PCC 7120, with the KaiA-associated protein (located between Ava_1020 and Ava_1021) located upstream of truncated KaiA; however, in this strain two putative transposase genes separate these two genes from the downstream *kaiBC* operon, and the last ~15 aa of KaiA from the remainder of the coding region. In other *Nostoc* species, such as *Nostoc punctiforme* PCC 73102, *Nostoc* sp. PCC 9709, and *Nostoc cycadae*, the KaiA-associated protein is fused to KaiA to form a novel protein with a size of ~190 aa. In another *Nostocaceae* family species *Nodularia spumigena* CCY9414, there is likely a miss-sense point mutation that disrupts the coding region of the KaiA-associated gene, with only the first 36 aa residues likely to be translated. This apparent change could be due to sequencing errors in this region since the whole genome sequence of *N. spumigena* CCY9414 has not been assembled and finalized.

No homologues of KaiA can be identified in *Gloeobacter violaceus* PCC 7421 and all *Prochlorococcus* strains. One exception is *Prochlorococcus* MIT 9313, a low-light-adapted strain, in which an ~54 aa C-terminal version of KaiA is still encoded upstream of *kaiBC*, while in other *Prochlorococcus* strains, such as a high-light-adapted strain *Prochlorococcus* MIT 9312, no ORF can be detected between the *kaiBC* operon and the upstream ribosome protein operon encoding L21 and L27. Thus, it can be speculated that originally there was KaiA in *Prochlorococcus* strains and it was lost gradually during the adaptation of these cells to relative constant and specialized open ocean environments (185, 186). Whether the residual copy of KaiA in *Prochlorococcus* MIT 9313 and the C-terminal versions of KaiA proteins in *Nostocaceae* family species are still functional is unknown. The function and origination of the KaiA-associated protein is also an interesting question to pursue (106).

The HPK SasA (*Synechococcus* adaptive sensor, 387 aa) functions very close to the central oscillator of the circadian clock in *S. elongatus* PCC 7942 (60). SasA contains an N-terminal KaiB-like sensory domain, which interacts with KaiC, and a C-terminal histidine kinase domain. SasA is well conserved among sequenced cyanobacterial strains, except *Gloeobacter violaceus* PCC 7421, which also has no *kai* genes. The cognate RR of SasA, RpaA (249 aa) (61) is an OmpR-like protein. It is highly conserved in all sequenced cyanobacteria, including *G. violaceus* PCC 7421, which suggests that it has non-clock functions as well.

CikA (ORF01892, 754 aa), a circadian input kinase, carries a GAF domain, an HPK domain, and a *pseudo*-receiver (PsR) domain. Its N-terminal ~160 aa has no

recognizable features and is not found in any other organisms. The other parts of CikA, as a whole, are conserved in many cyanobacteria strains, but not in *Nodularia spumigena* CCY9414 and all *Prochlorococcus* or any other *Synechococcus* species. There is a possibility that other proteins, with different domain structure and length, substitute for the CikA function in these strains, as every cyanobacterium has at least one protein that contains all CikA domains (dada not shown). Another gene of a circadian input pathway, *ldpA* (light-dependent period), encodes an Fe-S protein (352 aa) involved in light-dependent modulation of the cyanobacterial circadian clock (59). An *ldpA* gene has been identified in all sequenced cyanobacteria strains. It has been found that LdpA functions as a sensor for intracellular redox state and transfer the information to the central oscillator (76).

The period-extender protein, Pex (ORF01928, 126 aa), carries a PadR-like transcriptional regulator domain (62, 77). It has been shown that Pex specifically bound a 25 bp DNA sequence upstream of *kaiA* promoter *in vitro* (187).The distribution of Pex in sequenced cyanobacterial genomes is different from that of other clock proteins. It is not present in *Crocosphaera watsonii* WH 8501, *Cyanothece* sp. CCY0110, *Gloeobacter violaceus* PCC 7421, *Synechococcus* sp. JA-2-3B'a(2-13), *Synechococcus* sp. JA-3-3Ab, *Synechococcus* sp. WH 5701, *Synechocystis* sp. PCC 6803, or any *Prochlorococcus marinus* strains. There is no *kaiA* in *P. marinus* and *G. violaceus* strains. If the function of Pex is mainly through its binding to the *kaiA* upstream sequence, then Pex is not required in these strains. For other strains that possess *kaiA*, but not *pex*, there might be other proteins with similar function to Pex. CpmA (ORF02479, 260 aa), a circadian

phase modifier with a NCAIR mutase-like domain, is well conserved in all sequenced cyanobacterial genomes, which suggests a function outside of the clock. A recently identified protein for negative feedback regulation of KaiC, LabA (low-amplitude and bright, ORF00403, 186 aa), contains a conserved DUF88 domain of unknown function (pfam01936) and is not encoded in any marine cyanobacterial strains, except *Synechococcus* sp. WH 5701, which can live in both freshwater and seawater. In all the other sequenced genomes, there are usually two or more copies of *labA*. ORF02452 (198 aa) in *S. elongatus* PCC 7942, which shows ~35 % identity to LabA in aa sequence and also carries a pfam01936 domain, is likely a paralogue of LabA. The circadian clock function of this protein has not yet been tested.

At least five group 2 sigma factors, which are similar to the principle sigma 70 in protein sequence but are not essential for growth, have been identified in PCC 7942: RpoD2 (SigB, ORF00251, 320 aa), RpoD3 (SigD, ORF01924, 320 aa), RpoD4 (ORF01812, 311 aa), RpoD5 (SigC, ORF00360, 398 aa), and RpoD6 (ORF00032, 310 aa). The first four proteins were found to be involved in circadian regulation of expression of some genes (66, 67). They are very similar to each other and to the principle RNA polymerase sigma factor (RpoD1, SigA) at the aa sequence level (at least 44% identity), and well conserved among cyanobacterial genomes in which various copies of group 2 sigma factors, from three to nine, are usually present.

The first genes identified to be involved in circadian clock function that are essential for viability are *clpP2* and *clpX* (65). They are arranged as an operon with the *clpP2* gene, encoding the ATP-dependent protease subunit ClpP2 (ORF01095, 244 aa)

upstream of *clpX* (ORF01096, 449 aa), which encodes the ATPase subunit of the Clp protease, ClpX; the latter functions as a chaperone to bring substrates to the protease (188). Consistent with their essential functions, both ClpP2 and ClpX are highly conserved among cyanobacterial strains.

Cyanobacterial circadian clock proteins can be divided into two groups based on their conservation. The first group consists of proteins that are highly conserved among most, if not all, cyanobacterial strains: KaiB, KaiC, SasA, RpaA, LdpA, CpmA, ClpP2, ClpX, and the group 2 sigma factors. These proteins are essential for either clock functions or cellular metabolism. Proteins in the second group are not present in at least a subset of cyanobacteria, e.g., *Prochlorococcus*: KaiA, CikA, Pex, and LabA. None of them are essential genes. Even though their function in the circadian clock may be critical in strains that have them, such as KaiA in *Synechococcus* strains, there might be other proteins with similar function in strains that do not contain them. The KaiA-associated protein conserved in all *Anabaena* and *Nostoc* species is a good candidate for functional substitution of a circadian clock protein.

Conserved domains in circadian clock proteins

Identifying the functional conserved domains that participate in the regulation of circadian rhythms is necessary for elucidating the molecular mechanism of the clock. Quite a few conserved domains appear in known clock genes. For example, the PAS domain is present in proteins encoded by many eukaryotic clock genes, e.g., *bmal1*, *clock*, and *period* in *Drosophila* and Mammals, as well as *wc-1/wc-2* in *Neurospora*

(28). As mentioned above, most cyanobacterial circadian clock proteins also have conserved domains. KaiC has two KaiC/RAD55 domains (RecA-superfamily ATPases implicated in signal transduction). Each RAD55 domain consists of a P-loop or Walker's motif and a DXXG motif conserved in various GTP-binding proteins (95). SasA and RpaA are cognate pair of histidine kinase and response regulator. There are three domains in CikA: a GAF domain, an HPK, and a PsR (58). LdpA contains a 4Fe-4S binding domain; Pex carries a transcriptional regulator PadR-like domain; CpmA have a NCAIR mutase (PurE)-related domain. Comparative analyses indicate that the circadian clock may have evolved independently in bacteria, fungi, plants and animals (28). However, employing some conserved domains, such as the PAS domain, is common to many clock components in most circadian models because of similar molecular mechanisms used for sensing of environmental and metabolic information (29).

PsR domains lack one or more of the three conserved amino acid residues essential for phosphoryl-accepting function in true receivers: the N-terminal aspartate (D) surrounded by negatively charged amino acids, the central aspartate (D), and the C-terminal lysine (K) followed often by a proline (P) (189). There is a group of *Arabidopsis* Pseudo-Response Regulators (APRRs), including the central clock component TOC1/APRR1 (190, 191), which are involved in the plant clock. A subset of the APRRs contains an N-terminal PsR domain and a C-terminal CONSTANS motif (present in a family of plant transcription factors). The rhythmic transcription of the members in this subset (APRR1/TOC1, APRR3, APRR5, APRR7, and APRR9) shows a pattern of circadian waves so that it peaks with 2-3 hours intervals in the order of

APRR9→7→5→3→1 (191, 192). Certain light stimuli can induce the robust circadian waves, probably through a positive cascade of transcription mediated by the PsR domains (192). Among cyanobacterial circadian clock proteins, KaiA, the central clock component, has a PsR domain at its amino terminus that possibly receives environmental cues transmitted, probably indirectly, from CikA, which also contains a PsR domain (98). However, the precise function of the PsR domain in KaiA is still unknown. Because KaiA, like CikA, is degraded in the presence of plastoquinone analogs (76), it is possible that the KaiA PsR, like the CikA PsR, directly binds a quinone (73); this possibility has not yet been tested. There are four other PsR domains encoded in the PCC 7942 genome, including one in NblR, which is involved in the degradation of the light-harvesting complex (phycobilisome) in response to nutrient deprivation (193), and another one in PsfR, which regulates *psbAI* expression (194). Whether the other two PsR domains are involved in circadian clock function is not yet known.

Conclusions

The complete genome sequence of *S. elongatus* PCC 7942 has been determined to be ~2.7 Mb in length, with ~55.5% G+C content. A total of around 2,900 genes, including 45 RNA genes, have been identified with automated annotation from TIGR. More than 5% of the putative ORFs were manually curated using a web-based tool, Manatee. The putative replication origin, *oriC*, is located at the *dnaN* locus. There are 7,402 HIP1 sites (GCGATCGC) in the chromosome of *S. elongatus* PCC 7942, but no insertion sequences (IS) are present, and there are only one putative transposase and two *pseudo*-

transposases in the genome, suggesting a distinct mechanism for chromosomal rearrangement from many other cyanobacteria. There are 16 TA (toxin-antitoxin) genes, belonging to six TA families, in the genome. Many proteins participate in signal transduction and photoreception. There are 11 histidine protein kinases, 20 response regulators, 5 hybrid histidine kinases, 5 chemotaxis proteins, 5 serine/threonine protein kinases, as well as 61 one-component proteins that carry one or more functional conserved domains like PAS/LOV, GAF, GGDEF, and EAL. More than a dozen well-studied proteins, including KaiABC, are involved in circadian clock function. Most of them contain conserved domains that are also present in photosensory and signaling proteins, *e.g.*, histidine protein kinase domains, receiver/pseudo-receiver domains, PAS domains, and GAF domains. The availability of the complete genome sequence of *S. elongatus* PCC 7942 and its closely related strain *S. elongatus* PCC 6301 provides advantages for elucidating molecular mechanisms behind these cellular functions with strategies of reverse genetics, comparative genomics, and microarrays, among others. Chapter III will describe a functional genomic strategy that seeks to provide many answers regarding the mechanism of the cyanobacterial circadian clock.

Materials and Methods

Transposon-mediated sequencing of cosmids

A 960-cosmid genomic library was constructed previously by inserting *Sau*3AI partially digested *Synechococcus* PCC 7942 genomic DNA into the *Bam*HI site of the SuperCos I cosmid vector (Stratagene) (65). Each cosmid carries ~30-40 kb of genomic DNA. Both ends of the genomic insert on each cosmid were sequenced using cosmid end primers. A BLAST query of these cosmid end sequences against the GenBank database identified several of them that have significant homologous hits in published *S. elongatus* PCC 7942 and PCC 6301 sequences (>99.5% identity). Five cosmids, each containing known genes that are distributed evenly on the physical map of PCC 6301 (147), were chosen as the starting point of the genome project. For each cosmid, an *in vitro* transposition assay was performed with a commercial derivative of bacteriophage/transposon *Mu* (GeneJumper™ Primer Insertion Kit for Sequencing, Invitrogen). The kit uses a minimal *Mu*Cm, which contains only the inverted repeats of *Mu* right end sequences at both of its ends and an antibiotic resistance marker, *cat*, for chloramphenicol (Cm), as well as purified MuA transposase. In the reactions, MuA assembles onto the MuA binding sites R1 and R2 at the ends of *Mu*Cm to form a functional transposition complex, which then inserts the *Mu*Cm into the target cosmid randomly (131, 141). *Mu* is one of the transposons with the least target specificity. With proper molar ratio of the transposon to the target DNA, only one copy of the *Mu*Cm will be inserted into a random position of the cosmid in most cases.

After transforming an *E. coli* host strain with the transposition reaction and selecting with both chloramphenicol (Cm) and Kanamycin (Km) (resistance encoded by SuperCos I), a population of colonies that each carries usually one *Mu*Cm inserted into a different position on the cosmid will be obtained. Restriction digestion of the cosmid DNA extracted from the colonies was performed to identify *Mu*s located only on the genomic sequence, not the cosmid vector. The resulted sample set, in 96-well format, was used for both generating PCC 7942 insertional mutants and sequencing.

Mu-mediated sequencing was carried out as reported (65). When the whole sequence of a cosmid has been completed, the precise position of each *Mu*Cm can be annotated. Normally, 96×4 sequences were produced for each cosmid. These sequences were then analyzed using the ContigExpress program of Vector NTI bioinformatics software suite (Invitrogen). Each sequence was polished by trimming 5'-*Mu*Cm end sequence and 3'-unreadable sequence. The cleaned sequences were assembled using modified standard parameters of the program. The assembled contigs were edited to correct ambiguous bases and other sequence conflicts, such as extra bases or missed bases in overlapping sequences.

There were always several gaps for each cosmid between remaining contigs because only 3-5 fold sequence coverage was available. The "primer walking" strategy was used to fill these gaps. End sequences of contigs were analyzed by the BLAST function against the GenBank database to identify contigs share the same putative ORF or published sequence. This is helpful to orient contigs and reduce the sequencing reactions needed to finalize the cosmid during "primer walking."

Manual annotation of finished cosmid sequences

For gene prediction, putative ORFs were identified using FramePlot 2.3.2, an on-line program specific for predicting protein-coding regions of high G+C content in bacterial genomes (195). Each putative ORF was then compared against the non-redundant protein database in GenBank using blastp (196). Significant hits were defined as homologous with at least 30% identity over at least 70% of the ORF length. Eight of those cosmids were completely annotated and published in public databases, GenBank and EMBL (Table 2-2), including a cosmid from the large endogenous plasmid pANL. Another cosmid 8D8, showing an internal ~20 kb deletion, was also annotated and listed in the table.

Automated annotation of the complete genome sequence

The DNA sequence was submitted to the TIGR Annotation Engine (www.tigr.org/AnnotationEngine), where it was run through TIGR's prokaryotic annotation pipeline. Included in the pipeline is gene finding with Glimmer, Blast-extend-repraze (BER) searches, HMM searches, TMHMM searches, SignalP predictions, and automatic annotations from AutoAnnotate. All of this information is stored in a MySQL database and associated files which was downloaded to our site. The manual annotation tool Manatee was downloaded from SourceForge (manatee.sourceforge.net) and used to manually review the output from the TIGR prokaryotic pipeline of the Annotation Engine.

CHAPTER III

INSERTIONAL INACTIVATION AND FUNCTIONAL ANALYSIS OF THE *Synechococcus elongatus* PCC 7942 GENOME

Introduction

Even though many circadian clock loci in *S. elongatus* PCC 7942, the model organism for cyanobacterial circadian rhythms, have been discovered and well-studied, the partners of some known clock component are still unidentified (58). Identification of all clock components in *S. elongatus* is necessary for fully elucidating molecular mechanisms of cyanobacterial circadian clock, as well as its relationship to metabolism and other essential cellular activities. The best way to unveil circadian components is to screen for mutants with altered circadian phenotypes, clone the genes that have been mutated, and characterize the function and regulation of these genes. Because *S. elongatus* PCC 7942 is naturally transformable and mediates homologous recombination with high efficiency, most known clock-related loci in *S. elongatus* were identified by several rounds of traditional *in vivo* mutagenesis procedures. For example, the *kaiABC* locus was revealed through chemical mutagenesis using EMS (ethyl methanesulfonate) (49, 57), while *cikA* was identified from a Tn*5* transposon-mediated insertional mutagenesis (52, 58). We propose a transposon-mediated *in vitro* mutagenesis and sequencing strategy to determine the sequences surrounding insertion mutations in

essentially every *Synechococcus elongatus* PCC 7942 gene (65). These genes, which are disrupted by inserted transposons, can then be re-introduced into *Synechococcus elongatus* PCC 7942 cells through transformation to create mutants and analyze their functions, with a focus on the phenotypes associated with the circadian rhythmicity of gene expression.

The completion of the genome sequence by the Department of Energy Joint Genome Institute (JGI) greatly facilitated our functional genomics project, which is very close to the finish line with over 95% of the genome mutagenized and approximately 30% of loci screened for circadian function. More than 70 new clock loci that belong to different functional categories were discovered through a team effort, including the *clpP2X* and *trxM* loci. Additionally, functional analysis of insertion mutants revealed that Type-IV pilus assembly protein PilN and RNA chaperon Hfq are involved in transformation competence of *S. elongatus* cells.

Results and Discussion

Transposon-mediated mutagenesis of individual genes

As described in chapter II, a 960-cosmid genomic library provided the templates for transposon-mediated mutagenesis and partial sequencing of the genome of *S. elongatus* PCC 7942. A "cosmid walking" strategy was adopted to choose new cosmids whose end sequences overlap with sequenced cosmids to proceed around the genome. For each cosmid, an *in vitro* transposition assay was performed with a commercial derivative of bacteriophage/transposon *Mu* (GeneJumper™ Primer Insertion Kit for Sequencing, Invitrogen) (65). Due to temporary technical problems with the GeneJumper™ Primer Insertion Kit, we used the EZ::TN <Kan-2> transposon (Epicentre) for part of the project; this approach required us to subclone fragments from cosmid inserts into plasmids because of incompatibility of the Km^R Tn*5* with the cosmid. In total, 45 cosmids carrying chromosomal sequences were completely mutagenized and sequenced, and 9 of them were subcloned and mutagenized with Tn*5* (Table 3-1 & Fig. 3-1).

At this point in the project the DOE Joint Genome Institute (JGI) completed the genome sequence and sent us all the genomic libraries that they used for shotgun sequencing. JGI developed 3-kb and 8-kb plasmid libraries, as well as a fosmid library for large (~40kb) inserts. All shotgun sequences used for assembling the whole genome also were sent to us as output files from the Phred/Phrap program (148, 149). The coordinates of each sequence in the assembly can be accessed in Consed (197, 198), a Unix-based graphical editor and automated finishing program for Phrap sequence assemblies. Because plasmids are much more convenient for sequencing and

Table 3-1. Current progress* of the functional genomics project of *S. elongatus*.

	Transposon		Insertion Sites sequenced	Total kb sequenced
	Mu	Tn*5*		
Published genomic cosmids	8		~2,300	~268
Completed genomic cosmids	28	9	~9,500	~1,425
Completed JGI plasmid sets		30	~6,000	~900
Total	66		~17,800	~2,593

	Cosmids/plasmid sets	Insertions	Genes
Screened in *S. elongatus*	20	~1,400	~700

*Modified and updated from Holtman et al, 2005 (65).

transformation, we switched to using JGI 8-kb plasmids as templates for mutagenesis and sequencing for the remainder of the functional genomics project. For economy and convenience, we used a set of five 8-kb plasmids (termed JGI 8-kb plasmid sets, or plasmid sets, hereafter) in one *in vitro* mutagenesis reaction with Tn*5*. Each plasmid set carries ~40 kb genomic sequence, comparable to the size of a cosmid insert. A total of 31 JGI 8-kb plasmid sets (designated 8S1-8S31) that cover most gap regions between sequenced cosmids were identified using Consed. Almost all of these plasmid sets, excluding 8S2, have been fully mutagenized and sequenced (Table 3-1 & Fig. 3-1).

Annotation of insertion site was performed in the Vector NTI software package (Invitrogen, Carlsbad, CA). Transposon-mediated sequences were assembled with the JGI complete sequence. The exact site of each *Mu*/Tn*5* insertion was then localized to its footprint (duplicated host sequence due to insertion of transposons, 5-bp for *Mu* and 9-bp for Tn*5* (199, 200)), which is determined as the genomic sequence immediately downstream of transposon end sequences. Originally at least 384 (4 X 96-well plates) individual insertion-mediated sequences were used to assemble contigs or main scaffold

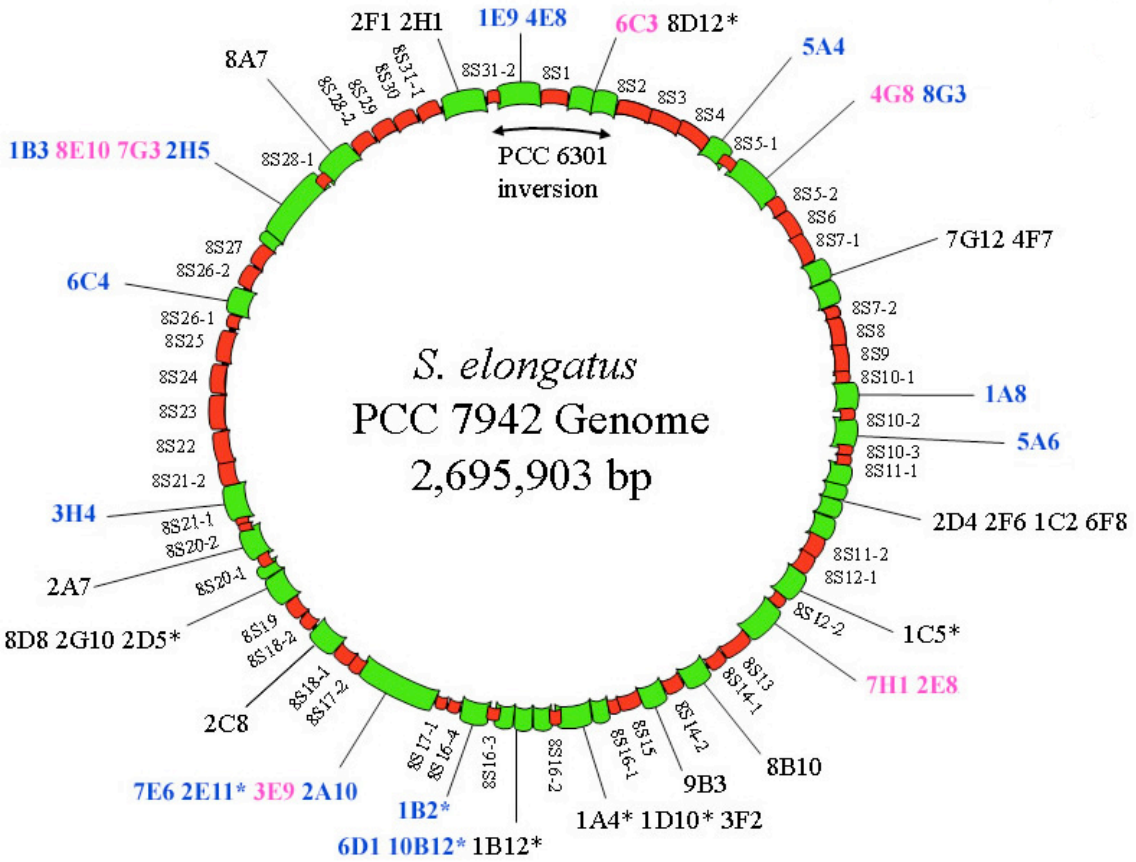


Fig. 3-1. Current status of the *S. elongatus* genome project. The circle represents the chromosome of *S. elongatus* PCC 7942. Green arched rectangles on the circle indicate the map positions, and lengths to scale, of mutagenized cosmids. An asterisk indicates the subset of cosmids that was mutagenized by Tn*5* rather than *Mu*. Cosmids with names in pink have been submitted into GenBank, while cosmids with insertion sites annotated, but not submitted, have names in blue. Where cosmids appear to abut, they overlap, and the cosmid names are indicated as a group. Red arcs represent the positions of JGI plasmid sets that were mutagenized to fill the gaps left after cosmid mutagenesis. The position and extent of a large inversion between two *S. elongatus* genomes is indicated by a double-headed arrow.

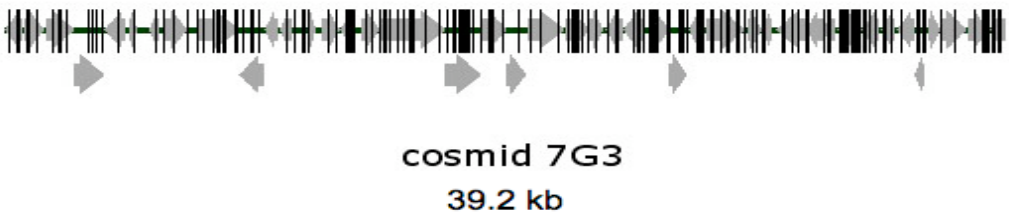


Fig. 3-2. Saturation of cosmid 7G3 with *Mu* transposon insertions. The 39.2 kb *S. elongatus* DNA segment in cosmid 7G3 is depicted. Gray arrows represent ORFs and those below the line overlap other ORFs. Black vertical lines indicate the positions of *Mu* insertions.

of a cosmid. This usually caused a cosmid to be over-saturated by transposon insertions as shown in cosmid 7G3 (Fig. 3-2). With the complete genome sequence available, only 192 (2 X 96-well plates) insertions are required for saturation of an ~40 kb cosmid or plasmid set.

Screening for circadian phenotypes of transposon insertion mutants

Transposon-inserted cosmid/plasmid DNA can be used directly to transform *S. elongatus* bioluminescent reporter strains. Because these cosmids or plasmids can not replicate autonomously in cyanobacterial cells, the only surviving cyanobacterial cells under the selection pressure of antibiotics carried by transposons are the ones harboring transposons integrated *via* homologous recombination into the corresponding region of the chromosome. In most cases, transformants are apparent double-recombinants, such that the vector backbone is lost (201). If the disrupted region is essential, a wild-type copy and a mutant allele are both maintained. This may happen through a single recombination event, resulting in both alleles in the same chromosome, or the maintenance of wild-type and mutant (double recombination event) chromosomes (201). Otherwise, no transformants will be obtained.

High-throughput transformation of *S. elongatus* strains was conducted in 24-well microplates using liquid BG-11 medium, instead of agar plates (65). This method greatly has a disadvantage in that no single colonies are isolated. Thus, a mixture of several or many independent colonies is used for further phenotypic analysis. Single recombinants and double recombinants can coexist in the same population. Even wild-type alleles may

be present if segregation has not completed. If disruption of a gene by insertion has a growth phenotype, single recombinants probably outgrow double recombinants during the course of selection. Thus, it will be more difficult to detect the attenuated phenotype. Despite this drawback, we have been able to identify mutant phenotypes with this high-throughput approach. Clones that show a mutant phenotype by this method are re-tested with plating for individual colonies.

We are focusing mainly on circadian clock phenotypes. An automated system has been developed, allowing continuous monitoring of the circadian rhythms in cyanobacteria (52, 202). In this system *S. elongatus* promoters are fused to promoterless *luxAB* (*Vibrio* luciferase gene) or *luc* (firefly luciferase gene) and then integrated into one of the Neutral Sites (see Materials and Methods) on the chromosome through homologous recombination. Bioluminescence of cyanobacterial cells produced by the luciferase reporter gene in 96-well black microplates can be counted automatically and continuously using a Packard TopCount luminometer (202). The reporter strains currently used for the functional genomics project are AMC1020 and AMC1300 for *Mu* mutants, and AMC462 for Tn*5* mutants. AMC1020 carries the promoter for *psbAI* fused to *luxAB* (P*psbAI*::*luxAB*) in NS1 and P*psbAI*::*luxCDE* in NS2. The operon *luxCDE* is from *Xenorhabdus luminescens* and produces aldehyde substrate for *Vibrio harvey* luciferase encoded by *luxAB* (52). The *psbAI* gene is one of three *psbA* paralogs that encode a critical photosystem II reaction center protein, D1 (203, 204). Northern blot results have shown that the expression of *psbAI* is rhythmic at the mRNA level (51), and peaks at dusk as do most *S. elongatus* genes (designated as class 1) identified by a

whole-genome screening of *luxAB* fusions using a cooled-CCD camera monitoring system (84). AMC1300, almost identical to AMC1020 except for the promoter that drives *luxAB*, is a P*kaiB* (class 1) reporter strain (65). AMC462 is another P*kaiB* reporter strain, with P*KaiB::luxAB* in NS1 and P*sbAI*::*luxCDE* in NS2 (63). A P*purF* (peaks at dawn, gene encodes glutamine PRPP amidotransferase)-based class 2 reporter gene may also be used (108, 205) to detect phenotypes of mutants that affect different aspects of the circadian clock.

We are currently looking for mutants with altered circadian periods or phasing, or arrhythmic expression under constant light conditions. Additional screening will be done to identify mutants with defects in entrainment or phase resetting. The latter mutants would be those that can be entrained by 12-h/12-h light/dark cycles, but cannot reset the relative timing of peak bioluminescence in response to a 5-h dark pulse applied at certain circadian time points.

Initially, phenotype screening was based on insertions of individual ORFs with annotated transposon insertions. For each putative ORF, usually two insertions were picked for transformation and screening. However, insertions located in the intergenic regions were overlooked. Because there might be unidentified small RNAs or small unannotated protein-coding ORFs located in intergenic regions, we decided to check all available insertion mutants. Around 20 cosmids or plasmid sets have been screened for the first-round screening, accounting for ~1,400 insertions in ~700 ORFs (Table 3-1).

The functional genomics project is a team work coordinated by Dr. C. Kay Holtman (65). I serve as the chief bioinformaticist for the project, and conduct

annotation of coding regions and transposon insertion sites. I also performed the mutagenesis and phenotypic screening for insertions in several cosmids: 7G3, 8D8, 4G8, 2H5, 8E10, and 6F8.

Novel clock-related loci in the genome

Preliminary results show that among ~700 ORFs screened, several known clock genes were confirmed, including the *kaiABC* locus. In addition, mutations in around 70 additional loci, not previously connected to the circadian clock, showed altered circadian phenotypes, primarily period changes. These loci belong to several different functional categories (Table 3-2).

A lengthened circadian period was observed in *Mu*-inserted mutants of the *clpP2clpX* operon, which contains the first two genes in the first cosmid (7H1) screened. The *clp* operon encodes an ATP-dependent Clp protease complex, which is ubiquitous in bacteria, plants, and animals. ClpP is a serine-type protease (206), while ClpX functions as an ATPase subunit, now known as a member of the Clp/Hsp100 family of chaperones (188). This protease complex may be involved in degradation of clock proteins, leading to the period phenotype in the mutants (65). These two *clp* genes are the first clock genes in the cyanobacterium found to be essential for cell viability.

There are two interesting insertion mutants in the *kai* locus in cosmid 7G3, the first cosmid I screened. All other insertion mutants in this locus were arrhythmic, as expected when the central oscillator genes are disrupted. One *Mu* transposon, 7G3GG5,

Table 3-2. Categories of novel clock ORFs.

Functional Category	# of ORFs
Transporters	5
Metabolic (synthases, isomerases, etc.)	11
Light harvesting antenna	10
Membrane-related	3
Redox-related	4
Cell division-related	2
Hypothetical (conserved or unique)	21
Intergenic regions	3
Proteases	2
Regulatory proteins	3
Other	7
Total	**71**

was inserted at the C-terminal coding region of *kaiA* and caused an atypical short period phenotype. The functional analysis of this mutant is reported in Chapter IV. The other one, 7G3DD8, is localized in the N-terminal domain region of *kaiC*. A premature stop codon was formed at the insertion site, which truncates the coding region of KaiC after the first 66 codons. This mutant displayed an occasional rhythmic phenotype. That is, during a TopCount bioluminescence assay, mutant cells in a subset of wells in a 96-well plate showed unstable and imperfect circadian rhythms. Because the first 66 aa of KaiC is not likely to sustain a normal clock function, there could be a *kai*-less clock that controls rhythmic expression from a reporter gene. However, a reconstructed *kaiC66* allele could not reproduce the original occasional rhythmic phenotype in a *kaiC* null strain (data not shown), and the mutant was not pursued.

Mu insertions in two other ORFs of cosmid 4G8 also showed slightly long period circadian phenotypes (~0.5-1.0 h longer than wild type, data not shown) in reporter strains AMC1020 and AMC1300. 4G8L6 and 4G8EEE10 disrupt the N-terminus of a

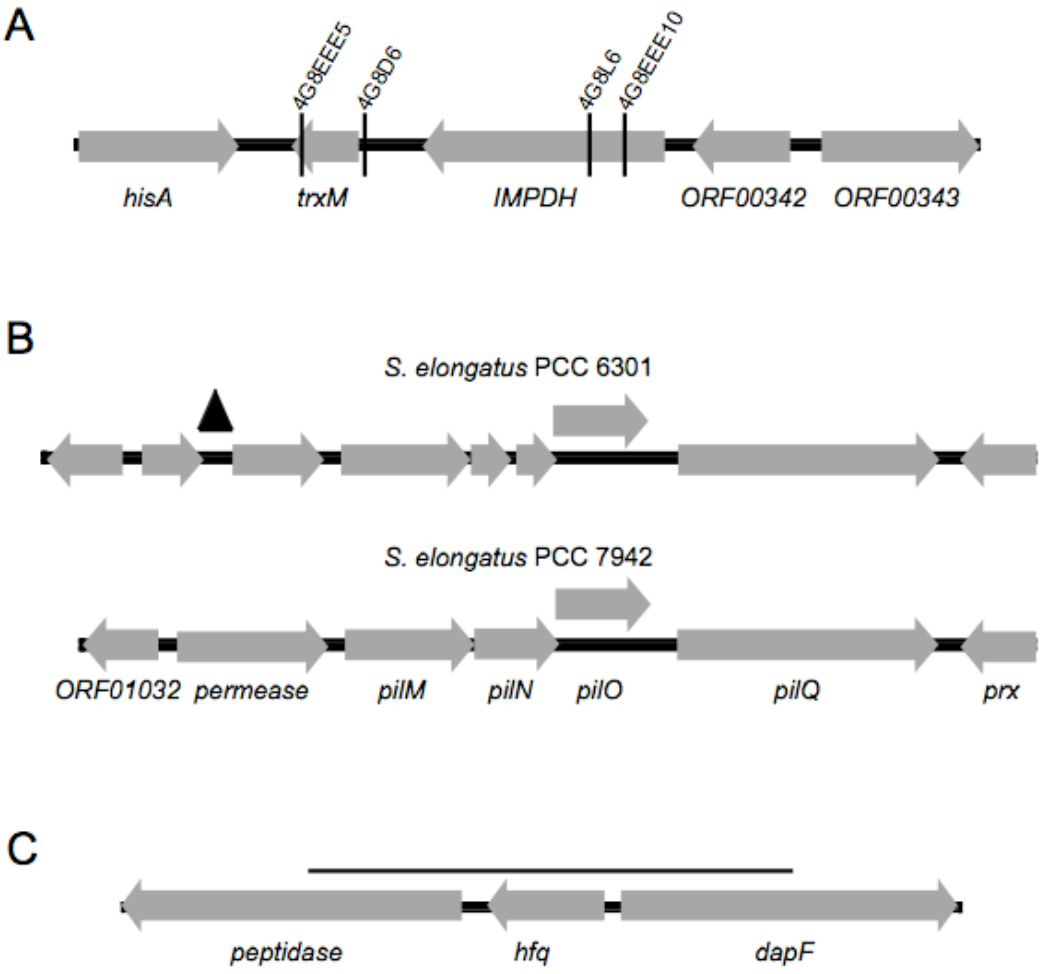


Fig. 3-3. Graphical representation of ORFs for insertional analysis. Gray arrows are ORFs and overlapping ORFs are shown on the top of the chromosome (long black horizontal bars). ***A)*** Thioredoxin (*trxM*) locus in cosmid 4G8. ORF00339 (*hisA*), ORF00340 (*trxM*), ORF00341 (*IMPDH*), ORF00342, and ORF00343 are labeled below the ORFs. *Mu* insertions 4G8EEE5, 4G8D6, 4G8L6, and 4G8EEE10 are shown at the top. Black vertical bars are *Mu* insertion sites. ***B)*** PilN locus of *S. elongatus* PCC 6301 (upper panel) and *S. elongatus* PCC 7942 (lower panel). Upper panel, from left to right, *syc1649_c, syc1650_d, syc1651_d, syc1652_d (pilM), syc1653_d, syc1654_d, syc1655_d (pilO), syc1656_d (pilQ),* and *syc1657_c* (*prx*, Peroxiredoxin). The black triangle depicts the 243 bp fragment that is deleted in *S. elongatus* PCC 7942. Lower panel, from left to right, ORF01032, ORF01031 (putative permerase subunit of ABC transporter), ORF01030 (*pilM*), ORF01029 (*pilN*), ORF01028 (*pilO*), ORF01027 (*pilQ*), and ORF01026 (*prx*) are labeled below the ORFs. The orientation of *S. elongatus* PCC 7942 ORFs was reversed for convenience of comparison with *S. elongatus* PCC 6301 ORFs. ***C)*** Hfq locus with gene names labeled below the ORFs. From left to right, ORF00437 (putative peptidase), ORF00438 (*hfq*), and ORF00439 (*dapF*, diaminopimelate epimerase). The black line above the ORFs indicates the 1.2 kb fragment used for plasmid construction.

putative inosine monophosphate dehydrogenase (IMPDH, ORF00341) gene (Fig. 3-3A). Two other *Mu* insertions are related to the downstream thioredoxin gene, *trxM* (ORF00340): 4G8D6, which sits ~40 bp upstream of the start codon, and 4G8EEE5, which sits in the C-terminal coding region of the ORF. The *Mu* insertions in IMPDH mutant strains could not be fully segregated, while segregation in *trxM* mutants was complete (data not shown). This result suggests that the IMPDH gene is essential for viability, whereas the *trxM* gene is not. Because the IMPDH gene is immediately upstream *trxM* and in the same orientation, we could not rule out a possible polar effect such that insertions in IMPDH gene affect function of the downstream thioredoxin gene. The *trxM* gene was reported previously to be essential in our organism, but not in *E. coli* (207). There are at least three thioredoxin genes (ORF00300, ORF00340, and ORF00492) in *S. elongatus* genome. They are very similar in size and protein sequence. There is a hypothetical protein encoded by ORF00342, which is upstream of the IMPDH gene and transcribed in the same orientation. Two other flanking genes that are in the opposite orientation of this locus are: ORF00338, *hisA,* which encodes a histidine biosynthesis protein (phosphorybosilformimino-5-amino-phosphorybosil-4-imidazolecarboxamideisomerase); and ORF00343, encoding a putative Soj-like ATPases involved in chromosome partitioning (Fig. 3-3A). *Mu* insertions in these ORFs have not yet been checked.

Other novel clock-related loci are still under second-round testing to confirm their phenotypes. Many are likely to be also involved in essential cellular functions, such as cell division and metabolism (Table 3-2). These genes, together with *clpP2clpX* and

the IMPDH gene mentioned above, are, for the first time, are linking circadian clock functions to essential housekeeping pathways in cyanobacteria, which indicates a more significant role of the circadian clock in cellular functions than previously demonstrated.

Genes involved in natural genetic transformation of *Synechococcus elongatus*

S. elongatus PCC 7942 is naturally transformable (208). This competence is completely lost in the closely related PCC 6301, which is the type strain for the species (1). It has been suggested that this distinction is due to the structure and sequence differences in two porin-like proteins located at opposite ends of an ~190 kb chromosome inversion between these two strains (7). However, it is well known that a subset of proteins involved in natural DNA uptake in bacteria are very similar to those of type IV pili and type II secretion systems (209, 210). Type IV pili, widespread among Gram-negative bacteria, are implicated in gliding- and twitching-like surface motility and phototaxis, pathogenesis-related bacterial virulence, cell-cell interaction, communication, and aggregation, as well as natural transformation (211-214). Type IV pili have been identified in many cyanobacterial species, such as the unicellular *Synechocystis* sp. PCC 6803 (215, 216), toxic unicellular *Microcystis aeruginosa* PCC 7806 (217), and filamentous *Nostoc punctiforme* (218). A dozen of core type IV pili genes are also present in the genome of *S. elongatus* strains: four *pilA* homologues (ORF01046, ORF01163, ORF01237, and ORF01238), which encode prepilin, precursor of basic pilus unit pilin; *pilB* (ORF00595) and *pilT* (ORF00594) that encode the motors for pilus extension and retraction, respectively; *pilD* (ORF00447), which encodes the prepilin

peptidase; several other genes involved in pilus assembly, including *pilC* (ORF00593), *pilM* (ORF01018), *pilN* (ORF01017), and *pilO* (ORF01016); as well as a pore-forming secretin gene *pilQ* (ORF01015). Other type-IV pili genes found in *Synechocystis* PCC 6803, such as *pilG*, *pilH*, *pilI*, *pilS*, *pilR*, and *pilP*, have not yet been identified in *S. elongatus*.

When scrutinizing the locus around a 243-bp *Pvu*I fragment, which is one of the two small regions of *S. elongatus* PCC 6301 deleted in *S. elongatus* PCC 7942, another obvious difference between these two genomes was discovered. Immediately downstream of this fragment sits the *pilMNO* operon and *pilQ*. The second gene of the operon in *S. elongatus* PCC 7942, *pilN*, is split into two genes in *S. elongatus* PCC 6301, syc1653_d and syc1654_d (Fig. 3-3B). Sequence alignment of the coding region identified a C→T nonsense point mutation in the middle of the PCC 6301 *pilN* coding sequence. The mutation changes a glutamine codon (CAG) in PCC 7942 to a stop codon (UAG) in PCC 6301 of syc1653_d. A putative start codon (GUG) of syc1654_d appears 36 bp downstream of the UAG stop codon. There are two other single-nucleotide differences between the two *S. elongatus* strains in the downstream *pilO* coding sequence. One of them causes a substitution of asparagine (AAU) in PCC 7942 with aspartate (GAU) in PCC 6301, and the other one does not change the encoded amino acid (UUG→CUG, both for leucine). Thus, the function of PilO in these two strains should be similar. When disrupted in *Synechocystis* PCC 6803, *pilN* mutants completely lost transformation ability (216). All *pilN* genes in other cyanobacterial species are well conserved as a non-split gene, like the one in *S. elongatus* PCC 7942. These pieces of

evidence enable us to speculate that the point mutation in *pilN* is related to the difference between these two *Synechococcus* strains in natural competence. To test this hypothesis, the *S. elongatus* PCC 7942 *pilN* was disrupted by an *Ω* cassette (*ΩpilN*, Sp^RSm^R), and the mutant indeed displayed the expected phenotype - the loss of transformation ability. A putative single recombinant strain (AMC1591) that is resistant to both SpSm (from the *Ω* cassette) and Km (from the plasmid vector) did not show this phenotype (Table 3-3). The transformation phenotype was rescued when a wild-type copy of the *pilN* coding region, under the control of IPTG-inducible P*trc,* was inserted into the NS1 locus of the *ΩpilN* mutant following conjugal transfer from *E. coli*, although the transformation efficiency is much lower than that of the wild type (Table 3-3). A copy of the P*trc*::*pilN* gene was also introduced into the NS1 locus of *S. elongatus* PCC6301 through conjugation. Preliminary data showed that transgenic *S. elongatus* PCC6301 cells gained transformation ability, but at a very low efficiency (Table 3-3). Because wild-type PCC6301 cells are completely non-transformable, the results, if confirmed, indicates that *pilN* is a key player, but probably not the only one, that contributes to the competence difference between these two strains. There might be other genes that affect the efficiency of transformation. The expression of *pilM* and *pilO* in the same operon could be affected by the point mutation in *pilN*, which might be one of the reasons for the low transformation yield. This hypothesis can be tested by introducing the complete *pilMNO* operon into *S. elongatus* PCC6301.

In addition to the loss of transformation ability (Table 3-3), the *ΩpilN* mutant also showed a suspension phenotype in liquid culture. Wild-type *S. elongatus* cells

Table 3-3. Transformation efficiency of wild-type and mutant strains.

Strain	Characteristics	Transformation efficiency (%)*
AMC06	*S. elongatus* PCC 7942 wild type	100
AMC17¶	*S. elongatus* PCC 6301 wild type	0
AMC18‖	*S. elongatus* PCC 6301 wild type	0
AMC1588	AMC17+P*trc::pilN* (NS1)	0.1†
AMC1589	AMC18+P*trc::pilN* (NS1)	0.1†
AMC1590	AMC06Ω pilN-double§	0
AMC1591	AMC06Ω pilN-single§	90
AMC1592	AMC06Ω pilN + P*trc::pilN* (NS1)	16
AMC1593	AMC06ΩGmhfq	0
AMC1594	AMC06Ωhfq	0
AMC1595	AMC06ΩGmhfq + P*hfq::hfq* (NS1)	60
AMC1596	AMC06Ωhfq + P*trc::pilN* (NS1)	90

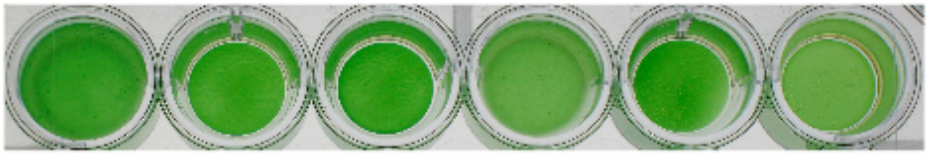
* Percentage of the frequency of transformation relative to wild type of *S. elongatus* PCC 7942 (AMC06); Transformation tests of all strains except AMC1588 and AMC1589 were conducted at least three times. The data from a representative experiment is shown for each. NS2 plasmids tested for transformation were AM1583 (Km^R) or AM2105 (Cm^R).
† Preliminary results from two independent experiments.
‡ For each plasmid, 1.0 μg DNA was used for transformation, which usually produced ≥1,000 individual colonies in AMC06.
§ Double, double recombinants; Single, single recombinants.
¶ AMC17 was obtained from S.E. Stevens (Pennsylvania State University, University Park, PA) as TX20(UTEX collection).
‖ AMC18 was obtained from M.M. Allen (Wellesley College, Wellesley, MA).

usually suspend in liquid medium homogeneously, while *pilN* mutant cells tend to sit at the bottom of the containers (Fig. 3-4). The suspension phenotype, however, was not complemented by an ectopic P*trc*::*pilN* gene (Fig. 3-4). Because *pilN* is in an operon, an insertion in the gene could also affect the expression of the downstream *pilO* gene, which is involved in pilus assembly.

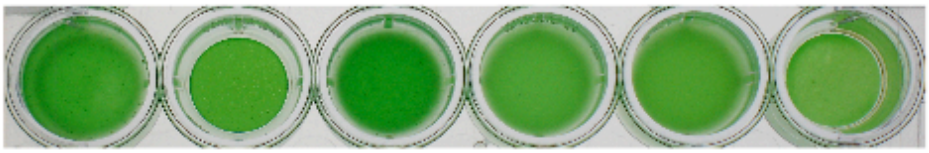
Surprisingly, an insertion mutant of a gene encoding a potential orthologue of the RNA chaperone Hfq (Fig. 3-3C) caused the same suspension phenotype. Hfq functions to help small regulatory RNAs form incomplete and imperfect duplexes with their target sequences in eubacteria, and has been identified in many cyanobacterial species, including *Synechococcus* (219, 220). The putative *hfq* gene (ORF00438) was originally disrupted by a gentamycin-resistance version of the *Ω* cassette (*Ω*Gm*hfq*) to check whether it is involved in circadian clock function, which was negative (data not shown). The transformation ability of *Ω*Gm*hfq* mutants was also lost in addition to the suspension defect. Both phenotypes were complemented when a copy of P*hfq*::*hfq* was inserted into NS1 of the mutant strain (Table 3-3 & Fig. 3-4). A copy of P*trc*::*pilN* (NS1) also rescued both transformation and suspension phenotypes in an *Ωhfq* mutant, in which a copy of original *Ω* cassette, not *Ω*Gm, was inserted in *hfq* due to an antibiotic conflict. The single-recombinant strain (AMC1591), in which the *pilMNO* operon is disrupted by insertion of the complete plasmid (pAM3755) sequence, did not show the suspension phenotype. Wild-type PCC 6301 strains suspend normally in liquid medium as does PCC 7942 wild type (data not shown).

No growth phenotype was observed in either *pilN* or *hfq* mutants and colonies of these mutants are normal on agar plates (data not shown). No IPTG was added in all complementation experiments and it is apparent that the *trc* promoter is leaky. The expression level of PilN or Hfq in these complemented strains is not known.

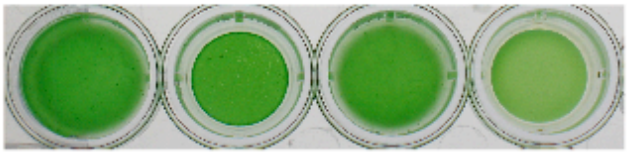
These results implicate *hfq* in natural competence and type IV pilus assembly in



AMC strain:	06	1590-1	1590-4	1591	1592-1	1592-4
Genotype:	WT	ΩpilN	ΩpilN	ΩNs	NΩN	NΩN
Tranformation:	+	-	-	+	+	+
Suspension:	+	-	-	+	-	-



AMC strain:	06	1593	1595-2	1595-6	1595-1	n.a.
Genotype:	WT	ΩGhfq	qΩGq	qΩGq	qΩGq	3861ΩGq
Tranformation:	+	-	+	+	+	-
Suspension:	+	-	+	+	+	-



AMC strain:	06	1594	1596	n.a.
Genotype:	WT	Ωhfq	NΩq	3955Ωq
Tranformation:	+	-	+	-
Suspension:	+	-	+	-

Fig. 3-4. Suspension phenotype of *S. elongatus* PCC 7942 strains in 24-well plates. The AMC strain numbers is shown below each well. Upper panel, from left to right: WT (AMC06), wild-type; ΩpilN (AMC1590-1/-4), *pilN* insertional null strains; ΩNs, single recombinant strain; NΩN (1592-1/-4), *pilN* insertional null strains complemented with P*trc*::*pilN* in NS1. Middle panel: WT (AMC06), wild-type; ΩGhfq (AMC1593), *hfq* insertional null strain (ΩGm); qΩGq (AMC1595-2/-6/-1), complementation strains with P*hfq*::*hfq* in NS1; 3861ΩGq (n.a., strain number not assigned), ΩGhfq strain transformed with a blank pAM3861 vector in NS1. Lower panel: WT (AMC06), wild-type; Ωhfq (AMC1594), *hfq* insertional null strain (Ω); NΩq (AMC1596), Ωhfq strain transformed with P*trc*::*pilN* in NS1; 3955Ωq (n.a., not assigned), Ωhfq strain transformed with empty pAM3955 vector in NS1. For transformation phenotypes: +, transformable; -, non-transformable. For suspension phenotypes: +, cells suspending; -, cells settling. Wells are viewed from the top; settling of strains to the bottom of the well can be visualized by an apparent clear halo around the green cell mass.

S. elongatus PCC 7942. As an RNA chaperone, *hfq* likely assists one or more small RNAs to stimulate the expression of *pilMNO* operon. It is also possible that only *pilN* is under the regulation of small RNA(s), as P*trc*::*pilN* can rescue both the competence and suspension phenotypes of *Ω*Gm*hfq*. Several sRNAs identified in *E. coli* and some other bacteria, *e.g.*, MicF, MicC, and MicA, participate in stress responses by post-transcriptionally regulating outer membrane proteins, mainly OMP family porins (221, 222). The implication of Hfq and small non-coding RNA in natural competence is not yet published.

Preliminary TEM results show that no pili assembled on the surface of *pilN* mutant cells, whereas both long (thick) and short (thin) pili, as previously observed in *Synechocystis* PCC 6803 (215, 216), were seen on wild-type cells (data not shown). These type IV pili are not only involved in natural transformation, but probably also help cyanobacterial cells to suspend and float in the water for better assimilation of light and nutrients.

Conclusions

More than 95% of the *S. elongatus* PCC 7942 genome was mutagenized by *Mu* or Tn*5* transposons. Among ~700 putative ORFs screened for circadian clock phenotypes, over 70 novel clock loci were discovered by the project team. Preliminary functional analysis has been performed for Clp protease (65) and thioredoxin loci (this study). Insertion analysis suggests the involvement of the Type IV pili assembly protein PilN and RNA chaperon Hfq in natural competence and cell suspension.

Materials and Methods

Transposon-mediated mutagenesis and sequencing

Transposon-mediated mutagenesis and sequencing of cosmids have been described previously (65), and is also summarized in Chapter II. Mutagenesis of JGI 8-kb plasmid sets is described in Results and Discussion. Sequencing of the sites of insertion for more than half of the clones was outsourced to High-Throughput Sequencing Solutions, a non-profit facility of the Department of Genome Sciences, University of Washington. Annotation of insertion sites is described in Results and Discussion.

Cyanobacterial strains, media, and culture conditions

The cyanobacterial strains used in this study are summarized in Table 3-4. All cyanobacterial wild-type reporter and mutant strains were created in *Synechococcus elongatus* PCC 7942. Cyanobacterial strains were grown in BG-11 medium (223) under continuous light conditions (~70 μE/m^2s) at 30°C with appropriate antibiotics. The bioluminescent reporter strains AMC1020, AMC1300, and AMC462 are described in Results and Discussion. Neutral sites mediate homologous recombination with the *S. elongatus* chromosome and cause no apparent phenotypes (52).

Cyanobacterial transformation and bioluminescence assay

Transformation of *S. elongatus* PCC 7942 strains and screening of circadian clock phenotypes were described previously (52, 65, 202).

Table 3-4. Cyanobacterial strains and plasmids used in Chapter III.

Strains	Genetic background	Plasmid/cosmid introduced	Antibiotic resistance†	Source or reference
AMC06	PCC 7942 WT*	None	None	Lab collection
AMC17	PCC 6301 WT	None	None	Lab collection
AMC18	PCC 6301 WT	None	None	Lab collection
AMC462	AMC06	pAM1887 (NS1)	Sp^RCm^R	Katayama et al. (1999)
AMC1020	AMC06	pAM1501 (NS1) pAM1619 (NS2)	Sp^RKm^R	Andersson et al. (2000)
AMC1300	AMC06	pAM1887(NS1) pAM1619 (NS2)	Sp^RKm^R	Lab collection
AMC1588	AMC17	pAM3956 (NS1)	Gm^R	This study
AMC1589	AMC18	pAM3956 (NS1)	Gm^R	This study
AMC1590	AMC06	pAM3755-double‡	Sp^RSm^R	This study
AMC1591	AMC06	pAM3755-single‡	$Km^RSp^RSm^R$	This study
AMC1592	AMC1590	pAM3956 (NS1)	$Gm^RSp^RSm^R$	This study
AMC1593	AMC06	pAM3759	Gm^R	This study
AMC1594	AMC06	pAM3758	Sp^RSm^R	This study
AMC1595	AMC1593	pAM3957 (NS1)	$Gm^RSp^RSm^R$	This study
AMC1596	AMC1594	pAM3956 (NS1)	$Gm^RSp^RSm^R$	This study

Plasmid	Characteristics	Antibiotic resistance†	Source or reference
pAM1303	NS1 cloning vector	Sp^RSm^R	Andersson et al. (2000)
pAM2202	pHP45Ω	Sp^RSm^R	Prentki & Krisch (1984)
pAM3515	pHP45ΩGm	Gm^R	This study
pAM3580	pCR-Blunt-pilN	Km^R	This study
pAM3755	pCR-Blunt-Ω pilN	$Sp^RSm^RKm^R$	This study
pAM2991	NS1 overexpression vector	Sp^RSm^R	Lab collection
pAM3955	pAM2991Gm	Gm^R	This study
pAM3956	pAM3955-pilN	Gm^R	This study
pAM3756	pLitmus29-hfq1.2	Ap^R	This study
pAM3758	pLitmus29-Ωhfq1.2	$Sp^RSm^RAp^R$	This study
pAM3759	pLitmus29-ΩGmhfq1.2	Gm^RAp^R	This study
pAM3861	pAM2428-α-lacZ (NS1)	Sp^RSm^R	Lab collection
pAM3957	pAM3861-hfq1.2	Sp^RSm^R	This study

∗ WT, *S. elongatus* wild-type strain

† Cm^R, chloramphenicol; Km^R, kanamycin; Sp^R, spectinomycin; Sm^R, streptomycin; Gm^R, gentamycin; Ap^R, ampicillin.

‡ Double, double recombinants; Single, single recombinants.

Transformation efficiency test and suspension ability assay

For checking transformation efficiency of wild-type and mutant strains, all strains were grown up to O.D. ~0.6 at 750 nm in 100 ml flasks. NS2 plasmids AM1583 (Km^R) or AM2105 (Cm^R) were used to test transformation ability with selection for different antibiotics. For each transformation, ~1.0 μg plasmid DNA was mixed with 300 μl washed and concentrated cells (202). After incubation overnight (~16 h) at 30°C in the dark, and resuspended cells (150 μl) were spread on BG-11 (223) agar plates with appropriate antibiotics. Numbers of colonies were counted or estimated after 7 days of incubation. For test suspension ability, all strains were grown up to O.D. ~0.6 at 750 nm in 100 ml flasks, and then transferred to 24-well plates. Each well contained 2 ml culture for each strain. Images were taken after 12-16 h of incubation without disturbance at room temperature (~20°C) on the bench.

Plasmid construction

All plasmids are described in Table 3-4. Unless otherwise stated, plasmids were constructed in *Escherichia coli* strain DH10B (Invitrogen, Carlsbad, CA). In order to create an inactivation allele of *pilN*, a blunt-end PCR product of the *pilN* coding region of *S. elongatus* PCC 7942 was first inserted into the pCR-Blunt vector (Invitrogen) to form pAM3580. Plasmid pAM3580 was then digested with *Sal*I, which cuts in the middle of *pilN*, and the sticky ends were filled by T4 DNA polymerase (NEB). A *Sma*I-digested fragment that carries the Omega cassette from pAM2202 was then inserted into the *pilN* gene, in the reverse orientation, to form pAM3755. To inactivate *hfq* a 1.2 kb

blunt-end PCR fragment, which includes the *hfq* coding region plus ~500 bp upstream and downstream sequences (Fig. 3-3C), was first inserted into *EcoR*V-digested pLitmus 29 (NEB) so that the *hfq* gene is in the same orientation as the Ap^R gene (pAM3756). Then a *Sma*I-digested Omega cassette from pAM2202 was inserted into the *EcoR*V site in *hfq*, with the *aadA* (Sp^RSm^R) gene is in the same orientation as the Ap^R gene (pAM3758). In pAM3759, the Omega cassette in pAM3758 was substituted by a gentamycin-resistant version of Omega from pAM3515, in which the *aadA* (Sp^RSm^R) gene in the Omega cassette is replaced by the *accC1* (Gm^R) gene. To make AM3955, the Omega cassette (Sp^RSm^R) in pAM2991, an NS1 overexpression vector, was replaced by the Omega cassette (Gm^R) from pAM3515. Both pAM3515 and pAM2991 were digested with *Hind*III, and then the Omega (Gm^R) fragment recovered from a gel was ligated into pAM2991ΔΩ. A copy of the *pilN* coding region was inserted into pAM3955 to construct pAM3956, such that *pilN* is under the control of P*trc*. Both 3955 and the *pilN* PCR fragment were digested by *Eco*RI and *Bam*HI, and then ligated with T4 DNA ligase. A 1.2kb blunt-end PCR fragment of the *hfq* region was inserted into the unique *Sma*I site in pAM3861, which is a NS1 overexpression vector that allows blue/white screening with X-gal, to construct pAM3957. The *hfq* is presumably under the control of its native promoter.

CHAPTER IV

AN ATYPICAL *kaiA* MUTANT THAT SHORTENS THE CIRCADIAN PERIOD OF *Synechococcus elongatus* PCC 7942

Introduction

The KaiA protein is a component of the central oscillator of the circadian clock in *Synechococcus elongatus* PCC 7942. KaiA contains a *pseudo*-receiver domain at its amino terminus that is proposed to receive environmental cues transmitted, probably indirectly, from circadian input pathways, and a carboxyl terminal domain functioning in enhancement of KaiC autokinase activity, which is attenuated by KaiB (80, 97-99). It has been shown that the period length of cyanobacterial circadian rhythms is determined mainly by phosphorylation status and degradation rate of KaiC (80, 83). A KaiA dimer, KaiB tetramer, and KaiC hexamer compose large protein complexes in a circadian pattern, which is closely correlated with the phosphorylation status of KaiC (70, 88, 89, 104, 105, 224) (Fig. 4-1). However, the molecular mechanism of how interactions among Kai proteins affect the phosphorylation status of KaiC and hence the circadian period is currently not fully understood.

Almost all point mutants in *kaiA* display either arrhythmia or a long-period phenotype (138). Here we report a new mutation in *kaiA* that causes an atypical short-period phenotype. The mutant was identified in a functional genomics project of *S. elongatus* PCC 7942, using a transposon-mediated mutagenesis and sequencing strategy to make insertions in essentially every gene (65). A 960-cosmid genomic library,

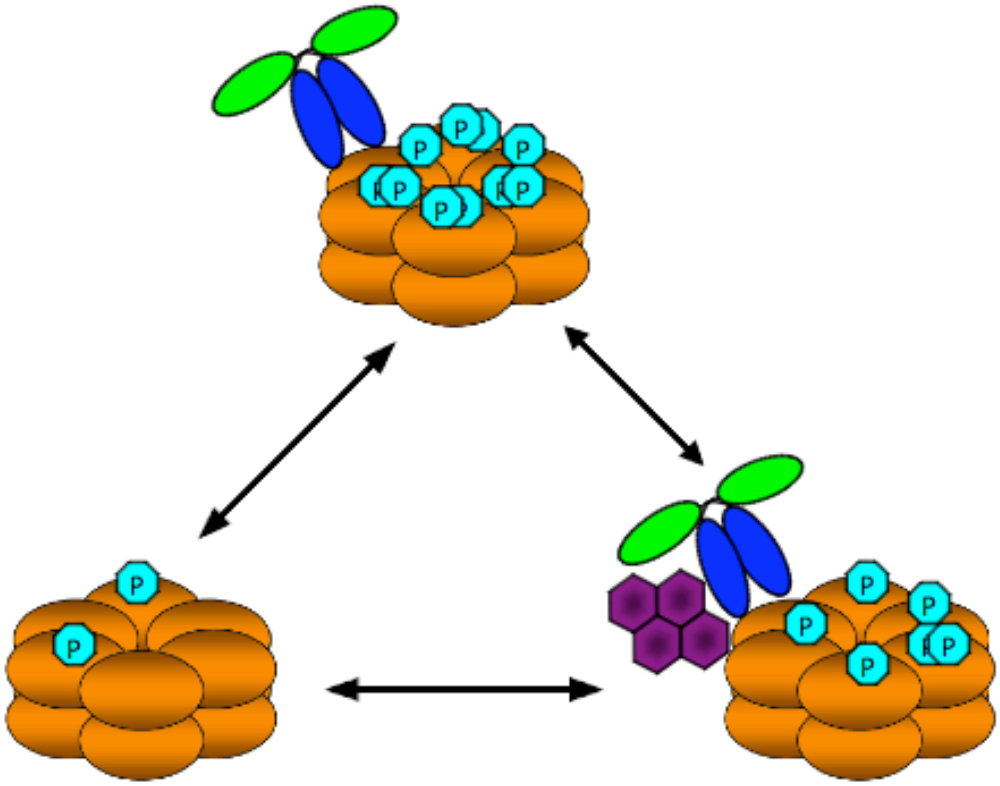


Fig. 4-1. Simplified model for molecular mechanism of cyanobacterial circadian clock central oscillator. The N-terminal domains of a KaiA dimer is in green, while the C-terminal domains are in blue. KaiB tetramer is in dark purple. KaiC hexamers are in orange. Phosphoryl groups are in light blue. The autophosphorylation of KaiC is stimulated by KaiA, which is attenuated by KaiB.

carrying 30-40 kb genomic DNA on each cosmid, serves as the template of *in vitro* mutagenesis. The genes disrupted by inserted transposons can then be re-introduced into *S. elongatus* reporter strains to create mutants for function analysis. During the screening for circadian phenotypes, which is our main interest, we identified an interesting transposon insertion in cosmid 7G3. The mini*Mu*Cm 7G3GG5, inserted at the C-terminal coding region of *kaiA*, causes an unusual shortened free-running period in constant light. Sequence analysis indicated that the last three amino acid residues, RET, are lost from KaiA when translation encounters a stop codon produced by the insertion of *Mu*Cm after polymerization of the first 281 residues.

In this study, we report that the short period phenotype in *S. elongatus* PCC 7942 is caused mainly by the truncation of KaiA. The disruption of a negative element upstream of the *kaiBC* promoter, which is another consequence of the insertion of the transposon, extends the circadian period. The overall circadian pattern of the accumulation of phosphorylated KaiC is roughly conserved in these mutants, but with some differences in relative levels of phosphorylated to unphosphorylated KaiC, as well as phase of circadian phosphorylation peak. No correlation between the KaiC phosphorylation pattern and bioluminescence rhythms in terms of phasing can be determined. The interaction between KaiC and the truncated KaiA is weakened as shown by fluorescence anisotropy analysis. Our data suggest the interaction between KaiA and KaiC, and the circadian pattern of KaiC autophosphorylation, are both important for determining the period, but not the phase, of circadian rhythms in *S. elongatus* PCC 7942.

Results

Identification of an atypical short period *kaiA* mutant

Most, if not all, *kaiA* point mutants in *S. elongatus* PCC 7942 show either arrhythmia or a long-period phenotype (138). However, during the screening of mutants that carry *Mu* insertional mutations in cosmid 7G3 for circadian phenotypes we identified an unusual short-period mutant in *kaiA*. This mini*Mu*Cm insertion, 7G3GG5, shortened the free-running period length of circadian clocks in the bioluminescent reporter strain AMC1020 (P*psbAI*::*lux*) from the ~24.4 h of the wild type to ~23.3 h in the mutant under constant light conditions (Fig. 4-2B & Table 4-1). This phenotype was confirmed in another reporter stain AMC1300 (P*kaiBC*::*luxAB*), in which the bacterial luciferase genes are under the control of the *kaiBC* promoter. The free-running period of the bioluminescence rhythm in wild-type AMC1300 is around 25.3 h, one hour longer than that of AMC1020 under the screening conditions used (see Discussion). With the insertion of the transposon, the period was also shortened about 1 h in the AMC1300 background (Fig. 4-2B & Table 4-1). The mutant strains showed no discernible growth phenotype compared to their corresponding wild-type strains. 7G3GG5 was found to be located at the carboxyl-terminal coding region of *kaiA* by sequencing. The insertion of 7G3GG5 separates the last 12 nucleotides, encoding RET plus the stop codon, from the rest of the coding region of *kaiA* (Fig. 4-2A). The end sequence of the transposon forms a new stop codon for a truncated *kaiA*, which would encode a protein with 281 amino acid residues instead of 284 aa in wild-type KaiA. The western blot analysis of KaiA protein levels confirmed the expression of a KaiA allele comparable to the wild-type

Table 4-1. Circadian periods of wild-type and mutated *kaiA* strains.

PCC 7942 Strains	Genetic Background*	Ectopic *kaiA* (NSI)†	Total *kaiA* alleles	Period±SEM‡ (h)	n
AMC1020	WT	-	*AWT*	24.40±0.05	39
AMC1483	*Mu*Cm	-	*A281*	23.32±0.05	57
AMC1300	WT	-	*AWT*	25.35±0.04	41
AMC1484	*Mu*Cm	-	*A281*	24.42±0.03	48
AMC541	WT	-	*AWT*	25.03±0.02	54
AMC1161	ΩKm	-	-	AR	20
AMC1485	WT	*AWT*	*AWT+AWT*	24.12±0.03	25
AMC1487	WT	*A281*	*AWT+A281*	23.81±0.03	26
AMC1486	ΩKm	*AWT*	*AWT*	25.02±0.03	28
AMC1488	ΩKm	*A281*	*A281*	24.49±0.05	23
AMC1491	*Mu*Gm	-	*A281*	24.51±0.22	23
AMC1492	*Mu*Gm+ΩKm	-	-	AR	24
AMC1531	*Mu*Gm	*AWT*	*A281+AWT*	24.48±0.10	17
AMC1532	*Mu*Gm	*A281*	*A281+A281*	23.81±0.08	16
AMC1533	*Mu*Gm+ΩKm	*AWT*	*AWT*	25.62±0.04	25
AMC1534	*Mu*Gm+ΩKm	*A281*	*A281*	NA§	16

* WT: wild-type; *Mu*Cm: original *Mu* insertion at the C-terminus of *kaiA*; ΩKm: omega cassette inserted at the N-terminus of *kaiA* for construction of a null strain; *Mu*Gm: reconstructed Gm version of the original *Mu* insertion.
† *AWT*: P*kaiA*::*kaiA*, wide-type allele; *A281*: P*kaiA*::*kaiA281*, the truncated allele.
‡ SEM: standard error of the mean
§ NA: not available; unstable traces, phase advanced and low amplitude (see Fig. 4-4A)

KaiA protein in size, with slightly faster mobility on SDS-PAGE gel (Fig. 4-2C).

A *kaiA* insertional knockout strain, AMC1161, was used as a negative control. This strain was generated by inserting a kanamycin-resistant version of the *Ω* cassette into a *Bam*HI site near the N-terminal coding region of *kaiA* (79), so that potential regulatory elements of *kaiBC*, which lie within the *kaiA* open reading frame, would not be affected (Fig. 4-2A). The insertion of 7G3GG5 not only truncated *kaiA* but also presumably disrupted a negative element (78), and resulted in noticeably elevated levels of KaiB and KaiC protein (Fig. 4-2C). Thus, the short period of 7G3GG5 mutant could be explained by either the truncation of KaiA or the disruption of the negative element, or both.

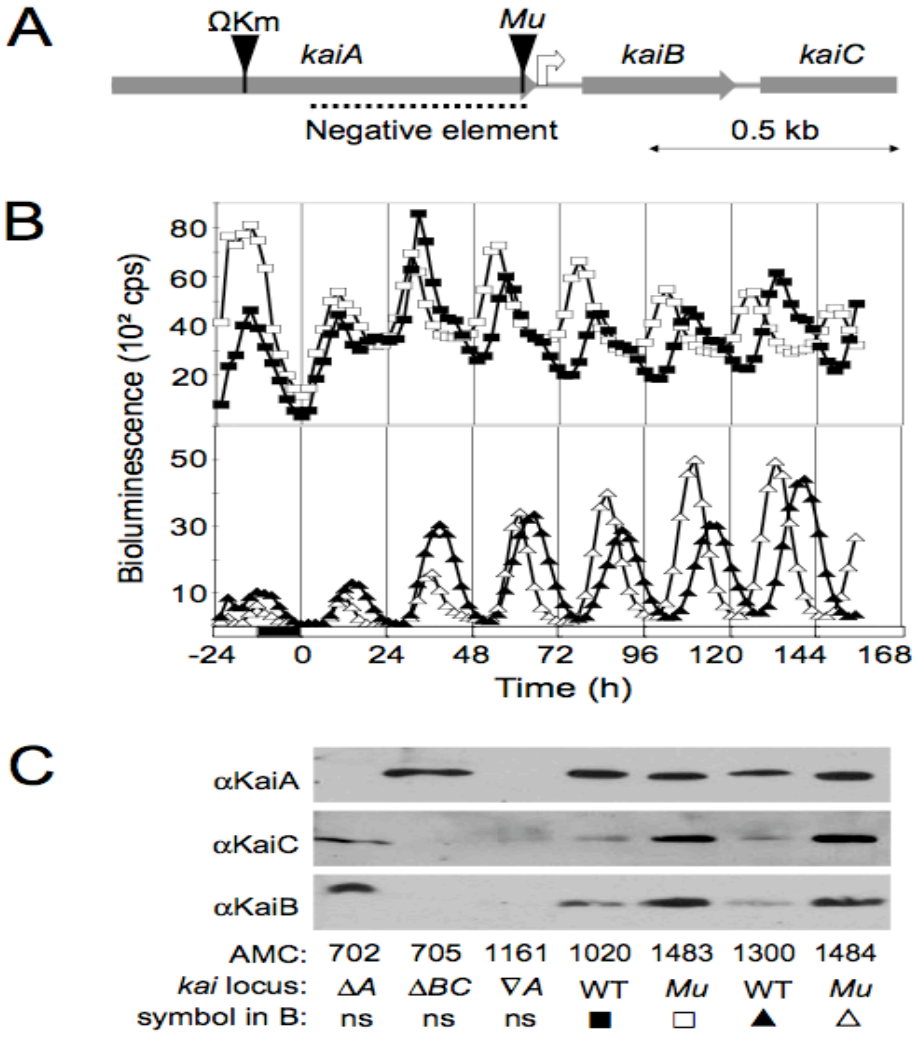


Fig. 4-2. An atypical short-period *kaiA* insertional mutant. ***A)*** Representation of the *kai* locus showing the relative positions of insertions: an Omega Cassette (*Ω*Km) and mini*Mu* transposon 7G3GG5 (*Mu*). *Ω*Km was inserted in the N-terminal coding region of *kaiA* to create AMC1161, a *kaiA* null. The *Mu* mutant encodes KaiA281, truncated by 3 residues, and disrupts the *kaiBC* negative element. A bent arrow shows the position of the *kaiBC* promoter. The dashed line under the C-terminal coding region of *kaiA* indicates a negative control element upstream of the *kaiBC* promoter. ***B)*** Representative bioluminescence traces from wildtype (closed symbols) and the *Mu* mutant (open symbols) in two different reporter backgrounds: *psbAI*::*luxAB* (squares) and *kaiBC*::*luxAB* (triangles). X-axis, time in hours: the blank bars represent light conditions; the black bar represents 12-h darkness. Y-axis, counts per second, cps. ***C)*** Immunoblot analysis of Kai proteins in wildtype and *Mu*-insertion mutant strains. Total soluble protein (20 μg) was loaded in each well. AMC702 and AMC705 carry in-frame deletions of *kaiA* (ΔA) or *kaiBC* (ΔBC), respectively; AMC1161 carries an insertional null allele of *kaiA* (∇A); WT, wild-type strain; *Mu* stands for the 7G3GG5 *Mu*Cm insertion strain. ns, not shown.

The short period phenotype is mainly due to the truncation of KaiA at its carboxyl terminus

To separate these two possible causes for the phenotype, two different mutant strains were made. One mutant carries an ectopic copy of the truncated allele of *kaiA* (*kaiA281*) with the negative element intact; in the other one, the negative element was disrupted at the same position as for the original transposon mutant in an insertional *kaiA* null strain, while a copy of wild-type *kaiA* was provided ectopically. We predicted that the former would show the effect of truncated KaiA alone, whereas the latter would recapitulate the effects of elevated KaiB and KaiC only.

The *kaiA281* and wild-type *kaiA* alleles were introduced into a wild-type reporter strain AMC541 (P*kaiBC::luc*) and the insertional *kaiA* knockout strain, AMC1161 (∇*kaiA*, P*kaiBC::luc*), respectively. As shown in Table 4-1 and Fig. 4-3A, the re-introduction of an ectopic copy of wild-type *kaiA* into NS1 locus (see Materials and methods) restored the rhythmicity of circadian clocks in AMC1161, the arrhythmic *kaiA* null, with a wide-type period of ~25 h. The *kaiA281* allele also restored rhythmicity to the AMC1161 background, but with a significantly shorter period (~24.5 h). The presence of two copies of wild-type *kaiA*, as in AMC1485, shortened the period by almost one hour (~24 h; comparing to about 25 h in both AMC541 and AMC1486). The period was also shortened when a copy of *kaiA281* was introduced in the NS1 locus of AMC541, the wild-type strain (~24 h).

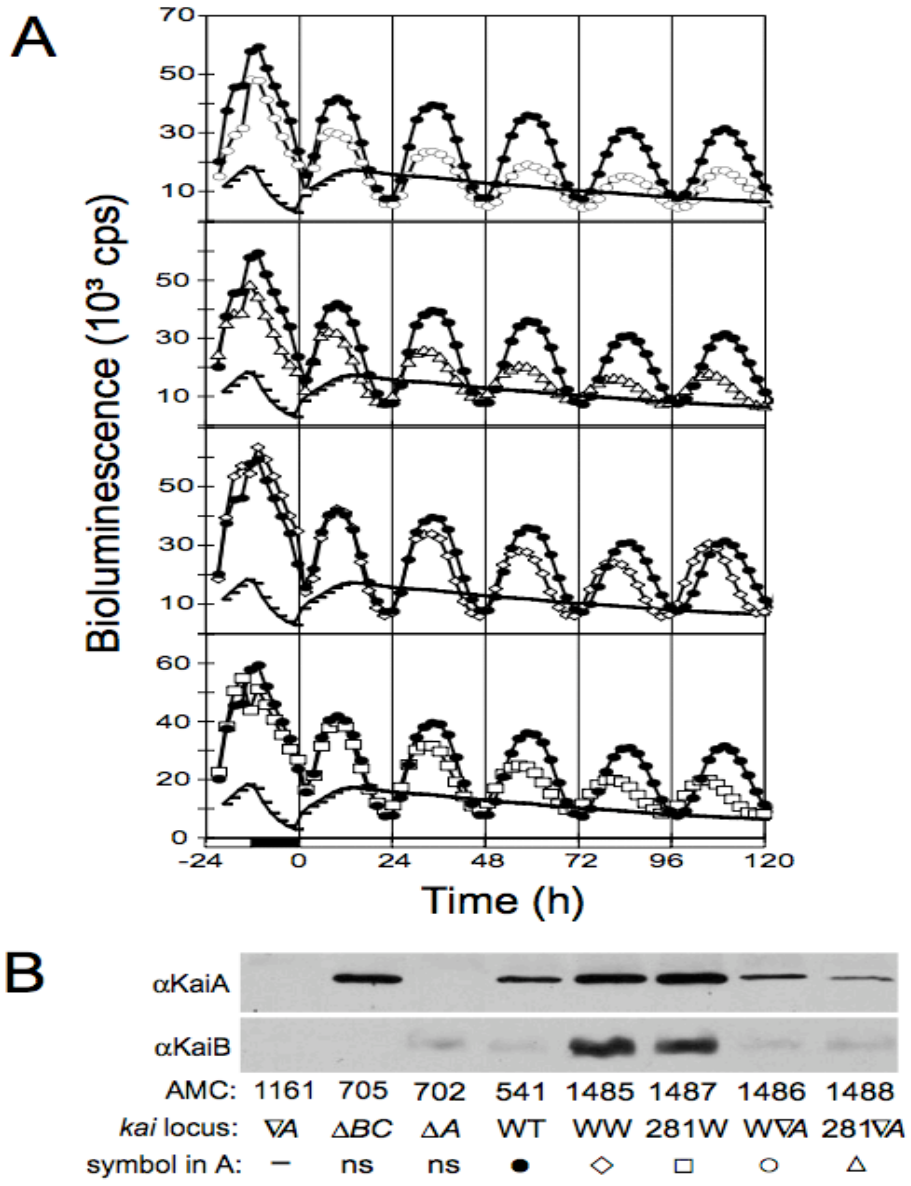


Fig. 4-3. Truncation of KaiA by 3 residues at its carboxyl terminus causes a short period. ***A)*** Representative bioluminescence traces from mutants that carry WT and/or truncated *kaiA* alleles. Closed circles, *kaiBC*::*luxAB* in a WT background (AMC541); dashed lines, an insertion knock-out mutant of *kaiA* (AMC1161)*;* open circles, an ectopic copy of WT *kaiA* in AMC1161 background (AMC1486); open triangles, an ectopic copy of truncated *kaiA* (*kaiA281*) in AMC1161 (AMC1488); open diamonds, an ectopic copy of WT *kaiA* in AMC541 (AMC1485); open squares, an ectopic copy of *kaiA281* in AMC541 (AMC1487). Axes are as for Fig. 1. ***B)*** Immunoblot analysis of KaiA and KaiB proteins performed as for Fig. 1. KaiC consistently behaved like KaiB and is not shown here. AMC702, AMC705, and AMC1161 are described in Fig. 1. WT, wild-type *kaiA*; WW, and 281W, an ectopic *kaiA* or *kaiA281* allele, respectively, in a WT background; W∇A and 281∇A indicate an ectopic wild-type *kaiA* or *kaiA281* allele, respectively, in a *kaiA* insertional null background. ns, not shown.

The protein levels of KaiA and KaiB in two complemented strains, AMC1486 and AMC1488, were not significantly different from those of wild-type strain AMC541 (Fig. 4-3B). Thus, the truncation of KaiA alone is not likely the cause of elevated KaiB and KaiC protein level in the original *Mu* Mutants. KaiA is known to stimulate *kaiBC* expression (57). When two copies of *kaiA* were expressed as in AMC1485 and AMC1487, we indeed saw notably higher KaiB protein levels in the cells (Fig. 4-3B), which suggests that the truncated KaiA281 protein retains the function to positively regulate *kaiBC* expression. These results clearly show that the truncation of *kaiA* could cause a short period phenotype.

Next we checked whether the disruption of the negative element also contributes to the shortened period in *Mu*Cm 7G3GG5 mutant strains. Due to the conflict of antibiotics resistance genes, the original *Mu* mutant was reconstructed so that the chloramphenicol marker, *cat*, in the middle of the mini*Mu* 7G3GG5 was replaced by the gentamycin resistance gene, *aacA*. The new mini*Mu*Gm was then introduced into both wild-type AMC541 and the *kaiA* null AMC1161. The insertion of *Mu*Gm into *kaiA* at the same position as in 7G3GG5 (AMC1491) again shortened the period, but by only about 0.5 h; in the AMC1161 background, cells remained arrhythmic after insertion of *Mu*Gm into *kaiA* (AMC1492) (Table 1 & Fig. 4-4A). The KaiB protein level was greatly elevated in AMC1491 due to the disruption of the negative element. In the absence of KaiA, only a trace amount of KaiB was detected even when the negative element was eliminated in AMC1492 (Fig. 4-4B). Wild-type *kaiA* or *kaiA281* was then introduced into both AMC1491 and AMC1492 as an ectopic copy. In the AMC1492 background,

AMC1533, which carries a copy of wild-type *kaiA* in NS1 site and a disrupted negative element, rhythmicity was restored with a slightly longer than normal period, 25.62 h (Fig. 4-4A & Table 1). The introduction of a copy of *kaiA281* in the AMC1492 background (AMC1534), however, could not restore a stable and perfect circadian rhythm in the cells. The circadian traces of these cells show reduced amplitude, advanced phase, and fast damping phenotypes (Fig. 4-4A, horizontal bars). No accurate period can be calculated from this strain. The presence of an ectopic copy of wild-type *kaiA* in the AMC1491 background (AMC1531) did not significantly change the period, whereas an extra copy of *kaiA281* in AMC1491 (AMC1532) slightly shortened the period (Fig. 4-4A & Table 4-1). Immunoblot data showed the expected increase in KaiB protein level when the negative element was disrupted in the presence of an ectopic copy of KaiA (Fig. 4-4B). Based on these results, we can conclude that the disruption of the negative element along with normal expression level of KaiA elevates *kaiBC* promoter activity, which also results in slightly extended circadian period.

In summary, the short period caused by the insertion of mini*Mu* 7G3GG5 in *kaiA* is mainly due to the truncation of KaiA. The *kaiA281* allele, without disrupting its basic function as a central oscillator component, shortens the period for ~0.5-1.0 h, depending on the background of the reporter strains. Because the phosphorylation status of KaiC is thought to be the major factor that determines the period length of cyanobacterial circadian clocks (80, 83), we speculated that the phosphorylation pattern of KaiC in the mutant strains would be different from that of wild-type strains.

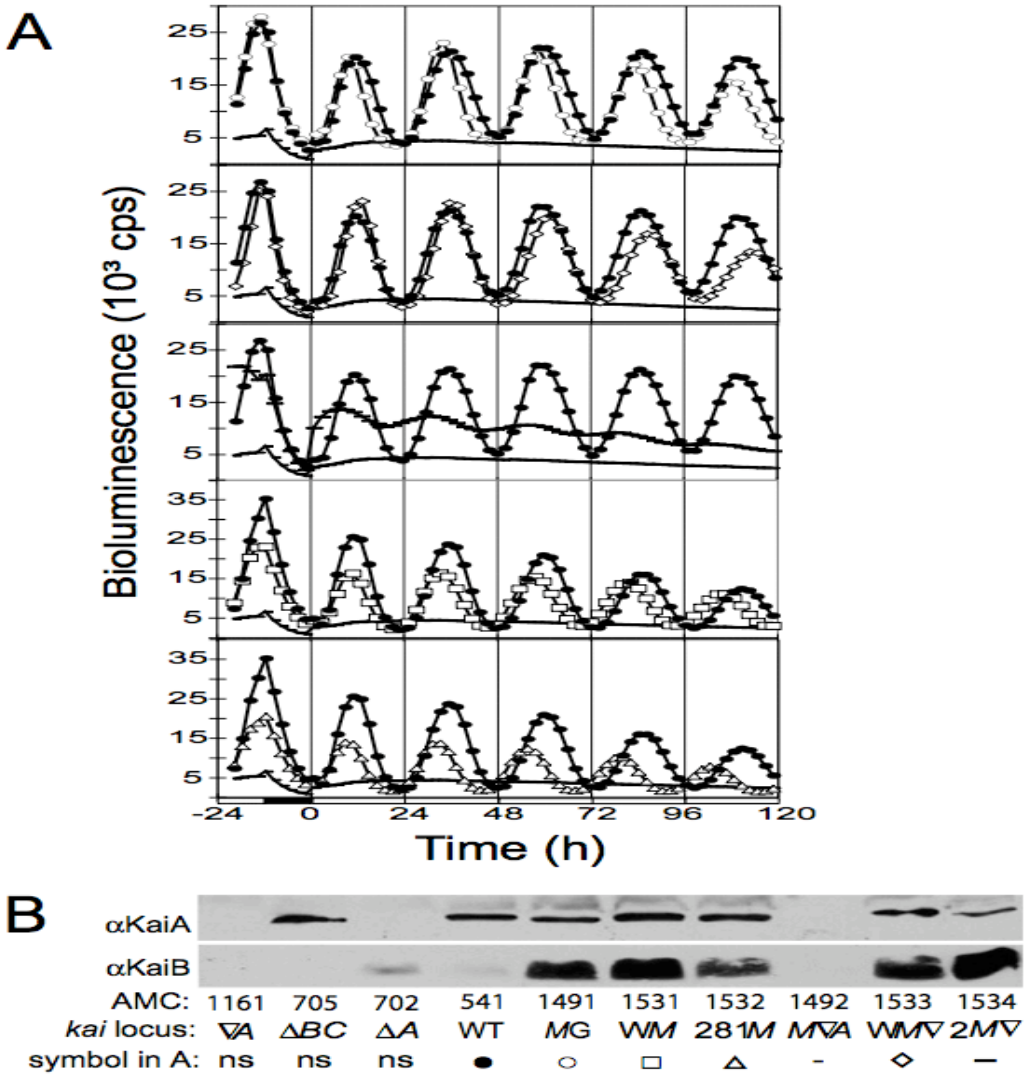


Fig. 4-4. Disruption of the negative element of the *kaiBC* promoter increases KaiB and KaiC levels and slightly slows the clock. ***A)*** Representative bioluminescence traces from *kaiBC*::*luxAB* strains that carry WT and/or mutant *kaiA* alleles. Closed circles, WT background (AMC541); open circles, 7G3GG5 mini*Mu*Gm mutant (AMC1491); dashed lines, *kaiBC* negative element disrupted by 7G3GG5 mini*Mu*Gm in *kaiA* insertional null (AMC1492); open diamonds (AMC1533) and horizontal bars (AMC1534) carry an ectopic WT or *kaiA281* allele of *kaiA,* respectively, in an AMC1492 background; open squares (AMC1531) and open triangles (AMC1532) carry an ectopic WT or *kaiA281* allele of *kaiA,* respectively, in an AMC1491 background. Axes are as described for Fig. 1. ***B)*** Immunoblot analysis of KaiA and KaiB proteins performed as for Fig. 1. AMC541, AMC702, AMC705, and AMC1161 are described in Fig. 1. WT, wild-type *kaiA*; *M*G, 7G3GG5 *Mu*Gm in a WT background; W*M* and 281*M indicate* an ectopic *kaiA* or *kaiA281* allele in a *Mu*Gm background, respectively; *M*∇*A*, 7G3GG5 *Mu*Gm in the *kaiA* insertional null background; W*M*∇ and 281*M*∇, an ectopic WT or *kaiA281* allele in the *Mu*∇*A* background, respectively. ns, not shown.

KaiC autophosphorylation pattern is altered in mutant strains

Time course analysis was performed to see whether KaiC phosphorylation status is altered in the *kaiA281* mutants. *S. elongatus* cultures were entrained with 2 cycles of either light/dark or dark/light (12-h/12-h). Synchronized *S. elongatus* cells were then sampled every 4 hours in constant light conditions. As shown in the top panel of Fig. 4-5A, the phosphorylated KaiC in wild-type strain AMC541 peaks around 8-12 h in constant light (LL8-12), while unphosphorylated KaiC peaks at LL0/24. Both complemented strains, AMC1486 and AMC1488, displayed an increased ratio of phosphorylated KaiC to unphosphorylated KaiC at LL0-4 and LL20-24. Samples taken at 1-h intervals clearly showed that the overall level of phosphorylated KaiC in these strains is elevated throughout the circadian cycle (Fig. 4-5A, lower panel). In AMC1486 and AMC1488 a significant amount of phosphorylated KaiC can be seen even at LL0 and LL24, when only unphosphorylated KaiC is present in wild-type strain AMC541. However, the overall circadian pattern of unphosphorylated KaiC in these complemented strains and the wild-type strain is very similar. There is no difference in the KaiC phosphorylation pattern between the *kaiA281*-complemented strain (AMC1488) and wild-type *kaiA*-complemented strain (AMC1486), even though they are slightly different in their circadian period.

We then checked the phosphorylation pattern of KaiC in AMC1533 and AMC1534, in which the negative element upstream of *kaiBC* promoter is disrupted. As in the reconstructed *Mu*Gm mutant AMC1491, the total level of KaiC protein is greatly increased in AMC1533 and AMC1534. AMC1491 and AMC1534 show similar patterns,

such that the elevated KaiC is mainly in the unphosphorylated state and accumulation cycle of phosphorylated KaiC is altered compared to that of the wild-type strain (Fig. 4-5B, upper panel). The rise in phosphorylated KaiC in these two strains is delayed until LL8, instead of LL4 in the wild-type AMC541. The phosphorylated KaiC level in AMC1491 and AMC1534, peaked around LL8-12 as in wild-type, then quickly diminished and was barely detectable at LL16, when wild type still has plenty of phosphorylated KaiC. The rise in phosphorylated KaiC in AMC1533 is further delayed until LL12 and peaks at LL16. It appears that the KaiC phosphorylation cycle has been shifted 4 hours later in this strain compared to the wild type. The results were confirmed in samples taken in a continuous 24-h sampling period instead of using two 12-h oppositely entrained cultures as in the previous time course experiments (Fig. 4-5B, lower panel).

In conclusion, the KaiC phosphorylation pattern in these mutant strains is different from the original wild-type strain in some aspects, such as protein level, relative ratio of phosphorylated to unphosphorylated KaiC, and phase of peak phosphorylation. However, the overall circadian pattern of rising and falling of phosphorylated KaiC and period of the phosphorylation cycle are retained. It seems that there is no direct phase correlation between the phosphorylation pattern of KaiC and the bioluminescence rhythms produced from the reporters.

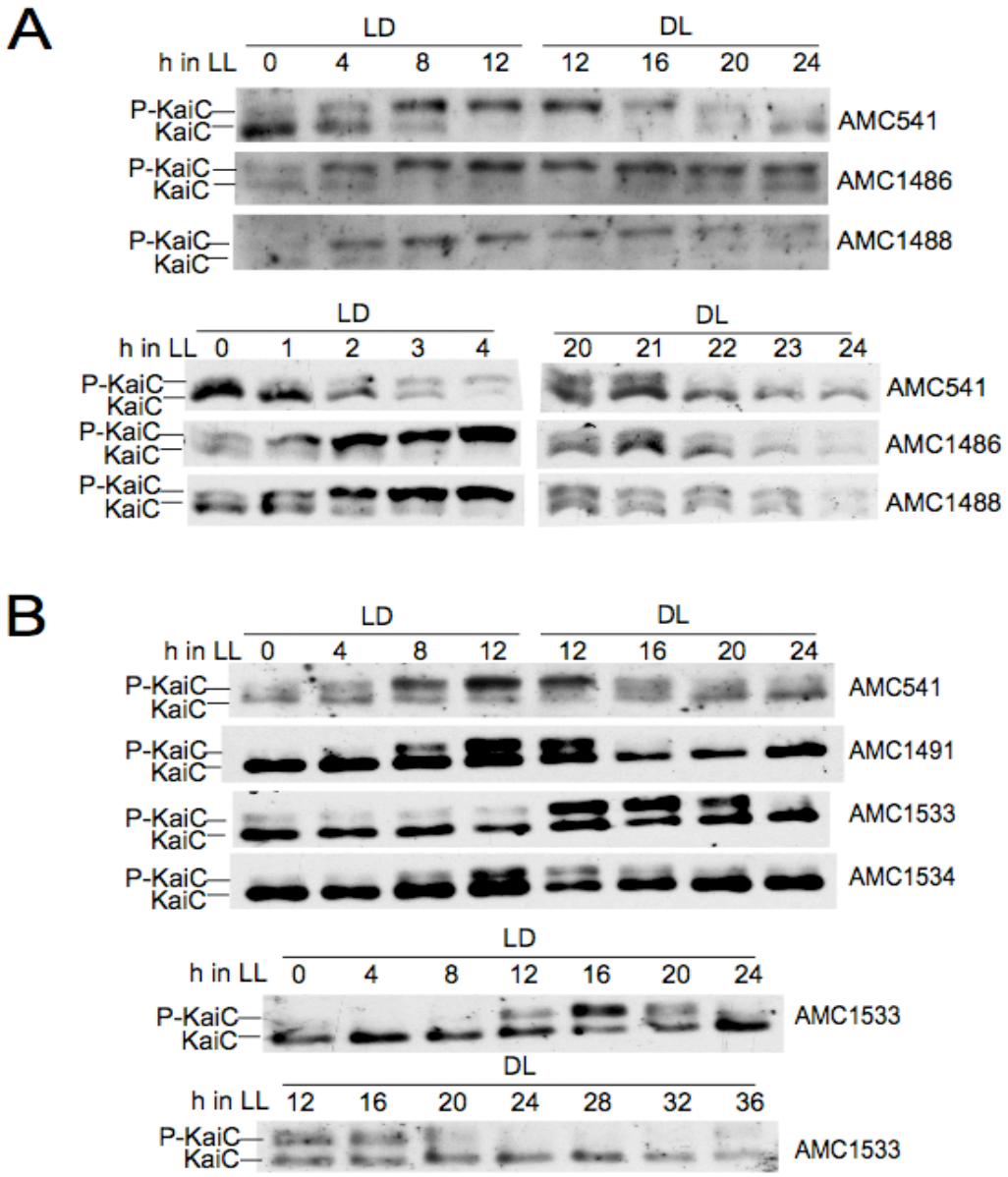


Fig. 4-5. KaiC autophosphorylation pattern in wild-type and mutant strains. ***A)*** Time course immunoblot analysis of *S. elongatus* strains AMC541 (wild-type), AMC1486 (wild-type *kaiA*-complemented strain) and AMC1488 (*kaiA281*-complemented strain). Upper panel: 4-h time point samples. Lower panel: 1-h time point samples. ***B)*** Time course immunoblot analysis of *S. elongatus* strains AMC541, AMC1491 (*Mu*Gm), AMC1533 (wild-type *kaiA*-complemented and negative element disrupted strain) and AMC1534 (*kaiA281*-complemented and negative element disrupted strain). Upper panel: 4-h time point samples taken in 12 hours. Lower panel: 4-h time point samples taken in 24 hours. LD (light/dark) and DL (dark/light) indicate opposite entrainment conditions. The hours released into constant light for each time point is shown on top of each lane. P-kaiC-, phosphorylated KaiC; KaiC-, unphosphorylated KaiC. Total soluble protein (40 μg) was loaded in each well.

The interaction between KaiA and KaiC is weakened in KaiA281

Those changes in the phosphorylation state of KaiC and the period differences in mutant strains are probably due to alternated protein-protein interaction between KaiA and KaiC. We performed fluorescence anisotropy assays to determine whether there is a change in the interaction between KaiA and a KaiC-derived peptide, referred as CII ATP binding domain (CIIABD), in the KaiA281 variant (see Materials and Methods). CIIABD locates at the C-terminus of KaiC and specifically binds KaiA (100, 101). Comparing with wild-type KaiA, KaiA281 exhibited less interaction with the KaiC peptide (Fig. 4-6). KaiA135N, the N-terminal domain of KaiA that possesses a *pseudo*-receiver like structure, did not exhibit a significant interaction with the KaiC peptide; conversely, the C-terminal domain of KaiA, KaiA180C, which has been shown greatly stimulated autokinase activity of KaiC (98), displayed a stronger interaction with the KaiC peptide than did full-length KaiA. The results suggest that the short period phenotype of *kaiA281* might result from the weakened KaiA-KaiC interaction.

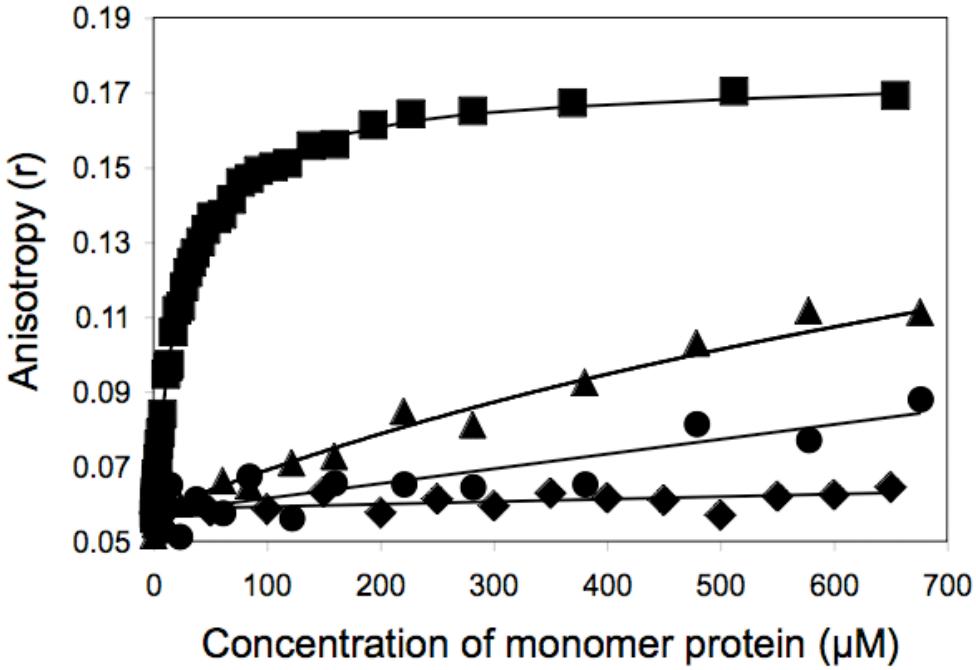


Fig. 4-6. Fluorescence anisotropy data of 6-iodoacetamido-fluorescein-labeled KaiC peptides as a function of KaiA180C (squares), wild-type KaiA (triangles), KaiA281 (circles), and KaiA135N (diamonds). The KaiC peptide consists of the C-terminal 32 residues of *S. elongatus* KaiC. The concentration of KaiC peptide was 100 μM.

Discussion

Among clock period mutants, both short and long period mutations have been obtained in *kaiC*; however, only short period mutants have been identified in *kaiB* and usually long period phenotypes result from mutagenesis of *kaiA* (57, 138). This phenomenon is probably related to the ability of these proteins to affect the autokinase activity of KaiC. Mutations in KaiA are predicted to attenuate its ability to stimulate KaiC autophosphorylation. A long period *kaiA* mutant, *kaiA2* (also called A30a), which carries a R249H missense point mutation, has a 30 h period (57). This mutant displays a reduced KaiC autokinase rate and accumulated unphosphorylated KaiC, which has been suggested to affect the degradation rate of KaiC and *KaiBC* expression (99). Period mutations in KaiA or KaiC have also been shown to alter the interaction between these two proteins in yeast, even though the results were somewhat contradictory (225).

The unusual short period *kaiA* deletion mutant, *kaiA281*, can be used for determining the relationships among circadian period, phosphorylation status of KaiC, and interactions among Kai proteins. If there is a direct correlation between KaiC phosphorylation status and the circadian period, the *kaiA281* short period mutant would be expected to show more accumulated phosphorylated KaiC. The time course analysis of KaiC protein samples, however, did not show notable differences in accumulation of phosphorylated KaiC between *kaiA281*-complemented *kaiA* insertional null strain and its corresponding wild-type complementation strain (Fig. 4-5A). Because the period difference between these two strains is less than 1 hour, the time course analysis may not have sufficient resolution to reveal differences in KaiC phosphorylation status. KaiC

autokinase assay is a more quantitative way to indirectly check the difference between KaiA281 and wild-type KaiA in stimulating KaiC autophosphorylation. Preliminary data showed that when KaiB was absent, the sample containing purified KaiA281 did show slightly increased stimulation of KaiC autophosphorylation activity compared with wild-type KaiA. More strikingly, in the addition of KaiB the KaiC dephosphorylation process appeared to be slowed down when mixed with KaiA281 instead of KaiA (YongIck Kim, Andy LiWang Lab, personal communications). It has been reported that long period KaiA mutants, including A30a/KaiA2, display a shorter lifetime of phosphorylated KaiC, but do not significantly differ from wild-type strains in the turnover rate of the unphosphorylated KaiC (80). The prolonged KaiC phosphorylation state in the presence of KaiA281 might also reflect differences in the degradation rate of KaiC in addition to altered KaiC phosphorylation status since both are likely to be important for maintaining the circadian period.

The fluorescence anisotropy data suggest that the loss of the three residues alters the interaction between KaiA and KaiC (Fig. 4-6). KaiA binds specifically to the C-terminal domain of KaiC possibly at two interfaces: a C-terminal peptide referred as CII ATP binding domain (CIIABD) and the ATP binding pocket (100, 101). Both *kaiA2* and *kaiA281* carry mutations that affect the C-terminal domain of KaiA, which is responsible for interaction with KaiC and stimulation of KaiC autophosphorylation (98, 226). The R249H substitution encoded by *kaiA2* is right in the KaiA-KaiC interaction interface and very close to the ATP binding pocket of KaiC; in contrast, the last three residues of KaiA are mapped at the edge of one of the dimer interfaces (88).

Two residues of the three that are missing in KaiA281, R282 and E283 are well conserved among cyanobacterial KaiA proteins, as shown in multiple protein alignment (data not shown), while T284 is not. The basic arginine and acidic glutamate on two different monomers in a KaiA dimer may form salt bridges**,** which may participate in stabilizing dimer structure. Thus, the deletion of these residues probably results in minor conformational change of KaiA dimer, which in turn weakens the interaction between KaiA and KaiC and somehow enhances its ability to stimulate the autokinase activity of KaiC. Since the binding of KaiB to KaiC is KaiA dependent (227), the KaiA281 protein may not only be weakened in its interaction with KaiC, but also affect the binding of KaiB to KaiC and hence the dephosphorylation of KaiC.

The original *Mu*Cm *kaiA* insertion, 7G3GG5, shortens the period in both P*psbAI*::*lux* and P*kaiBC*::*lux* reporter strains by almost 1 h. The period difference between the reconstructed *Mu*Gm *kaiA* insertion mutant and its corresponding wild-type strain carrying a P*kaiBC*::*luc* reporter is diminished to around 0.5 h. The two complemented *kaiA* strains, one with wild-type *kaiA*, and the other with *kaiA281*, also show period differences of ~0.5 h with each other (Table 4-1 & Fig. 4-7). This difference may reflect the variations in the promoter activity and/or post-transcriptional regulation of different reporter constructs. We have observed previously that the period length differs among some wild-type reporter strains. The P*psbAI*::*lux* reporter strain (AMC1020) is ~1 h shorter in circadian period than the P*kaiBC*::*lux* reporter strain (AMC1300). AMC541, a P*kaiBC*::*luc* reporter, shows a ~25 h period, which is between the periods of AMC1020 and AMC1300 (Table 4-1). One technical difference between

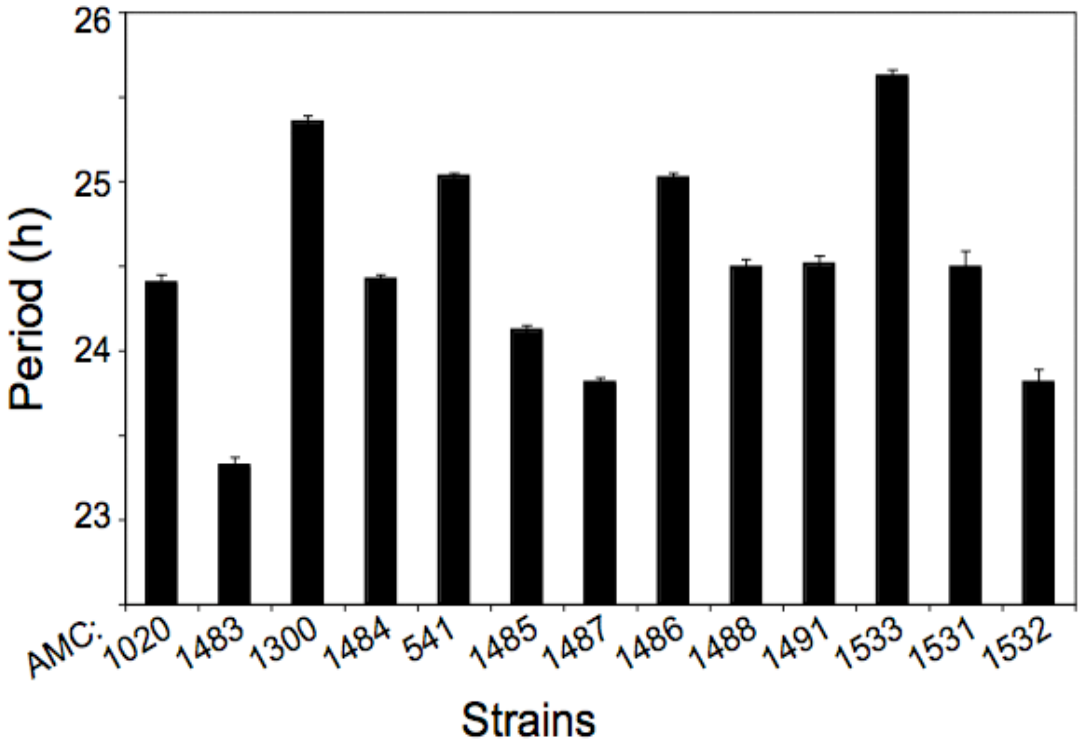


Fig. 4-7. Column chart of circadian period of cyanobacterial strains in Table 4-1. X-axis, cyanobacterial strains; Y-axis, circadian period in hours.

AMC541 and the other two reporter strains is that the substrate for the firefly luciferase (*luc*), luciferin, has to be provided exogenously.

Another interesting phenomenon we observed is the period difference between cells grown in high light and low light conditions (see Materials and methods) for some of the strains. As seen in Table 4-2 & Fig. 4-8, mini*Mu* insertion mutants AMC1483, AMC1484, and AMC1491 obey the Aschoff's rule for diurnal organisms (228), such that they exhibit a shorter circadian period in high light and longer period in low light. On the contrary, both P*kaiBC* reporter strains AMC1300 and AMC541, as well as AMC1531 and AMC1533 (mutant strains in which the negative element is disrupted), display a period significantly longer under high light conditions than in low light. Wild-type P*psbAI* reporter strain AMC1020, and reconstructed strains with the negative element intact (AMC1485, AMC1487, AMC1486, and AMC1488), as well as another negative element disruption strain AMC1532 showed no difference in period between high and low light conditions. The fact that even different wild-type reporter strains show different correlations between circadian period and light intensity suggests a need for closer analysis of circadian period phenotype, which has been overlooked in most previous studies. One consequence of the opposite response of a mutant strain and its wild-type partner, *e.g.*, AMC1491 and AMC541, is that their periods are much closer under low light than under high light (Table 4-2 & Fig. 4-8). Nonetheless, the relationship of circadian periods between mutant strains and their corresponding wild-type strains is constant, whether or not the basic light-dependent period difference in a given strain is considered in statistical analysis.

Table 4-2. Circadian periods of wild-type and mutated *kaiA* strains under high light and low light conditions.

PCC 7942 Strains	High Light Period±SEM‡ (h)	n	Low Light Period±SEM‡ (h)	n
AMC1020	24.38±0.13	8	24.41±0.06	31
AMC1483	23.07±0.05	26	23.54±0.05	31
AMC1300	25.53±0.05	20	25.17±0.03	21
AMC1484	24.31±0.03	24	24.52±0.04	24
AMC541	25.09±0.03	26	24.97±0.03	28
AMC1161	AR	8	AR	12
AMC1485	24.15±0.06	9	24.09±0.04	16
AMC1487	23.78±0.05	10	23.83±0.03	16
AMC1486	25.07±0.06	12	24.98±0.04	16
AMC1488	24.53±0.09	7	24.48±0.05	16
AMC1491	24.15±0.06	5	24.61±0.03	18
AMC1492	AR	12	AR	12
AMC1531	24.11±0.15	5	23.67±0.05	11
AMC1532	25.54±0.10	7	25.65±0.04	18
AMC1533	24.93±0.12	5	24.31±0.09	12
AMC1534	NA§	5	NA§	11

‡ SEM: standard error of the mean

§ NA: not available; unstable traces, phase advanced and low amplitude (see Fig. 4-4A)

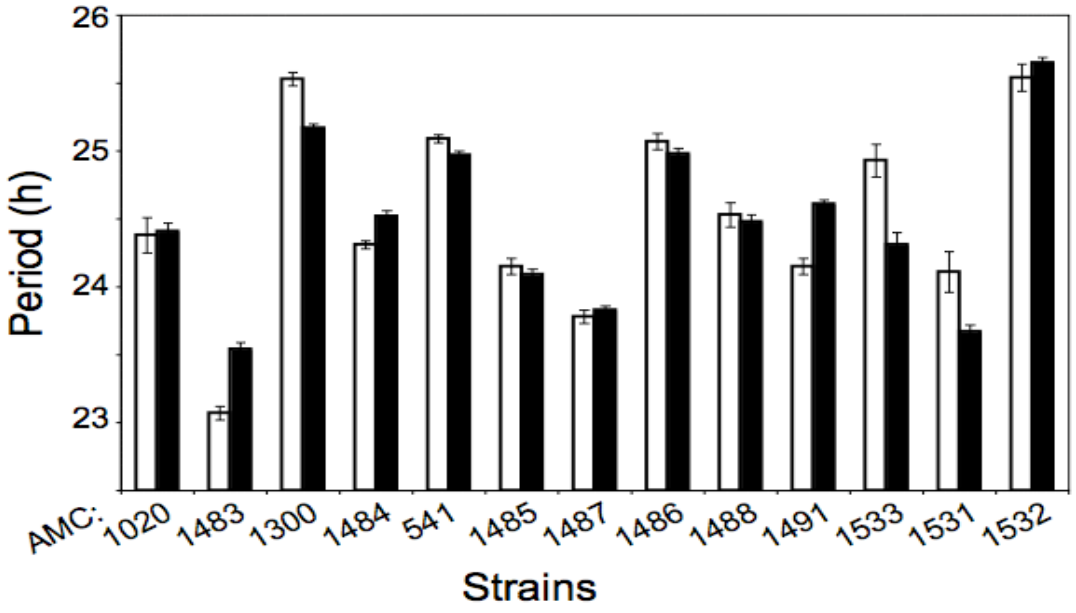


Fig. 4-8. Clustered column chart of circadian period of cyanobacterial strains in Table 4-2. X-axis, cyanobacterial strains; Y-axis, circadian period in hours. White bars, high light; black bars, low light.

The short period *kaiA281* allele does not lose its basic function as a central oscillator component. Our results suggest that *kaiA281* functions normally in enhancing *kaiBC* promoter activity and sustaining the circadian oscillation (Table 4-1 and Fig. 4-2). Nevertheless, KaiA281 is also different from wild-type KaiA in some other aspects, as it cannot complement AMC1492 completely, in which both *kaiA* and the negative element are disrupted by insertion, when provided ectopically (Fig. 4-4A). There might be a position effect of *kaiA* alleles (*cis* vs. *trans*); note that even the wild-type allele did not complement with a normal period in this elevated KaiB/KaiC situation. Ectopic *kaiA* or *kaiA281* was introduced into the NS1 locus, which is almost 180° from the native *kai* locus on the circular chromosome. In all cyanobacteria species that carry *kaiA*, it is always located immediate upstream of *kaiBC* (data not shown). The negative element of P*kaiBC* is located entirely within the coding region of *kaiA* in *S. elongatus*. Thus, this structural conservation of *kai* locus is likely to have significance in clock function. The ectopically expressed *kaiA* or *kaiA281* (*trans*) functions sufficiently well in AMC1486, AMC1488, and AMC1533, in which the oscillator is not otherwise compromised, to behave its *cis* counterpart. It is known that a promoter with peak expression in a completely opposite phase of P*kaiBC*, such as P*purF* (79), or even the heterologous *E. coli* promoter P*trc* (80), can successfully replace the original *kaiBC* promoter, expressed from a neutral site, to sustain a functional circadian clock in *S. elongatus*. Thus, there must be a 'buffering mechanism' in cyanobacterial clocks to tolerate certain alterations. In AMC1534, the combined effect of disrupting the negative element, together with a truncated KaiA expressed in *trans*, is probably beyond the robustness of the clock, and

thus, impairs circadian rhythms. Although they show robust rhythmicity, complemented strains AMC1486 and AMC1488 are different from the wild-type strain AMC541 in their circadian patterns of KaiC phosphorylation. The increased amount of phosphorylated KaiC across the time course may be another result of the ectopic expression of *kaiA* or *kaiA281*. This pattern is not seen in AMC1533, probably because of the presence of excess amount of unphosphorylated KaiC (Fig. 4-5).

The ratio of KaiA to KaiC seems to be important to sustain a normal circadian period (105). Here we show that when KaiA to KaiC ratio is high due to an extra copy of *kaiA*, more phosphorylated KaiC accumulates and circadian period is shortened (AMC1485 & AMC1487). Conversely, when the ratio is lowered due to the disruption of the *kaiBC* upstream negative element, there is more unphosphorylated KaiC and a longer period (AMC1533). We propose that this correlation is generally true in the cyanobacterial circadian clock. The *kaiA281* allele also displays a dominant effect: even with a wild-type *kaiA* allele present, a strain that encodes *kaiA281* always has a shorter circadian period than the wild-type strain on which it is based (Table 4-1 & Fig. 4-7). The disruption of the negative element upstream of *kaiBC* promoter alone slightly extends the period (AMC1533). In the presence of a *kaiA281* allele, however, the long period is masked by the short period phenotype, as shown in the *Mu* insertional strains (Table 4-1).

No correlation between the peak time of accumulation of phosphorylated KaiC and rhythmic bioluminescence traces can be deduced from our results. AMC1534, mimicking the original mini*Mu* insertion strain, is phase advanced in its

bioluminescence rhythms, but not in KaiC phosphorylation pattern. On the contrary, AMC1533 shows a normal phase in bioluminescence traces, but is phase delayed in time course immunoblot analysis (Fig. 4-5B). Previously, a point mutation of *kaiC* (C1265T) has been shown to cause the loss of the rhythmic KaiC accumulation and reduced circadian phosphorylation pattern. However, the *pr1* mutant still possessed robust bioluminescence rhythms with a wild-type period, but 3-4 hours phase advanced (109). Thus, there must be some other mechanisms for phase regulation, probably through a circadian output pathway. Nonetheless, the overall circadian pattern and period length of the KaiC phosphorylation cycle are conserved in these strains. Because the period difference among all the strains in this study (except AMC1534) is less than 1 hour, the KaiC phosphorylation and dephosphorylation cycle always recurs after approximately 24 hours despite variations in phosphorylation pattern during the cycle.

Materials and Methods

Cyanobacterial strains, media, and culture conditions

The cyanobacterial strains used in this study are summarized in Table 4-3. All wild-type reporter and mutant strains were created in *Synechococcus elongatus* PCC 7942. Cyanobacterial strains were grown in BG-11 medium (223) under continuous light conditions (~70 μE/m^2s) at 30°C with appropriate antibiotics. Two different versions of luciferase reporter genes were used: 1) firefly luciferase (*luc*) in neutral site II (NS2); 2) *Vibrio harveyi* luciferase (*luxAB*) in the neutral site I (NS1) locus and the aldehyde substrate synthesis genes *luxCDE* in NS2 (52). Neutral sites mediate homologous recombination with *S. elongatus* chromosome and cause no apparent phenotypes (52).

Plasmid construction

Plasmids are described in Table 4-4. Unless otherwise stated, plasmids were constructed in *Escherichia coli* strain DH10B (Invitrogen, Carlsbad, CA). Plasmid pAM2246 (79) carries a copy of wild-type *kaiA* with its native promoter region in a NS1 vector. The wild-type *kaiA* was then converted to *kaiA281* using a QuickChange mutagenesis method (229) to engineer a stop codon, which resulted in pAM3434. Plasmid pAM3582 was constructed by inserting a PCR-amplified fragment, containing the original mini*Mu*Cm (GeneJumper™ kit; Invitrogen, Carlsbad, CA) transposon 7G3GG5 (65) inserted at the C-terminal coding region of *kaiA* with its flanking regions, into cloning vector pLitmus 29 (New England Biolabs, Beverly, MA). The *cat* gene in the center of

Table 4-3. Cyanobacterial strains used in Chapter IV.

Strain #	Genetic background	Ectopic *kaiA*	Ectopic *kaiA* characteristics	Reporter plasmid#	Reporter characteristics	Source or reference
AMC541	Wild-type	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	Ditty et al. (2003)
AMC702	*kaiA* in-frame deletion	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	Ditty et al. (2005)
AMC705	*kaiBC* in-frame deletion	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	Ditty et al. (2005)
AMC1020	Wild-type	None	None	pAM1501 (NS1); pAM1619 (NS2)	P*psbAI*::*luxAB*, Sp^R; P*psbAI*::*luxCDE*, Km^R	Andersson et al. (2000)
AMC1161	*kaiA* ΩKm insertion	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	Ditty et al. (2005)
AMC1300	Wild-type	None	None	pAM1887 (NS1); PAM1619 (NS2)	P*kaiBC*::*luxAB*, Sp^R; P*psbAI*::*luxCDE*, Km^R	Lab collection
AMC1483	*kaiA Mu*Cm insertion	None	None	pAM1501 (NS1); pAM1619 (NS2)	P*psbAI*::*luxAB*, Sp^R; P*psbAI*::*luxCDE*, Km^R	This study
AMC1484	*kaiA Mu*Cm insertion	None	None	pAM1887 (NS1); pAM1619 (NS2)	P*kaiBC*::*luxAB*, Sp^R; P*psbAI*::*luxCDE*, Km^R	This study
AMC1485	Wild-type	pAM2246 (NS1)	P*kaiA*::*kaiA*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1486	*kaiA* ΩKm insertion	pAM2246 (NS1)	P*kaiA*::*kaiA*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1487	Wild-type	pAM3434 (NS1)	P*kaiA*::*kaiA281*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1488	*kaiA* ΩKm insertion	pAM3434 (NS1)	P*kaiA*::*kaiA281*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1491	*kaiA Mu*Gm insertion	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1492	*kaiA* ΩKm & *Mu*Gm insertions	None	None	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1531	*kaiA Mu*Gm insertion	pAM2246 (NS1)	P*kaiA*::*kaiA*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1532	*kaiA Mu*Gm insertion	pAM3434 (NS1)	P*kaiA*::*kaiA281*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1533	*kaiA* ΩKm & *Mu*Gm insertion	pAM2246 (NS1)	P*kaiA*::*kaiA*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study
AMC1534	*kaiA* ΩKm & *Mu*Gm insertion	pAM3434 (NS1)	P*kaiA*::*kaiA281*	pAM2105 (NS2)	P*kaiBC*::*luc*, Cm^R	This study

∗Cm^R, chloramphenicol; Km^R, kanamycin; Sp^R, spectinomycin.

mini*Mu*Cm was then substituted by a copy of the gentamycin resistance gene, *aacA*, from pAM3515 (pHP45ΩGm) to create pAM3613. The coding sequence of *kaiA* was inserted into pET-32a(+) (Novagen) at *Hind*III/*Eco*RV sites to build pAM3633. The last three amino acid residues of *kaiA* in pAM3633 were then deleted using the Quick-Change strategy to construct pAM3630.

Table 4-4. Bacterial plasmids used in Chapter IV.

Plasmid	Characteristics	Antibiotic resistance	Source or reference
pAM2246	P*kaiA*::*kaiA* (NSI)	Sp^R, Sm^R	Ditty et al. (2005)
pAM3434	P*kaiA*::*kaiA281* (NSI)	Sp^R, Sm^R	This study
pAM3515	pHP45ΩGm	Gm^R, Ap^R	This study
pAM3582	*kaiA::Mu*Cm	Gm^R, Ap^R	This study
pAM3613	*kaiA::Mu*Gm	Gm^R, Ap^R	This study
pAM3633	pET-32a(+)-kaiA	Ap^R	Vakonakis et al. (2004)
pAM3630	pET-32a(+)-kaiA281	Ap^R	This study

*Cm^R, chloramphenicol; Km^R, kanamycin; Sp^R, spectinomycin; Sm^R, streptomycin.

Bioluminescence assay and data analysis

The measurement of bioluminescence from the reporter strains was performed as described previously (65, 79, 202). The acquired bioluminescence data were processed and graphed by the Import and Analysis (I and A) Excel macro set (S. A. Kay Laboratory, The Scripps Research Institute, La Jolla, CA). To calculate the circadian periods of these strains, the "I and A" processed data sets were then exported to BRASS (Biological Rhythms Analysis Software System; A. J. Millar laboratory, University of Edinburgh, Scotland, UK), which is an Excel interface for FFT-NLLS (a Fast Fourier Transform statistical suite of programs; Dr. Martin Straume). The measurement of each strain was conducted in at least three independent experiments. Tables 4-1 & 4-2 show

representative data summarized from one of the two experiments except AMC541, which is measured from both experiments and the period was the same in both. Due to limited space, not all strains can be assayed simultaneously. For each experiment, the data of at least four circadian cycles in constant light were used to calculate the period. Standard error of the mean (SEM) was calculated in Excel (Microsoft, Redmond, WA). High light and low light data shown in Table 4-2 were combined in Table 4-1. Here 'high light' means the outside wells (#2, #3, #10, and #11) in one row of 96-well black plates used in the measurement of bioluminescence; while 'low light' refers to the inner wells in the same row (#5, #6, #7, and #8) (59).

Immunoblot analysis

Protein sample preparation and immunoblot analysis for single time point experiments were performed as describe previously with slight modifications (79). Cultures (O.D. 0.6-1.0 at 750 nm) for time course analysis were first entrained in opposite phases with at least two cycles of 12-h/12-h of either light/dark or dark/light, and then released to constant light conditions (30-50 $\mu E/m^2 s$). Samples were taken either every 4 hours for 12 to 24 hours or every 1 hour for 4 hours depending on the experimental design.

Peptide purification and fluorescein labeling

Segments of the *S. elongatus kaiC* gene that encode the desired peptide sequences were cloned in a pET-32a(+) vector (Novagen), thereby creating thioredoxin–polyHis–peptide fusions (100, 101). *Escherichia coli* BL21(DE3) (Novagen) was transformed with the

resulting plasmids and grown in Luria broth. Expression of peptide fusion constructs was induced by adding isopropyl-β-D-thiogalactopyranoside (Calbiochem) to a final concentration of 1 mM. The cells were harvested after 4 h, resuspended in a buffer containing 50 mM NaCl and 20 mM Tris-HCl (pH 7.4), and lysed. Cell lysates were separated by centrifugation, and the fusion peptide was purified from the supernatant fraction by metal affinity chromatography. The fusion peptide was labeled with fluorescein at the N terminus by the manufacturer's protocol (Molecular Probes, Eugene, OR). The sample was buffer exchanged to 50 mM NaCl, 20 mM Tris-HCl, pH 7.4, and the peptide was cleaved from the affinity tag by using enterokinase (Novagen). Cleavage by this enzyme results in the addition of three non-KaiC-derived residues (AMC) at the N terminus of the peptide. Peptides were isolated by reverse phase chromatography and lyophilized. Peptide identity and purity were confirmed by matrix-assisted laser desorption ionization-time-of-flight spectroscopy, and quantification was carried out by measuring the UV absorbance of fluorescein.

Fluorescence Anisotropy-Based Binding Experiments

Fluorescence anisotropy experiments were carried out with an ISS PC1 photon counting spectrofluorometer with fluorescein-labeled KaiC peptide. The peptide concentration in all cases was fixed at 100 nM in 20 mM Tris-HCl, 150 mM NaCl, 0.5 mM EDTA, pH 8.0. The initial volume of 1.8 ml of fluorescein-labeled KaiC peptide was used. Up to 2.2 ml of 1.2 mM KaiA-protein stocks were added in known aliquots up to a total concentration of ~700 μM protein.

CHAPTER V

THE LARGE ENDOGENOUS PLASMID OF *Synechococcus elongatus* PCC 7942, pANL

Introduction

As a group of photoautotrophic eubacteria, cyanobacteria are characterized by oxygenic photosynthesis, morphological diversity, as well as their blue-green color. Like most other prokaryotes, cyanobacteria usually have a single circular chromosome. In addition, many cyanobacterial strains also contain one or more endogenous plasmids (230-236). The complete sequences of about 30 cyanobacterial plasmids have been deposited into the GenBank database. They belong to strains of six cyanobacterial genera: *Anabaena*, *Leptolyngbya*, *Microcystis*, *Nostoc*, *Synechococcus*, and *Synechocystis*, with a size range from around 1.5 kb (pPBS1, GenBank accession No AF176225) to over 400 kb (pCC7120alpha, GenBank accession No BA000020). There are 7 plasmids in *Synechocystis* sp. PCC 6803 (237-240) and 6 in *Nostoc* (also known as *Anabaena*) sp. PCC 7120 (241). One of the species with no plasmid detected is the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1 (168).

Many cyanobacterial cellular functions have long been speculated to be determined by plasmids, such as antibiotics resistance, heavy-metal resistance, toxin production, and gas vacuolation (242, 243). Nonetheless, since most of the cyanobacterial plasmids are still cryptic, none of these biological roles has been experimentally verified.

Two endogenous plasmids, pANS (also called pUH24) and pANL (also called pUH25) in the fresh water unicellular cyanobacterium *Synechococcus elongatus* PCC 7942, formerly *Anacystis nidulans* R2, have been identified and fully sequenced (65, 244, 245); the complete sequence of pANL was determined as part of this dissertation research. These two plasmids were not only identified in strains closely related to *S. elongatus* PCC 7942, such as *S. elongatus* PCC 6301, but also found in *Cyanobium* (*Synechococcus*) sp. PCC 6707, which is different from *S. elongatus* PCC 7942 in both genome size and base composition (246, 247).

The small plasmid, pANS (GenBank accession No S89470), is 7,835 bp long and encodes 8 putative open reading frames (ORFs) (245). Because pANS can be easily cured and it has a relatively small size, extensive studies have been employed to successfully construct shuttle vectors between *E. coli* and *Synechococcus* strains (133, 248, 249). The large plasmid, pANL (GenBank accession No AF441790), has also been the subject of a number of investigations. Laudenbach and colleagues started the functional analysis of pANL by constructing its physical map of restriction sites and identifying the origin of replication (250, 251). Three *Bam*HI fragments of pANL were then sequenced (GenBank accession Nos U20224, U23436 and AF176824; 3.4 kb, 3.8 kb and 11.2 kb, respectively). A number of genes on these fragments were found to be functionally related to sulfur metabolism of *S. elongatus* PCC 7942, thus designated as sulfur-regulated plasmid-encoded (srp) genes (252, 253). This report provided the first clue of the possible function of a cyanobacterial plasmid. The complete pANL sequence revealed that an ~5 kb sequence fragment (GenBank accession No U70379), carrying a

putative *narL* gene and genes encoding a two-component system, is not on the chromosome as originally thought, but on pANL instead. Although the spontaneous loss (curing) of pANS has been reported (242), attempts to cure pANL were not successful (251).

Here we report: the complete nucleotide sequence of pANL determined by a combination of cosmid-end sequencing, transposon-mediated sequencing, primer-walking, as well as JGI shotgun sequencing; manual annotation of the putative ORFs by protein-coding region prediction and gene function assignment; identification of essential regions *via* cosmid-mediated deletion mapping and transposon-based mutagenesis; confirmation of the function of two toxin-antitoxin cassettes; attempts to cure pANL by deletion cloning and counter selection; and screening of pANL cosmid-based deletion mutants for circadian phenotypes in cyanobacterial reporter strains. Four modules were determined in pANL based on gene organization and functional assignment: the replication region, a signal transduction region, a plasmid maintenance region, and a sulfur-regulated region. Our results show that the difficulty to cure pANL is mainly due to two toxin-antitoxin cassettes in the plasmid maintenance region. The genes on pANL are not likely to be involved in circadian rhythmicity in *S. elongatus* PCC 7942.

Results and Discussion

Determination of the complete sequence of pANL

A 960-cosmid genomic library provided the templates for transposon-mediated mutagenesis and sequencing of the genome of *S. elongatus* PCC 7942 (65). Alignment of end sequences of these cosmids with published *S. elongatus* PCC 7942 sequences in GenBank showed that many of them were located on pANL, the large plasmid, instead of the chromosome, which means that pANL fragments were cloned into the cosmid vector during library construction. A hybridization assay using two large *Eco*RI fragments (~12 kb and ~16 kb, respectively) of cosmid 6G1 as probes to screen the genomic cosmid library identified pANL-related cosmids (Fig. 5-1). To our surprise, around one third of the cosmids in the library (~32%, 307 cosmids) carry pANL sequence. Because pANL is likely a low copy number plasmid due to the presence of the plasmid maintenance region mentioned below, the amount of pANL DNA in the cells should not reach 1/3 of the total DNA. One possible explanation is that most of the *Sau3*AI partial digestion fragments of pANL (~46 kb) are within the 30-40 kb size range of fragments that can be packaged into cosmid vector SuperCos I, while the majority of the chromosome *Sau3*AI fragments were discarded because of inappropriate size. Alternatively, pANL might carry some features that provide advantage and priority for packaging. As a result of the overrepresentation of pANL in the library, hundreds of cosmid end sequences could be used to provide good sequence coverage of pANL. Based on the analysis of these cosmid end sequences, one cosmid of pANL–origin, 2A8,

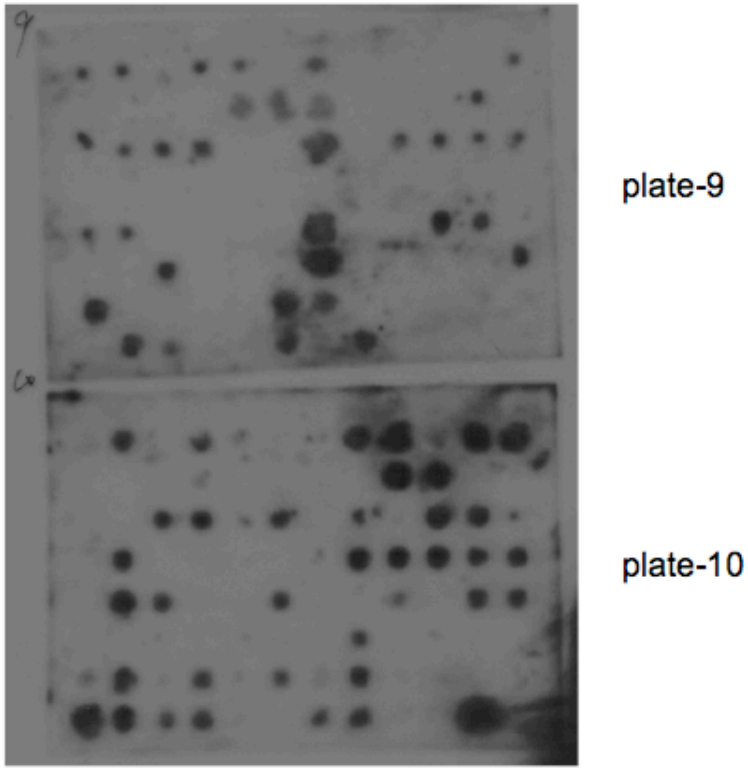


Fig. 5-1. Overrepresentation of pANL sequence in the genomic cosmid library of *S. elongatus* PCC 7942. Representative hybridization results of 96-well plates #9 & #10 are shown. The 12-kb *Eco*RI fragment of cosmid 6G1 was used as the hybridization probe. The experiment was also repeated with the 16-kb *Eco*RI fragment of 6G1 (data not shown).

was predicted to cover most of the gaps in the assembly, and was chosen as a template for *in vitro* mutagenesis and sequencing to complete pANL (65). The remaining gaps in the pANL sequence were filled by primer walking. The complete sequence of pANL was then annotated and submitted to GenBank (accession No AF441790.1). Later, the pANL complete sequence was refined with JGI shotgun sequences (65) as a reference and subsequently updated (GenBank accession No AF441790.2).

General features of pANL

The complete sequence of pANL is 46,366 bp in length with an ~53% GC content, lower than that of the chromosome (~55%, GenBank accession No CP000100) and pANS (~59%). There are 58 putative ORFs on pANL (Fig. 5-2 & Table 5-1). On average, there are ~1.25 coding regions per kb of the pANL sequence. The coding regions, ~54% in GC content, cover ~83% of the total sequence. The GC content in the intergenic regions is ~47%. As shown in Table 5-1, 27 ORFs are assigned with functions based on their significant hits in BLAST searches for conserved domains and close homologues; 26 ORFs encode conserved hypothetical proteins with homologues found (data not shown); 2 ORFs encoding small proteins (approximately 100 amino acid residues) are hypothetical ORFs without any significant hits in the protein database; additionally, there are three other small ORFs, pANL09, pANL26, and pANL27, that share similarity only with each other in both nucleotide and protein sequences, but with nothing else in the database. Because the GC content of these three ORFs is close to that of the total

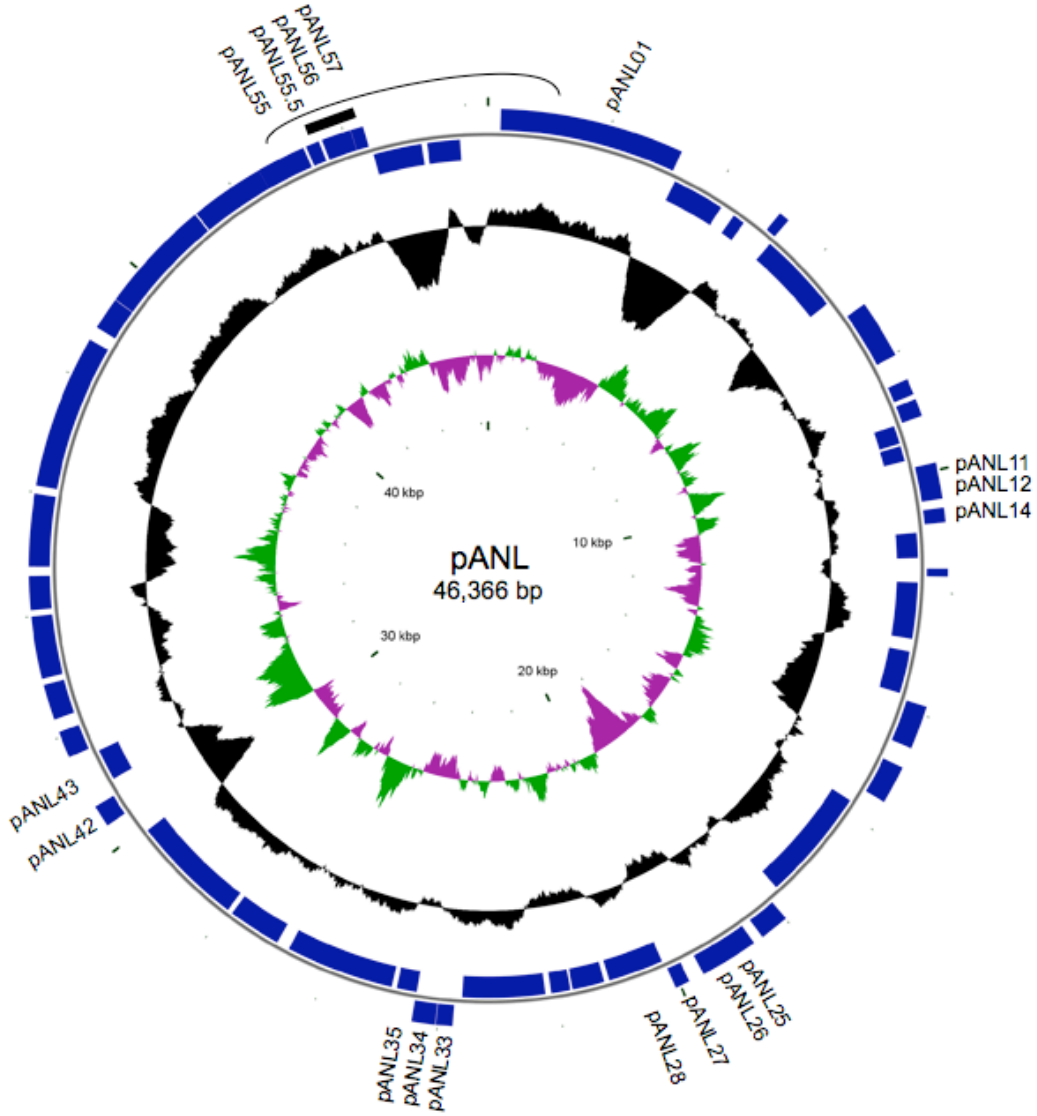


Fig. 5-2. Circular representation of pANL. Genomic features are shown concentrically. Shown as circles from the outermost to the inner most: putative Open Reading Frames (ORFs) in blue, GC content in black, GC skew ([G-C]/[G+C]) in green and purple, and the scale in kbp. Clockwise ORFs are shown outside the baseline, while ORFs expressed in the counterclockwise direction are inside. ORF01, ORF42, ORF43, and ORFs located at the borders of the four regions are labeled; For the second circle, peaks outside the center line correspond to regions with GC content above the average GC content value (0.5289), while the inside peaks refer to regions with below average GC content; In the third circle, GC skew values greater than the average (-0.0459) are in green, whereas GC skew values below the average are in purple. The black rectangle outside the ORF circle at top indicates the replication origin. The black arc above the rectangle shows the *Kpn*I-*Bam*HI fragment where the replication origin locates. Plotting of GC content and GC skew used a window size of 500 and a step of 1. The figure was generated using the CGVIEW program (161) with default settings.

Table 5-1. List of putative ORFs on pANL and their functional assignments.

ORF	Coordinates	% (G+C)	Size (aa)	Putative product	Conserved domains (CD)	Region* (aa)	% CD aligned†	E-value
pANL01	205-3204	57	999	Replication initiator protein RepA	COG1763 (MobB)	304-454	93	0.077
pANL02	3314-4228c	36	304	Conserved hypothetical protein	-	-	-	-
pANL03	4484-4699c	57	71	Conserved hypothetical protein	-	-	-	-
pANL04	5077-5250	59	57	Hypothetical protein	-	-	-	-
pANL05	5275-6699c	54	474	Conserved protein	COG5361	26-473	95	2e-71
pANL06	7072-8031	50	319	Conserved hypothetical protein	-	-	-	-
pANL07	8462-8722	56	86	Hypothetical protein	-	-	-	-
pANL08	8807-9094	59	95	Hypothetical protein	-	-	-	-
pANL09	9101-9418c	50	105	Conserved hypothetical protein	-	-	-	-
pANL10	9463-9732c	56	89	C-terminal domain of recombinase	cd01182 (INT_REC_C)	1-55	33	5e-7
pANL11	9877-10170	51	97	Conserved protein	COG2929	1-86	96	2e-11
pANL12	10136-10465	58	109	Conserved protein	COG3514	35-108	81	2e-11
pANL14	10610-10849	54	79	Hypothetical protein	-	-	-	-
pANL15	10980-11411c	56	143	Transcriptional regulator	COG3432	1-78	83	6
pANL15.5	11601-11726	45	41	Conserved hypothetical protein	-	-	-	-
pANL16	11835-12809c	58	324	Zn-dependent oxidoreductase	COG0604 (Qor)	1-324	100	1e-9
pANL18	13037-13762c	47	241	Type 1 glutamine amidotransferase	cd03139 (GATase1_PfpI_2)	6-184	100	1e-46
pANL19	13855-14562	43	235	One-component signal transduction protein	Smart00065 (GAF)/ Smart00421 (HTH_LUXR)	24-161/ 176-231	89/ 97	1e-10/ 5e-12
pANL21	14905-15528	55	207	Conserved protein	COG3544	49-207	78	8e-17
pANL22	15804-17120c	59	438	Histidine protein kinase	Smart00387 (HATPase_C)	330-433	100	1e-25
pANL23	17089-17763c	55	224	Response regulator	COG0745 (OmpR)	2-223	98	4e-51
pANL24	17986-18468	55	160	Conserved hypothetical protein	-	-	-	-
pANL25	18632-19222	56	196	Transcriptional regulator	Smart00420 (HTH_DEOR)	44-90	87	0.82
pANL26	19197-19586	48	129	Conserved hypothetical protein	-	-	-	-
pANL27	19841-20083	50	80	Conserved hypothetical protein	-	-	-	-
pANL28	20113-21075c	56	320	Site-specific recombinase	COG4974 (XerD)	24-318	92	2e-26
pANL29	21165-21515c	50	116	PemK-like toxin (SepT1)	Pfam02452 (PemK)	4-113	100	2e-4
pANL29.5	21506-21721c	49	71	Antidote (SepA1)	-	-	-	-
pANL30	21765-22094c	52	109	Nucleotidyltransferase	COG1669	18-101	80	6e-13
pANL31	22197-23021c	56	274	Partitioning protein ParB	Smart00470	36-76	45	0.87
pANL32	23018-23629c	56	203	Partitioning protein ParA	COG1192 (Soj)	4-201	99	1e-11

Table 5-1. Continued.

ORF	Coordinates	% (G+C)	Size (aa)	Putative product	Conserved domains (CD)	Region* (aa)	% CD aligned†	E-value
pANL33	23763-24029	55	88	Antitoxin (SepA2)	-	-	-	-
pANL34	24040-24423	52	127	VapC-like toxin (SepT2)	COG1487 (VapC)	1-123	97	3e-11
pANL35	24420-24770c	52	116	Rhodanese-related sulfurtransferase	COG0607 (PspE)	17-112	88	1e-16
pANL36	24860-26740c	55	626	Cysteine desulfurase	COG1104 (NifS)	243-619	96	2e-48
pANL38	26975-27895c	55	306	Conserved hypothetical protein SrpI	-	-	-	-
pANL39	27990-28949c	57	319	Serine acetyltransferase SrpH	COG1045 (CysE)	134-301	82	9e-41
pANL40	28946-29935c	56	329	Cysteine synthase SrpG	COG0031 (CysK)	8-314	100	5e-97
pANL42	30391-30750	48	119	Conserved hypothetical protein	-	-	-	-
pANL43	31000-31545c	53	181	Conserved protein	COG1791	47-175	75	4e-11
pANL44	31647-32081	54	144	Transcriptional regulator	COG1959	3-144	96	5e-21
pANL45	32300-32878	53	192	Carbonic anhydrase	cd03379	28-182	100	6e-49
pANL46	32980-33999	50	339	Catalase SrpA	Pfam00199 (catalase)	77-355	76	6e-31
pANL47	34100-34648	56	182	Conserved protein SrpB	Pfam02308 (MgtC family)	18-73	43	3e-13
pANL48	34800-35981	46	393	Chromate transporter SrpC	Pfam02417/ COG2059	14-190/ 221-378	100/ 82	7e-46 /4e-12
pANL49	36112-37101	57	329	Cysteine synthase SrpD	COG0031 (CysK)	4-304	100	8e-96
pANL50	37098-38624	59	508	Gamma-glutamyltranspeptidase SrpE	Pfam01019	31-503	100	8e-121
pANL51	38849-39376	57	175	Conserved hypothetic protein SrpF	-	-	-	-
pANL52	39387-40448	58	353	Acyl-CoA dehydrogenase SrpJ	COG1960 (CaiA)	4-324	87	1e-19
pANL53	40445-41257	58	270	Nitrate/sulfonate/bicarbonate ABC transporter, ATPase subunit SrpK	COG1116 (TauB)	17-264	97	3e-68
pANL54	41283-42506	56	407	Nitrate/sulfonate/bicarbonate ABC transporter, periplasmic subunit SrpL	COG0715 (TauA)	12-350	97	4e-44
pANL55	42508-43308	57	266	Nitrate/sulfonate/bicarbonate ABC transporter, permease subunit SrpM	COG0600 (TauC)	54-266	82	4e-30
pANL55.5	43355-43579	59	74	Conserved hypothetical protein	-	-	-	-
pANL56	43634-43912	55	92	Conserved protein	COG2929	1-85	95	6e-11
pANL56.5	43866-44117	57	83	Conserved protein	COG3514	22-78	61	2e-4
pANL57	44122-44325	54	67	Conserved hypothetical protein	-	-	-	-
pANL58	44375-45211c	38	278	Conserved hypothetical protein	-	-	-	-
pANL59	45309-45881c	58	190	Conserved hypothetical protein	-	-	-	-

* Region of the pANL protein that shows significant similarity to the relevant conserved domain in GenBank.

† Percent of the length of conserved domain that can be aligned with pANL-encoded protein in the CD search (rpsblast).

coding sequences, they are probably pANL-specific loci. Most of the proteins encoded by pANL have cyanobacterial homologues. There are 27 ORFs that have significant hits in the chromosome of *S. elongatus* PCC 7942 (data not shown). Both homologues of pANL03 are *S. elongatus* PCC 7942 chromosomal genes. On the contrary, pANL02, pANL25, pANL51, and pANL58 did not hit any cyanobacterial proteins in a BLAST search. The GC content of two of them, pANL02 and pANL58, is much lower than the other ORFs (36% and 38%, respectively). They are likely the consequences of lateral gene transfer events.

Repetitive sequences

An octameric highly iterated palindrome (HIP1), GCGATCGC, is over-represented in DNA sequences of many cyanobacterial strains, including *Synechococcus* species (165). The chromosome of *S. elongatus* PCC 7942 has 7,323 HIP1 sites, around 1 site in 368 bp sequence. There are 95 HIP1 sites found in pANL, averaging 1 site every 488 bp, higher than that in the chromosome, but similar to that of pANS (1 site/~412 bp). The distribution of HIP1 sites around pANL is roughly random, with a few hot spots and blank regions. One thing interesting is that the longest ~3.5 kb HIP1 gap corresponds to the putative replication of origin (data not shown).

Self-alignment of pANL sequence was performed with the 'BLAST 2 SEQUENCES' tool (bl2seq) to identify repetitive sequences (254). Several pairs of relatively large direct or inverse repeats were identified, all located in

Table 5-2. Large repetitive sequence pairs on pANL.

Repeat pair*	Type	Coordinates Sequence 1	Coordinates Sequence 2	Size (bp)	% identity
#1	direct	9517-9762c	20119-20364c	246	84
#2	direct	43412-43560	44161-44309	149	87
#3	direct	19474-19531	19956-20013	58	82
#4	direct	10398-10423	44062-44087	26	92
#5	inverse	9150-9341c	19828-20019	192	75

* Only repeat pairs larger than 20 bp are listed.

putative coding regions (Table 5-2). The first pair of 246 bp direct repeats (#1) correlates to coding sequences of pANL10 and pANL28. This pair of direct repeats is ~10 kb apart and shares ~84% sequence identity. The protein encoded by pANL10 shows ~95% identity to the C-terminal catalytic domain of the pANL28 protein, a homologue of an integrase or recombinase. The second direct repeat pair (#2), close to the putative replication origin, is 149 bp in length and shares 87% identity. These repeats, just ~0.6 kb apart, are located in two paralogous ORFs, pANL55.5 and pANL57, respectively. The third pair of direct repeats (#3), 58 bp long and sharing 82% identity, is located in adjacent ORFs, pANL26 and pANL27. The proteins encoded by these two genes are well conserved, accounting for 57% identity over the majority of their amino acid sequences. They both also show high similarity to the pANL09 protein (48% and 60%, respectively). A pair of inverse repeats (#5) was detected in ORF09 and ORF27. They are 192 bp long and have ~75% sequence identity. Both pANL26 and pANL27 are in the same orientation, while pANL09 has opposite orientation and is around 10 kb away from them. These three ORFs possibly resulted from successive duplication and inversion

events. The repeat sequences in these ORFs are very close to the direct repeats (#1) in pANL10 and pANL28. Together, they are likely the sites of recombination recognized by site-specific recombinases. There is another pair of repeats (#4) in pANL12 and pANL63 that is only 26 bp long, but shares 92% identity. All five pairs of repetitive sequences may be related to mobile elements and can serve as structural separators for different regions of pANL.

Taken together, pANL is likely to be divided into four modules by these structural separators and functional assignment of putative genes: the replication region; the signal transduction region; the plasmid maintenance region; and the sulfur-regulated region.

The replication region

This module starts from pANL55.5 and ends at pANL12, including 18 ORFs and several pairs of repeats. The replication origin of pANL has been narrowed down to an ~11.7 kb *Bam*HI fragment in this region (250). Even though two replication origins in two adjacent *Kpn*I fragments in this part of pANL were proposed (251), our experiment data (Table 5-3) suggest that there is only one replication origin (~1.4 kb) located in the *Kpn*I-*Bam*HI fragment that is closest to the first nucleotide of the pANL sequence (Fig. 5-2).

At the borders of this region, there are two pairs of overlapping genes. One pair of overlapping ORFs, pANL56 and pANL56.5, is very similar to the other pair of overlapping genes, pANL11 and pANL12, in their size and structure, as well as encoded

Table 5-3. Summary of cosmid-based deletion analysis of pANL.

Cosmids introduced*	SuperCos I vector†	Region shared by both pANL and cosmid‡	pANL region absent from cosmid§	ORFs not in cosmid	Coordinates of region not in cosmid (bp)
9H1	+	+	-	01-03	1776-4739
8A9	+	+	-	01-15	696-11492
2H7	+	+	-	06-23	6841-17582
2C10	+	+	+	18-30	13446-22164
9D9	+	+	+	19-30	13891-22171
2F10	+	+	-	21-29	14662-21293
2D12	+	+	+	22-50	15765-37343
2G5	+	+	+	23-34	17734-24298
9G2	+	+	+	31-36	22960-25049
4D1	+	+	+	31-42	22149-30520
4F10	+	+	+	32-36	23346-25770
2B10	+	+	-	33-39	23818-28499
4H2	+	+	-	33-46	23818-33455
3F2	+	+	+	38-56	27497-43702
10F3	+	+	+	43-16	31466-12193
2A8	+	+	+	45-59	32085-45552
2B4	+	+	-	46-48	33890-35060
2D11	+	+	-	48-50	35811-37343
9E12	+	+	+	48-56.5	35196-43990
9H7	+	+	+	48-59	35060-45705
6H1	+	+	-	49-55	36765-42913
2D1	+	+	+	52-01	39665-886
2F9	+	+	+	54-05	42003-6375
7G4	+	+	+	53-02	40591-4172
9B9	+	+	+	54-59	41738-45913
4F8	+	+	-	57-06	44307-7174
7F8	+	+	-	59-12	45658-10775

* Cosmid that was introduced into *S. elongatus* wild-type strain AMC06.
† PCR results using SuperCos I specific primer pairs (sci-1×sci-2 & sci-3×sci-4) to detect the presence of the cosmid.
‡ PCR results using primer pairs specific for the shared region present in both pANL and the cosmid.
§ PCR results using primer pairs specific for the pANL region absent from the cosmid.
+, Positive PCR result from boiled samples of *S. elongatus* cells transformed with cosmid; -, negative PCR result from those boiled samples (see Results and Discussion).

protein sequences. The putative products of ORF pANL56.5 and pANL12 are ~45% identical. Both carry a copy of the COG3514 conserved domain with no function characterized. A pair of 26 bp direct repeats (#4) lies in their coding regions. The similarity between pANL56 and pANL11 is too low to be detected at the protein sequence level. However, they both contain a COG2929 domain conserved in bacteria. Above all, these two pairs of overlapping ORFs, which are conserved among bacteria, are probably the consequence of either a gene duplication or transposition event. Beside pANL56 and pANL56.5 are two small paralogs, pANL55.5 and pANL57, sharing about 80% protein sequence identity and a pair of 149 bp direct repeats (#2). Among the other ORFs, the product of pANL01 (999 aa) shows remote but detectable similarity to the replication initiator protein RepA of other cyanobacteria. As mentioned above, pANL58 and pANL02 are very low in GC content (38% and 36%, respectively), and have no cyanobacterial homologues, but many significant hits in other bacteria. This suggests that they might be transferred from other bacteria not very long ago. Putative ORF pANL03 is only 71 aa long in protein sequence. No other homologues have been identified for this ORF except two chromosomal genes from *S. elongatus* PCC 7942, Synpcc7942_0207 and Synpcc7942_1560, with above 70% identities in protein sequence. Another small protein encoded by pANL09 (105 aa) shows high similarity only to pANL26 and pANl27, and to no other proteins in the database. These loci may have arisen from duplication events of ORFs as both direct and inverted repeats are detected in these ORFs (Table 5-2). The pANL10 locus encodes a protein almost identical (95% identities) to the C-terminal domain of the pANL28 protein, a site-

specific recombinase. The largest pair of direct repeats found in pANL is located in these two ORFs (Table 5-2).

The signal transduction region

This region starts with hypothetical protein pANL14. ORFs pANL26 and pANL27, which are closely related to pANL09, are at the other end. The most striking feature in this region is the presence of genes that encode a two-component system. ORF pANL22 and its upstream overlapping pANL23 encode a histidine protein kinase (HPK) and cognate response regulator (RR), respectively. The HPK encoded by pANL22, 439 amino acid residues in length, contains a C-terminal histidine protein kinase domain with the expected conserved histidine residue as the phosphoacceptor site. No identifiable domain is apparent at its N-terminal part except two transmembrane regions. The response regulator protein (225 aa) encoded by pANL23 carries a genuine receiver (REC) domain with the conserved aspartic acid residue and a C-terminal effector domain. Three two-component systems have been identified on the plasmids of *Synechocystis sp.* PCC 6803 (240), and they were shown not to be involved in the perception of hyperosmotic stress (255). The real function of these plasmid-born two-component systems is yet to be explored.

Other proteins encoded in this region may also participate in signal transduction and regulation. The pANL15 locus encodes a protein carrying a predicted transcriptional regulator domain. The protein encoded by ORF pANL16 looks like an NADPH-specific quinone reductase. There is a Type 1 glutamine amidotransferase homologue encoded by

pANL18. Putative gene pANL19 encodes a likely one-component system protein (159), carrying an N-terminal GAF domain and a C-terminal HTH_LUXR domain, which probably binds DNA directly. The ORF pANL25 encodes another protein that may also bind DNA through its C-terminal HTH_DEOR domain.

The plasmid maintenance region

As with many other low-copy number plasmids, pANL carries a full set of genes (Fig. 5-3) necessary for maintaining the stable inheritance of pANL in cells. These genes include: 1) one site specific recombinase for resolution of multimers, pANL28 (*xerC*), 2) a partitioning system for equally distribution of plasmids into daughter cells, pANL31 and pANL32 (*parB* and *parA* homologues, respectively), 3) two toxin-antitoxin cassettes for post-segregational killing of pANL-cured cells: pANL29 and pANL29.5, which encodes a PemK family toxin and its cognate antitoxin, respectively; and pANL33 and pANL34, which was also suggested as a candidate for a VapC-like toxin-antitoxin system (Dr. Roy David Magnuson, personal communications). They are designated *sepA1/T1* (pANL29.5/pANL29, PemK-like system) or *sepA2/T2* (pANL33/pANL34, VapC-like system) for *Synechococcus elongatus* plasmid-born antitoxin or toxin genes (Table 5-1, Fig. 5-3). No significant conserved domains were identified for putative antitoxins SepA1 and SepA2. Their functional assignment is mainly based on their gene context and conservation among other cyanobacteria, as well as deletion analysis (Table 5-3). Both of these ORFs not only are located upstream of, but also overlap with, their cognate toxin genes, as seen in other toxin-antitoxin cassettes. Protein SepA1

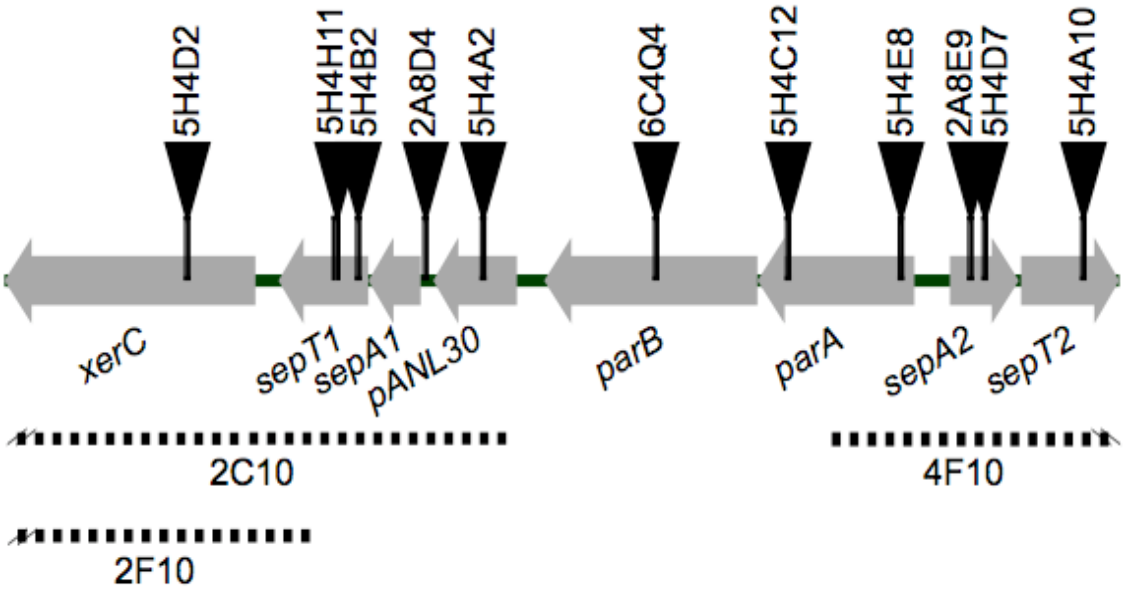


Fig. 5-3. Representation of the plasmid maintenance region showing the relative positions of mini*Mu* insertions. Putative ORFs are shown as grey boxes with arrowheads; black arrows are *Mu* insertions; the dashed lines under ORFs indicate regions that are absent from the cosmids labeled below.

(pANL29.5, 71 aa) shows ~48% identities to N9414_18895, a 79 aa small protein in *Nodularia spumigena* CCY9414. The gene downstream of N9414_18895, N9414_18900, also encodes a PemK-like protein. Thus both SepA1 and N9414_18895 are very likely antitoxins, even though they are both hypothetical proteins. The first 28 aa of SepA1 also shares ~64% identity with that of Ssl2420 in *Synechocystis* sp. PCC 6803, an 84 aa small hypothetical protein. Sll1225, downstream of Ssl2420 and carrying a PIN domain (COG4113), is possibly a toxin protein. It is interesting to see that antitoxin proteins for two different toxin families are very similar in their N-terminal parts. Thus, a composite and modular mechanism, which has been depicted from a toxin-antitoxin system on *Bacillus thuringiensis* plasmid pG11 (172), is also adopted for cyanobacterial toxin-antitoxin systems, in which similar antitoxins can be assembled with different toxins. There is no significant hit in the GenBank database for SepA2 (pANL33), the putative antitoxin for the SepT2 (pANL34) protein. The closest homologues of pANL34, the putative VapC family toxin, are RS9917_04395 from *Synechococcus* sp. RS9917 (37% identity) and SYNW0214 from *Synechococcus* sp. WH 8102 (36% identity). Upstream of each of these two toxins is a small ribbon-helix-helix protein. However, they are not homologous to SepA2 (pANL33) despite their similar size and gene context. The function of SepA1 or SepA2 as antitoxin was confirmed by deletion analysis (Table 5-3) and transposon-mediated mutagenesis as mentioned below.

The last gene in this region, pANL30, maps between the two putative operons of *pemIK* and *parAB*. The encoded protein contains a nucleotidyl transferase domain

(pfam01909) and may function by adding a nucleotidyl group onto its substrate. The real cellular function of this protein needs to be determined.

The sulfur-regulated region

Adjacent to the plasmid maintenance region is the sulfur metabolism related region, including 13 previously designated *srp* genes (*srpA*→*srpM*). This region can be divided into two gene clusters that are in opposite orientations. Within the clusters, all genes are expressed in the same direction. The counter-clockwise cluster comprises ORFs from pANL35 to pANL40. The clockwise cluster includes ORFs from pANL44 to pANL55. The two clusters are separated by two head-to-head genes: pANL42 and pANL43, both encode conserved hypothetical proteins (Table 5-1 & Fig. 5-2).

In the counter-clockwise cluster, a rhodanese-related sulfurtransferase domain (COG0607) is present in the pANL35 protein sequence. Its downstream gene pANL36 encodes a cysteine desulfurase domain (COG1104). Proteins encoded by pANL40 and pANL39, SrpG and SrpH, correspond to enzymes for catalyzing the two steps in the cysteine biosynthesis pathway, O-acetylserine (thiol) lyase and serine acetyltransferase, respectively (252).

Among the proteins encoded in the clockwise cluster are a putative transcriptional regulator pANL44 (COG1959), a putative carbonic anhydrase pANL45 (cd03379), a 36 kD putative catalase (pANL46/SrpA) that is induced in *S. elongatus* PCC 7942 cells under sulfur stress, and a likely chromate transporter (pANL48/SrpC) (253). ORF pANL47 (SrpB) carries an MgtC domain with no known function

(pfam02308). There is another O-acetylserine (thiol) lyase homologue encoded by pANL49 (SrpD). SrpE (pANL50) is a predicted gamma –Glutamyltranspeptidase (pfam01019), a key enzyme in glutathione metabolism. SrpJ (pANL52) is a homologue of acyl-CoA dehydrogenase (COG1960). The downstream three ORFs that are likely in an operon, *srpKLM*, encode subunits of putative TauABC complex, an ABC-type nitrate/sulfonate/bicarbonate transport system.

It was shown previously that genes from all but one *Bam*HI fragment of pANL display increased transcription under sulfur-deficient conditions (253). The only exception, Ba3, covers part of the replication region. Thus, the majority of pANL is involved in response to sulfur starvation. In the BG-11 medium (223) widely used for culturing cyanobacteria, the main sulfur source $MgSO_4 \cdot 7H_2O$ is in a concentration of 0.3 mM. If converted to the amount of sulfate, it is around 30 mg/L. Even though many rivers and lakes today are enriched by nitrate and sulfate due to human activities, natural fresh water systems usually contain much lower levels of sulfate (256). In contrast, sulfate concentration in average seawater is much higher (~185mg/L). Deficiency of sulfur or other major essential nutrients can result in degradation of the cyanobacterial light-harvesting complex, the phycobilisome (photobleaching), as well as modification of other metabolic pathways and physiological processes (257, 258). Thus, the essential plasmid pANL provides important advantages for the fresh water cyanobacterium *S. elongatus* PCC 7942 because it encodes almost all of the components required for adaptation to sulfur deficiency. These include genes functioning in signal transduction, gene expression regulation, and sulfate metabolism. They probably enhance the ability

of cyanobacterial cells to sense environmental changes in sulfate, elevate the expression levels from genes to increase the import of sulfate, make more cysteine, and speed up other steps in the sulfur metabolism to overcome the sulfur stress. Many of these genes have paralogs on the chromosome. However, the compact and efficient organization of these genes on pANL presumably facilitates cells' reaction in response to sulfur limitation. To preserve these advantages, pANL adopts all available strategies for plasmid maintenance and makes itself hard to cure.

Cosmid-based deletion analysis of pANL

To experimentally define the essential regions on pANL, deletion analysis was performed using pANL-bearing cosmids (Table 5-3). Usually, a cosmid carries a part of pANL sequence randomly produced during *Sau3*AI partial digestion in the process of cosmid library construction (65). If the pANL sequence on a cosmid includes all essential parts for pANL maintenance, the cosmid should be able to completely replace pANL in the cyanobacterial cell under constant selection for antibiotic-resistance markers on the cosmid vector. If not all essential components are present, both pANL and the cosmid would co-exist in the cells even under strong selection, and would probably form recombinant variants of them (251). The composition of the pANL/cosmid pool in the cyanobacterial cells can be detected by PCR amplification using primer pairs specific for the region of pANL that is lost in the cosmid or for the cosmid vector.

With both ends checked by sequencing, 27 cosmids missing different parts of pANL were chosen for deletion analysis (Table 5-3). Transformation and selection of these cosmids into *S. elongatus* cells of wild-type strain AMC06 were done in 24-well microplates. Total DNA was then extracted using a boiling method (see Materials and Methods) directly from *S. elongatus* transformed cells (not plated for single colonies) after ~100 generations of selection with kanamycin (>3 months). PCR amplification using primer pairs specific for the cosmid vector SuperCos I confirmed the presence of cosmids in all transformed strains (Table 5-3). Then for each cosmid-transformed strain, one or more pairs of primer specific for the region shared by pANL and the transformed cosmid as well as one or more pairs of primer specific for the region of pANL absent from the cosmid were used to detect the segregation of pANL in the cells (Fig. 5-4). As summarized in Table 5-3, all cosmid strains preserved the common regions, while some strains lost the region that is absent from the transformed cosmid, which indicates that no endogenous pANL is present in these cells. Other strains, however, still maintained the region that is lost in the cosmid; that is, pANL could not be fully segregated. The combined results of PCR amplification in different cosmids identified two regions that are likely indispensable for pANL: *1)* the region that includes ORFs pANL29.5, pANL30, pANL 31, and pANL32, *2)* the region that includes ORFs pANL55.5, pANL56, and pANL56.5. The putative antitoxin gene *sepA2* (pANL33) was also found to be essential for pANL, which will be discussed below. The first segment corresponds to the plasmid maintenance region, while the second is likely the replication origin. Other sequences, including the signal transduction and *srp* regions, are not

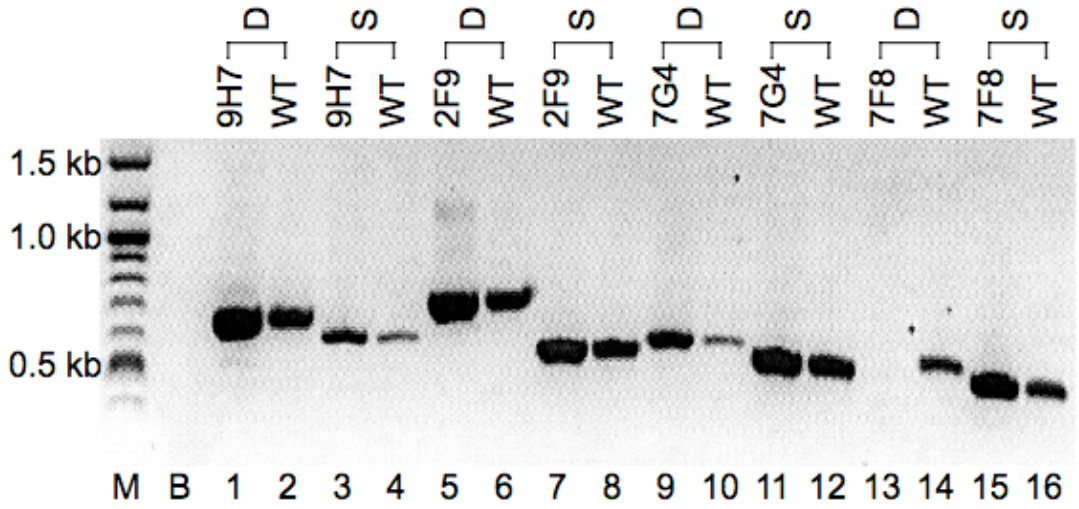


Fig. 5-4. The segregation of pANL in cosmid-transformed strains. Representative PCR amplification from boiled samples is shown. Primer pairs used for strains transformed with different cosmids: 9H7, co21×co22 (D) and gap55×gap53 (S); 2F9, co30×co31 (D) and 2ie3×614ie5 (S); 7G4, gap55×gap53 (D) and 2ie3×614ie5 (S); 7F8, gap55×gap53 (D) and 2ie3×614ie5 (S). M, 100-bp DNA ladder (Invitrogen); B, blank lane; 1-16, boiling samples. WT, wild-type strain AMC06. D, primer pair specific for region deleted in the cosmid; S, primer pair specific for region shared by both pANL and the cosmid.

essential for the maintenance of pANL in the cells. The results suggest that even though these sulfur-regulated genes may be important for the cellular functions of *S. elongatus*, they are not indispensable. One possible reason is functional redundancy, which is supported by the fact that many genes in this region have clear homologues on the chromosome. However, the results suggest that these genes indeed provide significant advantages for the cells in adaptation to sulfur starvation, because all known plasmid maintenance mechanisms are adopted for avoiding the loss of pANL.

Transposon-mediated mutagenesis of plasmid maintenance genes

To further investigate the essentiality of individual genes in the plasmid maintenance region, mini*Mu*Cm transposons inserted into pANL cosmids were used for mutagenesis of these genes. *S. elongatus* PCC 7942 favors double homologous recombination. Once transformed and selected by chloramphenicol (Cm), those transposons will be integrated into their corresponding loci on pANL and disrupt the function of the putative ORFs, if applicable. The cosmid itself will be lost because no selection pressure for the cosmid vector is exerted. If the disrupted pANL gene is not essential, then original pANL with no insertion will be fully segregated eventually. PCR amplification of the insertion locus was used to detect the presence of intact pANL in transposon-inserted strains. Mini*Mu*Cm transposons in pANL cosmids 2A8, 6C4 and 5H4 were introduced into wild-type *S. elongatus* strain AMC06 (Fig. 5-3). Transformation of AMC06 with *Mu*

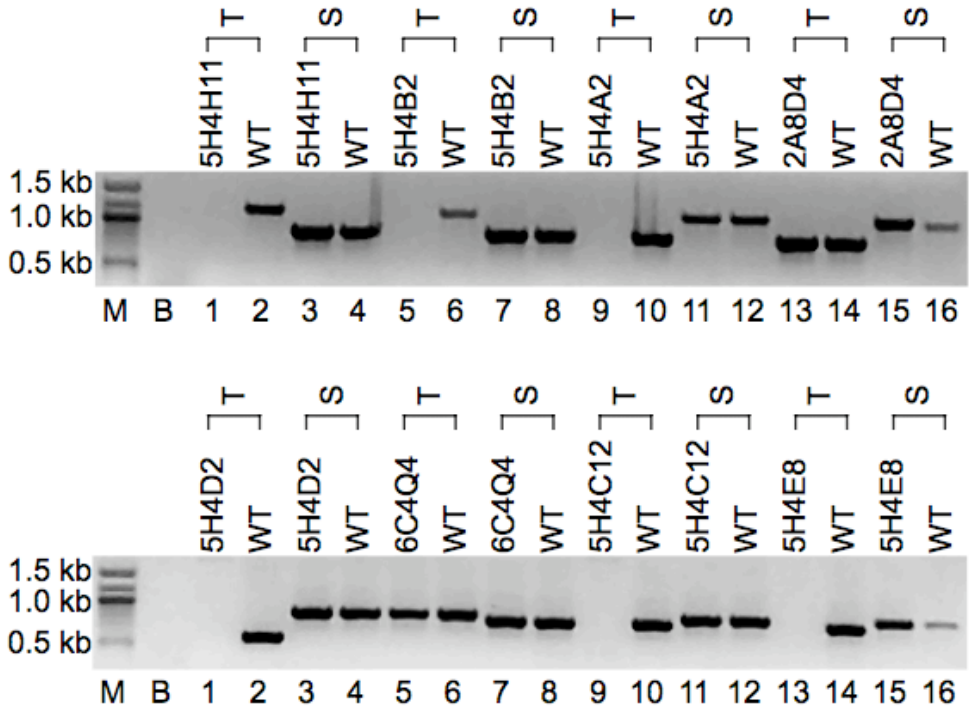


Fig. 5-5. The segregation of pANL in transposon-mediated mutant strains of plasmid maintenance genes. PCR amplification data from boiled samples are shown. Primer pairs used for detecting the insertion of mini*Mu* and wild-type copy of pANL sequence are: 5H4H11, 2ie3×pan10-2 (T) and pan11×pan12 (S); 5H4B2, 2ie3×pan10-2 (T) and pan11×pan12 (S); 5H4A2, pan11×pan12 (T) and 2ie3×pan10-2 (S); 2A8D4, pem5×pemi3 (T) and 2ie3×pan10-2 (S); 5H4D2, 2ie3×614ie5 (T) and pan11×pan12 (S); 6C4Q4, pan11×pan12 (T) and pan13×pan14 (S); 5H4C12, pan13×pan14 (T) and pan11×pan12 (S); 5H4E8, pan13×pan14 (T) and pan11×pan12 (S). In both upper and lower panels, lower labels: M, 100-bp DNA ladder (Invitrogen); B, blank lane; 1-16, boiled samples; upper labels: WT, wild-type strain AMC06; T , primer pair specific for regions surrounding *Mu* insertion; S, primer pair specific for region with no insertion, thus shared by both pANL and the cosmid.

insertions located in *sepA2* (pANL33), encoding antitoxin for the VapC-like toxin, resulted in either no colony or tiny colonies with very slow growth (2A8E9 and 5H4D7, Fig. 5-3), suggesting a likely antitoxin function. Transformants with *Mu*s in other ORFs were normal in growth. Primer pairs specific for the insertion locus were used in PCR amplification of boiled cyanobacterial samples. In the wild-type strain a PCR product, usually smaller than 1 kb as designed, should be produced. In insertional mutant strains, the expected PCR product is more than 2 kb due to the insertion of mini*Mu* (~1.4 kb). For unknown reasons, PCR amplification with boiled samples usually could not produce products bigger than 1.5 kb. Thus, no signal can be seen on the agarose gel if all pANL molecules in the cells carry the transposon insertion. Otherwise, a wild-type band should be detected. As shown in Fig. 5-5, no original pANL could be detected in *Mu* insertional strains carrying: 5H4H11 and 5H4B2 (both in pANL29, a *pemK* homologue), 5H4A2 (pANL30, encoding putative nucleotidyl-transferase), 5H4D2 (pANL28, encoding putative site-specific recombinase), as well as 5H4C12 and 5H4E8 (both in pANL32, a *parA* homologue). Preliminary data showed that pANL was also completely segregated in a strain transformed with 5H4A10 (pANL34, encoding VapC-like toxin) (data not shown). The results suggest that these genes are not essential for maintaining pANL. The *sepA1* (pANL29.5) gene, encoding the antidote for PemK, and pANL31, whose product is a ParB-like nuclease, are essential as indicated by the presence of a PCR product from un-segregated pANL in strains transformed with the respective mini*Mus* 6C4Q4

(pANL31) and 2A8D4 (pANL29.5). Together with the transformation results, bioinformatics shows two toxin-antitoxin cassettes on pANL. The *sepA2* was not identified as essential in cosmid-based deletion analysis with cosmid 2B10 or 4H2, both missing the C-terminal part of *sepA2, sepT2* and some downstream ORFs, which showed complete segregation of pANL (Table 5-3). These results suggest that the N-terminal part of SepA2 is sufficient for inhibiting the activity of SepT2.

Only toxin-antitoxin cassettes are essential

Toxin-antitoxin cassettes prevent curing of plasmids by a post-segregational killing mechanism which is based on the fact that toxins are much more stable than antitoxins. These genes are widely distributed in plasmids and chromosomes of many free-living bacteria with functions including plasmid maintenance, programmed cell death, or stress response (169, 259). Our mutagenesis data show that both antitoxins (SepA1 and SepA2) for PemK-like and VapC-like toxins are essential. To confirm their functions, complementation-mediated pANL segregation assays in strains that carry essential cosmids were performed (Fig. 5-6). Cosmid 2C10 is lacking ORFs 18-30, including the entire *sepA1T1* operon (Fig. 5-3). The endogenous pANL could not be completely segregated out by introduction of this cosmid (Table 5-3, Fig. 5-6A upper panel). When a copy of the *sepA1T1* operon was inserted into NS1 in the chromosome, PCR reactions from all four independent clones were negative for the pANL-specific product that is

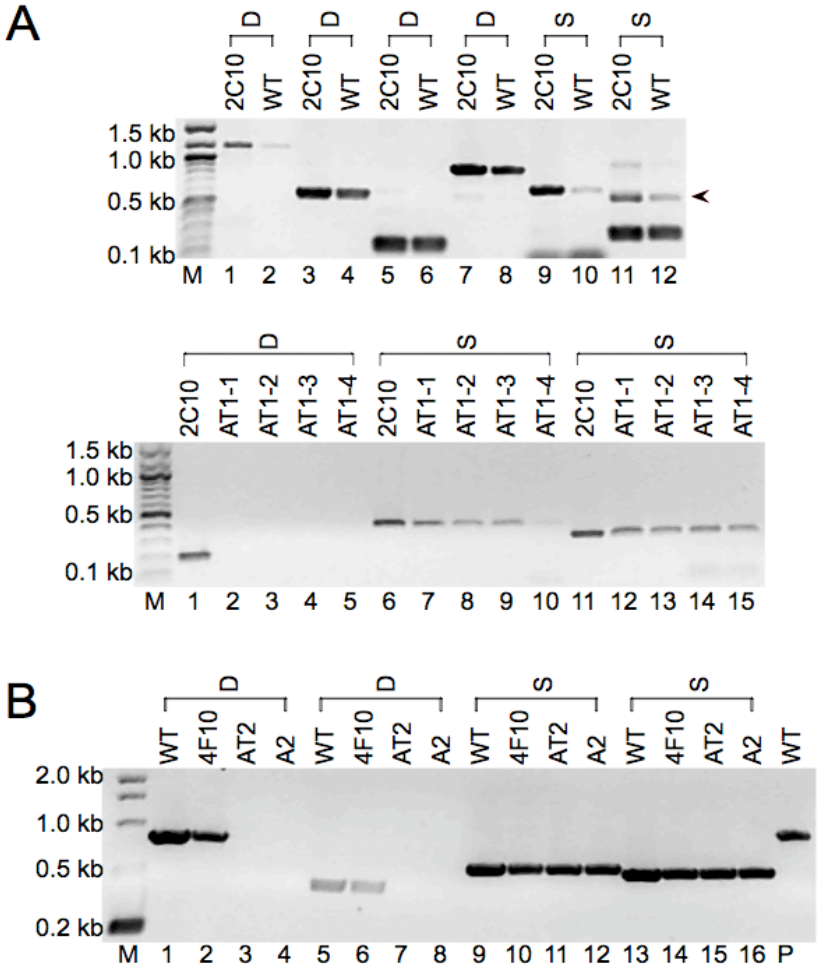


Fig. 5-6. Complementation analysis of toxin-antitoxin cassettes. ***A)*** The complementation with *sepA1T1* operon in NS1 resulted in full segregation of pANL in 2C10-transformed strain (AMC1584). Upper panel, segregation of pANL in AMC1584 is not complete; Lower panel, pANL was completely lost in all four independent *sepA1T1*-complemented 2C10 strains. Primer pairs used, in left-to-right order, for upper panel: pan5×pan6 (D), 2ie3×614ie5 (D), 8e3×5e5 (D), pan11×pan12 (D), gap55×gap53 (S), and gap45×gap43 (S); for lower panel: 8e3×5e5 (D), gap15×gap13 (S), and co4×co8 (S). 2C10, 2C10-transformed strain; WT, wildtype AMC06; AT1, *sepA1T1*-complemented AMC1584 strains. D, primer pair specific for deleted region in the cosmid; S, primer pair specific for region shared by both pANL and the cosmid; M, 100-bp DNA ladder (Invitrogen); 1-12 (upper panel) and 1-15 (lower panel), boiled samples. ***B)*** Segregation of pANL is complete in 4F10-transformed strains (AMC1505) when a copy of *sepA2T2* or *sepA2* was inserted into NS1. Primer pairs used, in left-to-right order: 614ie3×gap43 (D), gap45×1ie5 (D), 2ie3×614ie5 (S), and gap35×10ie5 (S). 4F10, 4F10-transformed strain; WT, wildtype AMC06; AT2, *sepA2T2*-complemented AMC1505 strain; A2, *sepA2*-complemented AMC1505 strain. D, primer pair specific for deleted region in the cosmid; S, primer pair specific for region shared by both pANL and the cosmid; M, 1kb DNA ladder (Invitrogen) enhanced with a 0.2-kb PCR products; 1-16, boiled samples. P, positive control (CTAB-minipreped total DNA sample amplified with 614ie3×gap43).

absent from the cosmid, indicating loss of pANL in the strain that had the cassette encoded on the chromosome (Fig. 5-6A lower panel). Complementation by the antitoxin gene *sepA1* alone did not succeed for unknown reasons (data not shown). Another strain carrying cosmid 4F10, which is missing the *sepA2T2* operon and hence also could not compete out pANL (Fig. 5-3 & Table 5-3), was complemented by inserting either *sepA2T2* operon or just *sepA2* into NS1 (Fig. 5-6B). These results confirmed that both toxin-antitoxin cassettes are active and indispensable. Results suggest that nothing else is required to account for the essentiality of pANL, which is contradictory to the previous mutagenesis analysis data that suggested *parB* (pANL31) is essential. Because *parB* is not adjacent to the *sepA1T1* operon and insertion in the upstream pANL30 did not show any effect on pANL segregation (Figs. 5-3 & 5-5), polar effects on *parB* are unlikely to be responsible. Further analysis need to be performed to figure out the reasons.

Attempts to cure pANL *via* counter-selection

With the knowledge of essential regions of pANL, attempts to cure pANL were conducted with a *sacB*-based counter-selection strategy. Selection with *sacB*, which encodes the secretory levansucrase from *Bacillus subtilis* (260), has been used successfully in cyanobacteria (52, 261). First, a copy of *sacB* was inserted onto cosmid 2F10, in which the toxin gene in operon *sepA1T1* is partially deleted and the antitoxin gene is intact (Fig. 5-3). The segregation of pANL was complete in strains transformed

with 2F10 (Table 5-3). Then the *sacB*-2F10 construct was introduced into a strain that carries a copy of the *sepA2T2* operon in NS1. In this strain the requirement for both toxin-antitoxin cassettes should be abolished, as was demonstrated individually for each, and *sacB*-2F10 was expected to be dispensable to the cells. However, spontaneous loss of the construct did not occur after culturing the cells in media with no selective antibiotics, probably due to the presence of other plasmid maintenance components, such as partitioning system (data not shown). Next, cells were transferred to medium containing 5% sucrose. Only cells that lost the *sacB*-2F10 construct should be able to survive. However, due to technical problems, no resistant colonies have been recovered on sucrose-containing medium (data not shown). Trouble-shooting analysis is under way and other counter-selection markers, such as *glcP*, are being tested.

Circadian clock phenotype screening of cosmid-transformed strains

To find out whether pANL genes are involved in regulation of circadian clocks in *S. elongatus* PCC 7942, cosmids carrying different fragments of pANL sequence, such that some pANL genes are missing, were introduced into a bioluminescent reporter strain, AMC1020. Transformed mutant strains were then screened on a TopCount luminometer for circadian phenotypes. All mutant strains displayed wild-type circadian rhythms under constant conditions (data not shown). Thus, no fragment of pANL was identified as affecting activity of the circadian clock in *S. elongatus*.

Conclusions

The large endogenous plasmid of *S. elongatus* PCC 7942, pANL, is ~46 kb in length and encodes 58 putative ORFs. Five pairs of relative large repetitive sequences are present in pANL. It can be divided into four structural and functional regions: the replication region, the signal transduction region, the plasmid maintenance region, and the sulfur-regulated region. Only one replication origin was identified based on deletion analysis. In addition to the replication origin, the partitioning system, and the site-specific recombinase, two functional toxin-antitoxin cassettes are also involved in maintenance of pANL in the cells. Genes in the signal transduction region, together with the sulfur-regulated genes, may also participate in the adaptation of *S. elongatus* cells to sulfur starvation. Attempts to cure pANL of *S. elongatus* cells were not successful yet. It is not likely that pANL genes are involved in circadian clock functions in *S. elongatus* PCC 7942.

Materials and Methods

Cyanobacteria strains, media, and culture conditions

All cyanobacterial wild-type reporter and mutant strains were created in *Synechococcus elongatus* PCC 7942. Cyanobacterial strains were grown in BG-11 medium (223) under continuous light conditions (~70 μE/m^2s) at 30°C with appropriate antibiotics. The bioluminescent reporter strain AMC1020 carries the *psbAI* promoter driving *Vibrio harvey* luciferase (*luxAB*) in the neutral site I (NS1) locus and the aldehyde substrate synthesis genes *luxCDE* in NS2 (52). Neutral sites mediate homologous recombination with *S. elongatus* chromosome and cause no apparent phenotypes (52). The cyanobacterial strains used in this study are summarized in Table 5-4.

Plasmid construction

Unless otherwise stated, plasmids were constructed in *Escherichia coli* strain DH10B (Invitrogen, Carlsbad, CA). All plasmids are described in Table 5-4. pAM3558 is a gentamycin-resistance version of pAM1303, which is a NS1 cloning vector. The *aadA* gene (Sp^RSm^R) in the middle of the *Ω* cassette of pAM1303 was replaced by a copy of the *accC1* gene (Gm^R) from pAM3511 (Table 5-4). The *sepA1T1* operon or *sepA1* alone was amplified using primer pairs pem5×pemk3 or pem5×pemi3, respectively. The PCR products were then inserted into pAM1303 to form pAM3716 and pAM3715, respectively. PCR products of the *sepA2T2* operon or *sepA2* alone, which were amplified with primer pairs add2-5×add2-34 or add2-5×add2-33, respectively, were inserted into pAM3558 to construct pAM3617 and pAM3618, respectively.

Table 5-4. Cyanobacterial strains and plasmids used in Chapter V.

Strains	Genetic background	Plasmid/cosmid transformed	Antibiotic resistance*	Source or reference
AMC06	Wild-type	None	None	Lab collection
AMC1020	Wild-type	pAM1501 (NS1) pAM1619 (NS2)	Sp^RKm^R	Andersson et al. (2000)
AMC1502	Wild-type	pAM3617	Gm^R	This study
AMC1503	Wild-type	pAM3618	Gm^R	This study
AMC1505	Wild-type	4F10	Km^R	This study
AMC1529	Wild-type	pAM3715	Sp^RSm^R	This study
AMC1530	Wild-type	pAM3716	Sp^RSm^R	This study
AMC1584	Wild-type	2C10	Km^R	This study
AMC1585	AMC1584	pAM3716	$Sp^RSm^RKm^R$	This study
AMC1586	AMC1505	pAM3617	Gm^RKm^R	This study
AMC1587	AMC1505	pAM3618	Gm^RKm^R	This study

Plasmid	Characteristics	Antibiotic resistance*	Source or reference
pAM1303	NS1 cloning vector	Sp^RSm^R	Andersson et al. (2000)
pAM3511	Source of Gm-resistance gene (*accC1*)	Gm^RKm^R	This study
pAM3558	Gm resistant version of pAM1303	Gm^R	This study
pAM3617	*sepA2T2* in pAM3558	Gm^R	This study
pAM3618	*sepA2* in pAM3558	Gm^R	This study
pAM3715	*sepA1* in pAM1303	Sp^RSm^R	This study
pAM3716	*sepA1T1* in pAM1303	Sp^RSm^R	This study

*Resistance to: Cm^R, chloramphenicol; Km^R, kanamycin; Sp^R, spectinomycin; Sm^R, streptomycin; Gm^R, gentamycin; Ap^R, ampicillin

Hybridization analysis

Cosmid 6G1 DNA (~200 ng/μl) was digested with *Eco*RI. Two large digestion fragments (~12 kb and ~16 kb), gel-purified using a DNA purification kit (BIO-RAD, Hercules, CA), were labeled with ^{32}P-dCTP as hybridization probes to blot 10 pieces of nitrocellulose membranes (Dr. Colleen Thomas), each blotted with DNAs from one of the ten 96-well plates of the cosmid library. Membranes were incubated first in prehybridization solution (5× SSPE, 1%SDS, 0.1 mg/ml SSDNA) at 65°C for 4 h, and then in hybridization solution (5× SSPE, 1%SDS) at 60°C overnight after addition of

boiled probes. Next, membranes were rinsed (0.5× SSPE, 0.2% SDS) 3 times for 3 min each at 60°C, and then washed in the same solution 3 times for 30 min each at 62-65°C. Finally, air-dried membranes were wrapped and exposed to x-ray film at -80°C overnight.

Sequence determination and annotation

The strategy for determination of the complete nucleotide sequence of pANL is described in the Results and Discussion. *In vitro* mutagenesis of cosmid 2A8, screening of the mutant pool, and sequencing of cosmid ends and flanking regions of *Mu* insertions were performed as previously described (65). Primers used in primer walking assays are underlined in Table 5-5.

The individual sequences were arranged and assembled using Vector NTI software package (Invitrogen, Carlsbad, CA). Protein-coding regions were predicted with FramePlot (195). Gene functional assignment was conducted by similarity searching in protein database and conserved domain database of GenBank (196, 262).

Boiling method for preparation of cyanobacterial PCR templates

Samples (1-2 ml) of *S. elongatus* PCC 7942 cells (OD 0.6-0.9 at 750 nm) were collected by centrifugation at 16,000 x g for 1 min and then resuspended in 150-200 μl TE (pH 8.0). The resuspended cells were boiled for 5 min, pelleted at 16,000 x g for 15 min. The supernatant fractions were then transferred to a new 1.5 ml tube. For a 50 μl PCR reaction, 2-5 μl boiled samples (~5-10 ng/μl) were used.

Table 5-5. Primers used in this study.

Primer	Sequence (5'→3')	Primer	Sequence (5'→3')
pan1	gcgacggcagatttcagacc	pan14	tgattggcggtgttctggtg
pan2	gcaatgcggaccgagcaccc	614ie3	aggctccccgtttgaatcc
Co1	gatccagtcactcacgaggtg	gap43	cagctcagtgctgagacagg
2H8-B	ggaggtatcgccatcgaagc	1ie5	gtgcggcatttccttaatgc
pan3	gatcaaaggagcacgacggg	gap45	tggacatactccagcgctg
pan4-2	cggacttgctgcgtggctgc	co28	ccggtactgaacgaggaactgaggc
2H8-A	ctctcaaagccagcaaagcc	co29	cgccagtggaagcgggtatctaccg
1ie3	cctgggaattcctcgtctgg	co24	gtcttcaattactggatggcgagcc
gap53	agctattgctggaggccatg	co23	gctcaagaggaaggactgatttcgg
gap55	cgctactaccagctcgacatg	co27	cgttaaatgtcggattggtgcagcc
7ie3	cttacaattaaagtaattcctctcacactcc	co2	gtctgctcatgaatcgctgg
7ie5	agaggcgcattgtgagattg	co26	ttgtagagcctgtgctgatgcctgc
gap13	ttgatatgaagacgccaccg	co25	actcgtctccgctctttgctatggc
3e3	cccaagacttcctcctttcacc	co6	cgggaccaagcagaaggactcagcg
gap15	tgtccgcaacattgagcg	co3	caagggcgatcgctaacc
gap23	tggcattcgttgaactgaag	co7	tcgcttcttgaacatggcactgctc
gap25	tgaatctggcctatcgaacg	co8	gagacggaaggagagacctagccag
411e5	gaaccttgacgtgaccacgt	co4	ctttgaccttgacggctgc
411e3	gcatgacgcttgaagaaatgc	co5	gctgtgcagtttcttggagg
gap33	gggctgatcaatggtttgg	co13	ctgaagcgatgctgggctgtattgc
10ie5	ccaagccacacctgaacaga	co14	ctgggccagttctggattgcgttgc
gap35	tgtctctctctgcaaagacagc	co19	ccgctgtatctcgttgaggcaaccc
8e5	tcggccttggtagcaaaac	co20	acgagattggaggcccagggcaagg
10ie3	gccactggccgaaaagc	co21	tcgccatccccatctgctgttgctg
5e5	ggtggaatcggacaaatcg	co22	gtagcgcagcgaaatggtagcaggc
8e3	acttcaacgcccagcatg	co30	gctctgcgtcggcttccttgatggc
pan5	gcaactgccctgtccaatcg	co31	cagatatgggtaatgctcgtagtgaggc
pan6	gactgctcttgctctgtgcg	pem5	cactgagaacgagaagcggctcgac
pan7	agcaagcgcacttgagcagc	pemi3	gacgtagcttcgttcgactgaggtc
pan8	ggtaacggcgaaagaccctc	pemk3	tctgaggcaggcttaaccagtaggc
pan9	cctcaccccagtagcttggc	add2-5	ctgcctgccatggaaagtgcgtc
614ie5	ggtaaggcattgttttaaccacc	add2-33	tcaggaagacagcctttgaggtgg
2ie3	cagagcagccgcagaatc	add2-34	gtgctactggatgtcagagagtcg
pan10-2	gcctcttgtacctgatttggagctg	sci-1	cgaccgatgcccttgagagc
pan11	ccgcttctcaatcaagtccg	sci-2	gcttccattcaggtcgaggtgg
pan12	cacccgcgtcagtttgaacg	sci-3	caacctatggaactgatgaatgggagc
pan13	gcctctcgctgtccgcctgc	sci-4	cagtcgcttcacgttcgctcg

* Underlined primers were used for primer-walking.

Cyanobacterial transformation and bioluminescence assay

S. elongatus PCC 7942 cells were grown to OD 0.6-0.9 (750 nm). Transformation assays were carried out as described previously (52). For transformation in 24-well plates, a 200 μl aliquot of washed and concentrated culture was distributed into each well and a DNA sample (usually 0.5-1.0 μg) was added. After mixing each well by pipetting, the plate was wrapped with aluminum foil and incubated stationary at 30°C for 12-16 h. Then 1-2 ml BG-11 medium (223) with appropriate antibiotics was added into each well. After 5-7 days of selective incubation at 30°C in constant light, fresh green transformant dots were resuspended and allowed to grow up for 2-3 more days. Transformed strains were routinely sub-cultured into new 24-well plates with fresh selection medium. The measurement of bioluminescence from the reporter strains was performed as described previously (65, 79). The acquired bioluminescence data were processed and graphed by the Import and Analysis (I and A) Excel macro set (S. A. Kay Laboratory, The Scripps Research Institute, La Jolla, CA).

CHAPTER VI

DISCUSSION AND CONCLUSION

Discussion

The Power of Comparative Genomics

Whole genome sequence comparison not only discovers gross chromosomal differences, like genome organization and composition, but also reveals nucleotide level polymorphisms, such as insertions and deletions in specific genes. In addition, it can also identify commonly shared regions, which enhances gene function assignment. Most important of all, multi-species sequence comparisons can elucidate genome loci that are involved in differences among related species in physiology, morphology, and geology, as well as history of evolution and speciation (263, 264).

The availability of the complete genome sequences of *S. elongatus* PCC 7942 and PCC 6301 (7), as well as dozens of other cyanobacterial species (Table 2-1), provides an excellent opportunity for comparative genomic analysis. Extensive comparative genomics studies in cyanobacteria have been conducted aiming at identifying a core gene set or signature genes of photosynthesis or cyanobacteria, to reveal the origin of photosynthesis or cyanobacteria, and to understand niche adaptation and genome evolution (265-273). Phylogenetic analysis, however, showed that *S. elongatus* strains are not clustered closely with other *Synechococcus* species; instead, this species is closer to *Microcystis* strains, whose genomes have not been sequenced yet (274, 275).

Nevertheless, genome comparisons can still be productive among *S. elongatus* and other *Synechococcus* species and remotely related *Synechocystis* for answering various questions. As an obligate photoautotroph, *S. elongatus* PCC 7942 displays many differences from the facultative photoheterotrophic *Synechocystis* PCC 6803 in gene regulation, such as the perception and response to light (276, 277). Global comparative analysis of these two genomes will give clues on the fundamental genetic and metabolic differences between photoautotrophy and photoheterotrophy in cyanobacteria. As a freshwater cyanobacterium, *S. elongatus*, together with *Synechococcus* CC9311 (273) and WH8102 (269), are good candidates for studying the physiological and ecological adaptation of unicellular cyanobacteria to freshwater, coastal, or marine territories. As a mesophile with optimal growth temperature at ~30°C, *S. elongatus* should have distinct metabolic processes and gene content from thermophilic cyanobacteria *Thermosynechococcus elongatus* (168) and hot spring *Synechococcus* strains (145). More fundamental questions can be asked because cyanobacteria are a diverse group, regarding such properties as non-motile *vs*. motile, unicellular *vs.* filamentous and diazotroph *vs.* non-diazotroph.

Freely available genome alignment tools, *e.g.*, MUMmer, from the J. Craig Venter Institute (previously TIGR) are efficient in identifying shared genes among genomes and unique gene sets for each genome. Scrutinizing regional changes, such as point mutations, sometimes also provides insights into cellular functional differences among closely related species or strains. As described in Chapter III, a single nonsense nucleotide substitution in the pilus assembly gene *pilN* is likely to be involved in the

difference in natural competence between *S. elongatus* PCC 7942 and PCC 6301. These two *S. elongatus* strains share 99.93% identity in their nucleotide sequences. The majority of these sequence differences are single nucleotide changes, which can result from point mutations that create single nucleotide polymorphisms (SNP), or may reflect merely sequencing errors. Analysis of several GAF domain proteins in *S. elongatus* genomes has shown that there are many sequencing errors in the PCC 6301 genome sequence (data not shown). There are likely to be some sequencing errors in PCC 7942 genome, too. Thus, any possible point mutation detected between these two *S. elongatus* genomes, which causes frame-shift or premature stop codons in ORFs, should be verified by sequencing the same locus in both strains.

Evolution and Divergence of *kai* Genes

The origin and evolution of circadian clock genes in prokaryotes have been analyzed by others (106, 278, 279). The *kaiC* gene predicted to be the oldest member, originated in the ancestor of both *Archaea* and eubacteria as a single-domain gene, then duplicated and fused into a double-domain gene, which was subsequently inherited in cyanobacteria. The *kaiB* gene likely originated later in cyanobacteria and fused with *kaiC* to form an operon in which they evolve together. Finally, *kaiA* appeared in cyanobacteria and clustered with *kaiBC* (106). As a whole unit, the *kaiABC* cluster in *Nostoc linckia* was reported to undergo rapid evolution, mainly gene duplication and diversification, under local environmental stress (280).

Extra copies of *kaiBC* or solitary *kaiB* and *kaiC* are present in four of the sequenced cyanobacterial genomes, as described in Chapter II; these orthologs probably resulted from duplication of the gene or whole genome, or lateral transfer events, or both (106, 281). In all these genomes, only one of the *kaiBC* operons is grouped with *kaiA*, which suggests the *kaiABC* cluster functions as its counterpart in *S. elongatus* PCC 7942. The functions of other copies of *kaiB* and *kaiC* may have two major possibilities: ***1)*** they have no function or functions other than in the circadian clock; ***2)*** they function in alternative clock oscillators with input pathways that do not require *kaiA*.

Phylogenetic analysis predicted that extra copies of KaiB and KaiC in *Synechocystis* PCC 6803 are not likely to have circadian clock function because they clustered phylogenetically with KaiB and KaiC proteins from other prokaryotes, such as *Rhodobacter*, in which many amino acid residues critical for clock functions have been replaced with other residues with different chemical polarity (279). However, rhythmic gene expression with circadian or untradian period has been demonstrated in purple bacteria (282). Thus, these proteins may still have clock function when experimentally tested. They may function independently or cooperatively, and may respond to different environmental or intracellular signals. Disruption of the *kaiABC* cluster in *Synechocystis* PCC 6803, which possesses another *kaiBC* operon and one pair of solitary *kaiB* and *kaiC*, will be able to test whether these extra copies of *kaiB* and *kaiC* can complement clock function, at least partially, under different conditions. Alternatively, those extra copies of *kaiB* and *kaiC* can be knocked out one by one or together to see whether the *kaiABC* locus can still function normally.

An "Essential" Circadian Clock

Despite tremendous work and great progress in identifying clock-related loci and elucidating molecular mechanism of central oscillator in cyanobacteria (61, 64, 65, 72, 76, 77, 83, 85, 101, 104), there is little known about how the circadian clock connects to cellular physiology. The overall goal of our functional genomics project is to identify all of the loci that participate in circadian clock function.

During the process of screening transposon-mediated mutants, many of the mutants that have circadian clock phenotypes were recovered in so-called merodiploid cells, in which the mutant alleles were not fully segregated from wild-type alleles. Even though one cell usually contains multiple copies of the chromosome (144, 283, 284), a non-essential mutant allele is usually completely segregated within the first streak of the original cyanobacterial transformants on selective agar plates. Thus, it is very likely that essential loci were targeted in those merodiploid cells and single recombination events occurred to preserve a wild-type allele (8, 201, 285). Presence of the vector sequence, which resulted from the single crossovers, was detected in 11 of the first 71 mutants (~15%), including the *clpP2X* operon (65).

Previous *in vivo* mutants hunts may have identified almost all non-essential clock loci. However, the clock has been shown to be involved in metabolic processes in cyanobacteria, such as photosynthesis (30, 35, 49, 73), nitrogen fixation (21, 33, 37), cell division (34, 72, 134, 286), amino acid uptake (287), and protein turnover (65). Moreover, the circadian clock globally controls gene expression and chromosome compaction (84, 85). All of these data suggest links between circadian clock and

fundamental cell functions. Our comprehensive *in vitro* mutagenesis strategy, which screens each locus in the genome individually, allows us to identify these missing links. Of course, not every mutant allele that disrupts essential locus can give transformants, and not every merodiploid can show a mutant phenotype. But the identification of 11 essential clock loci, in the first run of high throughput screening of just a quarter of the whole genome, is a very promising start to assemble a systems-level view of circadian clock in cyanobacteria.

In addition to our functional genomics project, recent studies in identification of proteins that interact with the clock input pathway component CikA also revealed novel clock loci that are essential for cell viability. Mutants of *cikA*, which is not essential for viability, display a cell-division defect (72) and CikA is closely connected to photosynthetic electron transport (73). The partners of CikA that transmit input signals to central oscillator are still missing. Some of the CikA-interacting proteins identified through a yeast two-hybrid assay are essential for viability, and thus provide new links among the circadian clock, metabolism, and cell division (S.R. Mackey et al, PNAS, in press).

A conditional antisense RNA-based gene suppression method for further analysis of essential genes has been successfully applied in studying *clpP2X* operon (65). The expression of gene-specific antisense RNA is under the control of *lacI/lacO* system, which can be induced by IPTG. This approach enables us to manipulate any essential locus that functions in any fundamental cellular processes.

Conclusion

This functional genomics project has already yielded valuable information, and promises much more to be learned upon its completion. The project has also generated an exceptional resource for the greater community, as the tools created here can be applied to the study of other cellular processes as well.

The complete genome sequence was determined through a combination of shotgun sequencing by the DOE Joint Genome Institute (JGI) and my assembly of transposon-mediated sequencing; the latter sequencing was performed by Dr. C.K. Holtman and her team, who also conducted the majority of mutant screening and clock phenotype analysis. Dr. P.A. Youderian participated in assembling and annotation of first several cosmids deposited into GenBank. Much of the content of Chapter III has been published (65). Immunoblot assays for detecting level of Kai proteins in Figs. 4-1, 4-2, and 4-3 were done by Dr. S.R. Mackey. Y.-I. Kim performed fluorescence anisotropy experiments for detecting interaction between a KaiC peptide and KaiA variants (Fig. 4-6). Dr. R.A. Magnuson (University of Alabama in Huntsville) assisted in identification of toxin-antitoxin cassettes described in Chapter V.

REFERENCES

1. Herdman, M., Castenholz, R. W., Iteman, I., Waterbuty, J. B., & Rippka, R. (2001) in *Bergey's Manual of Systematic Bacteriology*, eds. Boone, D. R., Castenholz, R. W., & Garrity, G. M. (Springler-Verlag, New York). Volume 1, 493-513.

2. Yamanaka, G., Lundell, D. J., & Glazer, A. N. (1982) *J. Biol. Chem.* **257,** 4077-4086.

3. Miki, K., Tamada, T., Nishida, H., Inaka, K., Yasui, A., de Ruiter, P. E., & Eker, A. P. (1993) *J. Mol. Biol.* **233,** 167-169.

4. Shestakov, S. V. & Khyen, N. T. (1970) *Mol. Gen. Genet.* **107,** 372-375.

5. Wilmotte AMR, S. W. (1984) *J. Gen. Microbiol.* **130,** 2737-2740.

6. Golden, S. S., Nalty, M. S., & Cho, D. S. (1989) *J. Bacteriol.* **171,** 24-29.

7. Sugita, C., Ogata, K., Shikata, M., Jikuya, H., Takano, J., Furumichi, M., Kanehisa, M., Omata, T., Sugiura, M., & Sugita, M. (2007) *Photosynthesis Research*. [Epub ahead of print]

8. Golden, S. S., Brusslan, J., & Haselkorn, R. (1987) *Methods Enzymol.* **153,** 215-231.

9. Capuano, V., Thomas, J. C., Tandeau de Marsac, N., & Houmard, J. (1993) *J. Biol. Chem.* **268,** 8277-8283.

10. Sippola, K., Kanervo, E., Murata, N., & Aro, E. M. (1998) *Eur. J. Biochem.* **251,** 641-648.

11. Schaefer, M. R. & Golden, S. S. (1989) *J. Biol. Chem.* **264,** 7412-7417.

12. Spiegel, S. & Bader, K. P. (2003) *Z. Naturforsch. [C]* **58,** 93-102.

13. Soitamo, A. J., Zhou, G., Clarke, A. K., Oquist, G., Aro, E. M., & Gustafsson, P. (1994) *Plant Mol. Biol.* **26,** 709-721.

14. Fadi Aldehni, M., Sauer, J., Spielhaupter, C., Schmid, R., & Forchhammer, K. (2003) *J. Bacteriol.* **185,** 2582-2591.

15. Hirani, T. A., Suzuki, I., Murata, N., Hayashi, H., & Eaton-Rye, J. J. (2001) *Plant Mol. Biol.* **45,** 133-144.

16. Sugimoto, Y., Tanaka, K., Masuda, S., & Takahashi, H. (1997) *J. Gen. Appl. Microbiol.* **43,** 17-21.

17. Bovy, A., de Kruif, J., de Vrieze, G., Borrias, M., & Weisbeek, P. (1993) *Plant Mol. Biol.* **22,** 1047-1065.

18. van Waasbergen, L. G., Dolganov, N., & Grossman, A. R. (2002) *J. Bacteriol.* **184,** 2481-2490.

19. Collier, J. L. & Grossman, A. R. (1994) *EMBO J.* **13,** 1039-1047.

20. Vazquez-Bermudez, M. F., Herrero, A., & Flores, E. (2003) *FEMS Microbiol. Lett.* **221,** 155-159.

21. Bradley, R. L. & Reddy, K. J. (1997) *J. Bacteriol.* **179,** 4407-4410.

22. Durham, K. A., Porta, D., McKay, R. M., & Bullerjahn, G. S. (2003) *Arch. Microbiol.* **179,** 131-134.

23. Garcia-Dominguez, M. & Florencio, F. J. (1997) *Plant Mol. Biol.* **35,** 723-734.

24. Richaud, C., Zabulon, G., Joder, A., & Thomas, J. C. (2001) *J. Bacteriol.* **183,** 2989-2994.

25. Green, L. S., Laudenbach, D. E., & Grossman, A. R. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 1949-1953.

26. Golden, S. S. (2003) *Curr. Opin. Microbiol.* **6,** 535-540.

27. Dunlap, J. C., Loros, J. J., & DeCoursey, P. J. eds. (2003) *Chronobiology: Biological Timekeeping* (Sinauer Associates, Inc., Sunderland. MA).

28. Bell-Pedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E., Thomas, T. L., & Zoran, M. J. (2005) *Nat. Rev. Genet.* **6,** 544-556.

29. Young, M. W. & Kay, S. A. (2001) *Nat. Rev. Genet.* **2,** 702-715.

30. Kondo, T., Strayer, C. A., Kulkarni, R. D., Taylor, W., Ishiura, M., Golden, S. S., & Johnson, C. H. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5672-5676.

31. Stal LJ, K. W. (1985) *Arch. Microbiol.* **143,** 67-71.

32. Mitsui, A., Kumazawa, S., Takahashi, A., Ikemoto, H., Cao, S., & Arai, T. (1986) *Nature* **323,** 720-722.

33. Huang T-C, a. C. T.-J. (1986) *FEMS Microbiol. Lett.* **36,** 109-110.

34. Sweeney BM, a. B. M. (1989) *J. Phycol.* **25,** 183-186.

35. Schneegurt, M. A., Sherman, D. M., Nayar, S., & Sherman, L. A. (1994) *J. Bacteriol.* **176,** 1586-1597.

36. Aoki, S., Kondo, T., & Ishiura, M. (1995) *J. Bacteriol.* **177,** 5606-5611.

37. Chen, Y. B., Dominic, B., Mellon, M. T., & Zehr, J. P. (1998) *J. Bacteriol.* **180,** 3598-3605.

38. Berman-Frank, I., Lundgren, P., Chen, Y. B., Kupper, H., Kolber, Z., Bergman, B., & Falkowski, P. (2001) *Science* **294,** 1534-1537.

39. Chen, Y. B., Dominic, B., Zani, S., Mellon, M. T., & Zehr, J. P. (1999) *Plant Mol. Biol.* **41,** 89-104.

40. Lorne, J., Scheffer, J., Lee, A., Painter, M., & Miao, V. P. (2000) *FEMS Microbiol. Lett.* **189,** 129-133.

41. Ouyang, Y., Andersson, C. R., Kondo, T., Golden, S. S., & Johnson, C. H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 8660-8664.

42. Johnson, C. H., Golden, S. S., & Kondo, T. (1998) *Trends Microbiol.* **6,** 407-410.

43. Woelfle, M. A., Ouyang, Y., Phanvijhitsiri, K., & Johnson, C. H. (2004) *Curr. Biol.* **14,** 1481-1486.

44. Dodd, A. N., Salathia, N., Hall, A., Kevei, E., Toth, R., Nagy, F., Hibberd, J. M., Millar, A. J., & Webb, A. A. (2005) *Science* **309,** 630-633.

45. Michael, T. P., Salome, P. A., Yu, H. J., Spencer, T. R., Sharp, E. L., McPeek, M. A., Alonso, J. M., Ecker, J. R., & McClung, C. R. (2003) *Science* **302,** 1049-1053.

46. Beaver, L. M., Gvakharia, B. O., Vollintine, T. S., Hege, D. M., Stanewsky, R., & Giebultowicz, J. M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2134-2139.

47. Mori, T. & Johnson, C. H. (2001) *Semin. Cell Dev. Biol.* **12,** 271-278.

48. Golden, S. S. & Canales, S. R. (2003) *Nat. Rev. Microbiol.* **1,** 191-199.

49. Kondo, T., Tsinoremas, N. F., Golden, S. S., Johnson, C. H., Kutsuna, S., & Ishiura, M. (1994) *Science* **266,** 1233-1236.

50. Kondo, T. & Ishiura, M. (1994) *J. Bacteriol.* **176,** 1881-1885.

51. Liu, Y., Golden, S. S., Kondo, T., Ishiura, M., & Johnson, C. H. (1995) *J. Bacteriol.* **177,** 2080-2086.

52. Andersson, C. R., Tsinoremas, N. F., Shelton, J., Lebedeva, N. V., Yarrow, J., Min, H., & Golden, S. S. (2000) *Methods Enzymol.* **305,** 527-542.

53. Ikeuchi, M. (1996) *Tanpakushitsu Kakusan Koso* **41,** 2579-2583.

54. Aoki, S., Kondo, T., & Ishiura, M. (2002) *J. Microbiol. Methods* **49,** 265-274.

55. Aoki, S., Kondo, T., Wada, H., & Ishiura, M. (1997) *J. Bacteriol.* **179,** 5751-5755.

56. Johnson, C. H., Hastings, J. W. (1986) *Am. Sci.* **74,** 29-36.

57. Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., Golden, S. S., Johnson, C. H., & Kondo, T. (1998) *Science* **281,** 1519-1523.

58. Schmitz, O., Katayama, M., Williams, S. B., Kondo, T., & Golden, S. S. (2000) *Science* **289,** 765-768.

59. Katayama, M., Kondo, T., Xiong, J., & Golden, S. S. (2003) *J. Bacteriol.* **185,** 1415-1422.

60. Iwasaki, H., Williams, S. B., Kitayama, Y., Ishiura, M., Golden, S. S., & Kondo, T. (2000) *Cell* **101,** 223-233.

61. Takai, N., Nakajima, M., Oyama, T., Kito, R., Sugita, C., Sugita, M., Kondo, T., & Iwasaki, H. (2006) *Proc. Natl. Acad. Sci. USA* **103,** 12109-12114.

62. Kutsuna, S., Kondo, T., Aoki, S., & Ishiura, M. (1998) *J. Bacteriol.* **180,** 2167-2174.

63. Katayama, M., Tsinoremas, N. F., Kondo, T., & Golden, S. S. (1999) *J. Bacteriol.* **181,** 3516-3524.

64. Taniguchi, Y., Katayama, M., Ito, R., Takai, N., Kondo, T., & Oyama, T. (2007) *Genes Dev.* **21,** 60-70.

65. Holtman, C. K., Chen, Y., Sandoval, P., Gonzales, A., Nalty, M. S., Thomas, T. L., Youderian, P., & Golden, S. S. (2005) *DNA Res.* **12,** 103-115.

66. Tsinoremas, N. F., Ishiura, M., Kondo, T., Andersson, C. R., Tanaka, K., Takahashi, H., Johnson, C. H., & Golden, S. S. (1996) *EMBO J.* **15,** 2488-2495.

67. Nair, U., Ditty, J. L., Min, H., & Golden, S. S. (2002) *J. Bacteriol.* **184,** 3530-3538.

68. Xu, Y., Mori, T., & Johnson, C. H. (2000) *EMBO J.* **19,** 3349-3357.

69. Vakonakis, I., Klewer, D. A., Williams, S. B., Golden, S. S., & LiWang, A. C. (2004) *J. Mol. Biol.* **342,** 9-17.

70. Kageyama, H., Kondo, T., & Iwasaki, H. (2003) *J. Biol. Chem.* **278,** 2388-2395.

71. Mutsuda, M., Michel, K. P., Zhang, X., Montgomery, B. L., & Golden, S. S. (2003) *J. Biol. Chem.* **278,** 19102-19110.

72. Zhang, X., Dong, G., & Golden, S. S. (2006) *Mol. Microbiol.* **60,** 658-668.

73. Ivleva, N. B., Gao, T., LiWang, A. C., & Golden, S. S. (2006) *Proc. Natl. Acad. Sci. USA* **103,** 17468-17473.

74. Gao, T., Zhang, X., Ivleva, N. B., Golden, S. S., & LiWang, A. (2007) *Protein Sci.* **16,** 465-475.

75. Dvornyk, V. (2005) *J. Mol. Evol.* **60,** 105-112.

76. Ivleva, N. B., Bramlett, M. R., Lindahl, P. A., & Golden, S. S. (2005) *EMBO J.* **24,** 1202-1210.

77. Takai, N., Ikeuchi, S., Manabe, K., & Kutsuna, S. (2006) *J. Biol. Rhythms* **21,** 235-244.

78. Kutsuna, S., Nakahira, Y., Katayama, M., Ishiura, M., & Kondo, T. (2005) *Mol. Microbiol.* **57,** 1474-1484.

79. Ditty, J. L., Canales, S. R., Anderson, B. E., Williams, S. B., & Golden, S. S. (2005) *Microbiology* **151,** 2605-2613.

80. Xu, Y., Mori, T., & Johnson, C. H. (2003) *EMBO J.* **22,** 2117-2126.

81. Nakahira, Y., Katayama, M., Miyashita, H., Kutsuna, S., Iwasaki, H., Oyama, T., & Kondo, T. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 881-885.

82. Tomita, J., Nakajima, M., Kondo, T., & Iwasaki, H. (2005) *Science* **307,** 251-254.

83. Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., & Kondo, T. (2005) *Science* **308,** 414-415.

84. Liu, Y., Tsinoremas, N. F., Johnson, C. H., Lebedeva, N. V., Golden, S. S., Ishiura, M., & Kondo, T. (1995) *Genes Dev.* **9,** 1469-1478.

85. Smith, R. M. & Williams, S. B. (2006) *Proc. Natl. Acad. Sci. USA* **103,** 8564-8569.

86. Woelfle, M. A. & Johnson, C. H. (2006) *J. Biol. Rhythms.* **21,** 419-431.

87. Lakin-Thomas, P. L. (2006) *J. Biol. Rhythms* **21,** 83-92.

88. Ye, S., Vakonakis, I., Ioerger, T. R., LiWang, A. C., & Sacchettini, J. C. (2004) *J. Biol. Chem.* **279,** 20511-20518.

89. Hitomi, K., Oyama, T., Han, S., Arvai, A. S., & Getzoff, E. D. (2005) *J. Biol. Chem.* **280,** 19127-19135.

90. Iwase, R., Imada, K., Hayashi, F., Uzumaki, T., Morishita, M., Onai, K., Furukawa, Y., Namba, K., & Ishiura, M. (2005) *J. Biol. Chem.* **280,** 43141-43149.

91. Hayashi, F., Suzuki, H., Iwase, R., Uzumaki, T., Miyake, A., Shen, J. R., Imada, K., Furukawa, Y., Yonekura, K., Namba, K*., et al.* (2003) *Genes Cells* **8,** 287-296.

92. Pattanayek, R., Wang, J., Mori, T., Xu, Y., Johnson, C. H., & Egli, M. (2004) *Mol. Cell* **15,** 375-388.

93. Nishiwaki, T., Satomi, Y., Nakajima, M., Lee, C., Kiyohara, R., Kageyama, H., Kitayama, Y., Temamoto, M., Yamaguchi, A., Hijikata, A*., et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 13927-13932.

94. Xu, Y., Mori, T., Pattanayek, R., Pattanayek, S., Egli, M., & Johnson, C. H. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 13933-13938.

95. Nishiwaki, T., Iwasaki, H., Ishiura, M., & Kondo, T. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 495-499.

96. Hayashi, F., Itoh, N., Uzumaki, T., Iwase, R., Tsuchiya, Y., Yamakawa, H., Morishita, M., Onai, K., Itoh, S., & Ishiura, M. (2004) *J. Biol. Chem.* **279,** 52331-52337.

97. Kitayama, Y., Iwasaki, H., Nishiwaki, T., & Kondo, T. (2003) *EMBO J.* **22,** 2127-2134.

98. Williams, S. B., Vakonakis, I., Golden, S. S., & LiWang, A. C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 15357-15362.

99. Iwasaki, H., Nishiwaki, T., Kitayama, Y., Nakajima, M., & Kondo, T. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 15788-15793.

100. Vakonakis, I. & LiWang, A. C. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 10925-10930.

101. Pattanayek, R., Williams, D. R., Pattanayek, S., Xu, Y., Mori, T., Johnson, C. H., Stewart, P. L., & Egli, M. (2006) *EMBO J.* **25,** 2017-2028.

102. Garces, R. G., Wu, N., Gillon, W., & Pai, E. F. (2004) *EMBO J.* **23,** 1688-1698.

103. Iwasaki, H., Taniguchi, Y., Ishiura, M., & Kondo, T. (1999) *EMBO J.* **18,** 1137-1145.

104. Mori, T., Williams, D. R., Byrne, M. O., Qin, X., Egli, M., McHaourab, H. S., Stewart, P. L., & Johnson, C. H. (2007) *PLoS Biol.* **5,** e93.

105. Kageyama, H., Nishiwaki, T., Nakajima, M., Iwasaki, H., Oyama, T., & Kondo, T. (2006) *Mol. Cell.* **23,** 161-171.

106. Dvornyk, V., Vinogradova, O., & Nevo, E. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 2495-2500.

107. Dunlap, J. C., Loros, J. J., Liu, Y., & Crosthwaite, S. K. (1999) *Genes Cells* **4,** 1-10.

108. Min, H., Y. Liu, C. H. Johnson, & Golden., S. S. (2004) *J. Biol. Rhythms* **19**.

109. Kiyohara, Y. B., Katayama, M., & Kondo, T. (2005) *J. Bacteriol.* **187,** 2559-2564.

110. Mills, D. A. (2001) *Curr. Opin. Biotechnol.* **12,** 503-509.

111. Pajunen, M. I., Pulliainen, A. T., Finne, J., & Savilahti, H. (2005) *Microbiology* **151,** 1209-1218.

112. Gehring, A. M., Nodwell, J. R., Beverley, S. M., & Losick, R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 9642-9647.

113. Michielse, C. B., Hooykaas, P. J., van den Hondel, C. A., & Ram, A. F. (2005) *Curr. Genet.* **48,** 1-17.

114. Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R.*, et al.* (2003) *Science* **301,** 653-657.

115. Tadege, M., Ratet, P., & Mysore, K. S. (2005) *Trends Plant Sci.* **10,** 229-235.

116. Granger, L., Martin, E., & Segalat, L. (2004) *Nucleic Acids Res.* **32,** e117.

117. Carlson, C. M. & Largaespada, D. A. (2005) *Nat. Rev. Genet.* **6,** 568-580.

118. Kitada, K., Ishishita, S., Tosaka, K., Takahashi, R., Ueda, M., Keng, V. W., Horie, K., & Takeda, J. (2007) *Nat. Methods* **4,** 131-133.

119. Balu, B. & Adams, J. H. (2006) *Cell Microbiol.* **8,** 1529-1536.

120. Bourhy, P., Louvel, H., Saint Girons, I., & Picardeau, M. (2005) *J. Bacteriol.* **187,** 3255-3258.

121. Su, J., Yang, J., Zhao, D., Kawula, T. H., Banas, J. A., & Zhang, J. R. (2007) *Infect. Immun.* **75**, 3089-3101.

122. Singer, B. & Kusmierek, J. T. (1982) *Annu. Rev. Biochem.* **51,** 655-693.

123. Lawrence, C. W., Gibbs, P. E., Borden, A., Horsfall, M. J., & Kilbey, B. J. (1993) *Mutat. Res.* **299,** 157-163.

124. Maloy, S. R. (2007) *Methods Enzymol.* **421,** 11-17.

125. Hayes, F. (2003) *Annu. Rev. Genet.* **37,** 3-29.

126. Pobigaylo, N., Wetter, D., Szymczak, S., Schiller, U., Kurtz, S., Meyer, F., Nattkemper, T. W., & Becker, A. (2006) *Appl. Environ. Microbiol.* **72,** 4329-4337.

127. Vidan, S. & Snyder, M. (2001) *Curr. Opin. Biotechnol.* **12,** 28-34.

128. Goryshin, I. Y. & Reznikoff, W. S. (1998) *J. Biol. Chem.* **273,** 7367-7374.

129. Kirby, J. R. (2007) *Methods Enzymol.* **421,** 17-21.

130. Biery, M. C., Stewart, F. J., Stellwagen, A. E., Raleigh, E. A., & Craig, N. L. (2000) *Nucleic Acids Res.* **28,** 1067-1077.

131. Haapa, S., Taira, S., Heikkinen, E., & Savilahti, H. (1999) *Nucleic Acids Res.* **27,** 2777-2784.

132. Cohen, M. F., Wallis, J. G., Campbell, E. L., & Meeks, J. C. (1994) *Microbiology* **140,** 3233-3240.

133. Koksharova, O. A. & Wolk, C. P. (2002) *Appl. Microbiol. Biotechnol.* **58,** 123-137.

134. Miyagishima, S. Y., Wolk, C. P., & Osteryoung, K. W. (2005) *Mol. Microbiol.* **56,** 126-143.

135. McCarren, J. & Brahamsha, B. (2005) *J. Bacteriol.* **187,** 4457-4462.

136. Zhang, S., Laborde, S. M., Frankel, L. K., & Bricker, T. M. (2004) *J. Bacteriol.* **186,** 875-879.

137. Bhaya, D., Takahashi, A., Shahi, P., & Grossman, A. R. (2001) *J. Bacteriol.* **183,** 6140-6143.

138. Nishimura, H., Nakahira, Y., Imai, K., Tsuruhara, A., Kondo, H., Hayashi, H., Hirai, M., Saito, H., & Kondo, T. (2002) *Microbiology* **148,** 2903-2909.

139. Lavoie, B. D. & Chaconas, G. (1996) *Curr. Top. Microbiol. Immunol.* **204,** 83-102.

140. Mizuuchi, K. (1992) *Annu. Rev. Biochem.* **61,** 1011-1051.

141. Haapa, S., Suomalainen, S., Eerikainen, S., Airaksinen, M., Paulin, L., & Savilahti, H. (1999) *Genome Res.* **9,** 308-315.

142. Haapa-Paananen, S., Rita, H., & Savilahti, H. (2002) *J. Biol. Chem.* **277,** 2843-2851.

143. Mizuuchi, M. & Mizuuchi, K. (1993) *Cold Spring Harb. Symp. Quant. Biol.* **58,** 515-523.

144. Herdman, M., Janvier, M., Waterbury, J. B., Rippka, R., Stanier, R. Y., & Mandel, M. (1979) *J. Gen. Microbiol.* **111,** 63-71.

145. Allewalt, J. P., Bateson, M. M., Revsbech, N. P., Slack, K., & Ward, D. M. (2006) *Appl. Environ. Microbiol.* **72,** 544-550.

146. Ferris, M. J., Muyzer, G., & Ward, D. M. (1996) *Appl. Environ. Microbiol.* **62,** 340-346.

147. Kaneko, T., Matsubayashi, T., Sugita, M., & Sugiura, M. (1996) *Plant Mol. Biol.* **31,** 193-201.

148. Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998) *Genome Res.* **8,** 175-185.

149. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186-194.

150. Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007) *Bioinformatics*. **23,** 673-679.

151. Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636-4641.

152. Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998) *Nucleic Acids Res.* **26,** 544-548.

153. Badger, J. H. & Olsen, G. J. (1999) *Mol. Biol. Evol.* **16,** 512-524.

154. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195-197.

155. Eddy, S. R. (1998) *Bioinformatics* **14,** 755-763.

156. Kuhsel, M. G., Strickland, R., & Palmer, J. D. (1990) *Science* **250,** 1570-1573.

157. Sugita, M., Luo, L., Ohta, M., Itadani, H., Matsubayashi, T., & Suguira, M. (1995) *DNA Res.* **2,** 71-76.

158. Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I.*, et al.* (2006) *Nucleic Acids Res.* **34,** D344-348.

159. Ulrich, L. E., Koonin, E. V., & Zhulin, I. B. (2005) *Trends Microbiol.* **13,** 52-56.

160. Ulrich, L. E. & Zhulin, I. B. (2007) *Nucleic Acids Res.* **35,** D386-390.

161. Stothard, P. & Wishart, D. S. (2005) *Bioinformatics* **21,** 537-539.

162. Liu, Y. & Tsinoremas, N. F. (1996) *Gene* **172,** 105-109.

163. Bustos, S. A. & Golden, S. S. (1992) *Mol. Gen. Genet.* **232,** 221-230.

164. Morrow, I. C. & Parton, R. G. (2005) *Traffic* **6,** 725-740.

165. Gupta, A., Morby, A. P., Turner, J. S., Whitton, B. A., & Robinson, N. J. (1993) *Mol. Microbiol.* **7,** 189-195.

166. Robinson, N. J., Robinson, P. J., Gupta, A., Bleasby, A. J., Whitton, B. A., & Morby, A. P. (1995) *Nucleic Acids Res.* **23,** 729-735.

167. Robinson, P. J., Cranenburgh, R. M., Head, I. M., & Robinson, N. J. (1997) *Mol. Microbiol.* **24,** 181-189.

168. Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T*., et al.* (2002) *DNA Res.* **9,** 123-130.

169. Gerdes, K., Christensen, S. K., & Lobner-Olesen, A. (2005) *Nat. Rev. Microbiol.* **3,** 371-382.

170. Hayes, C. S. & Sauer, R. T. (2003) *Cell* **112,** 2-4.

171. Pandey, D. P. & Gerdes, K. (2005) *Nucleic Acids Res.* **33,** 966-976.

172. Fico, S. & Mahillon, J. (2006) *BMC Genomics* **7,** 259.

173. Stock, A. M., Robinson, V. L., & Goudreau, P. N. (2000) *Annu. Rev. Biochem.* **69,** 183-215.

174. Ryjenkov, D. A., Tarutina, M., Moskvin, O. V., & Gomelsky, M. (2005) *J. Bacteriol.* **187,** 1792-1798.

175. Schmidt, A. J., Ryjenkov, D. A., & Gomelsky, M. (2005) *J. Bacteriol.* **187,** 4774-4781.

176. Schmitz, O., Boison, G., & Bothe, H. (2001) *Mol. Microbiol.* **41,** 1409-1417.

177. van der Horst, M. A. & Hellingwerf, K. J. (2004) *Acc. Chem. Res.* **37,** 13-20.

178. Lin, C. & Todo, T. (2005) *Genome Biol.* **6,** 220.

179. Hitomi, K., Okamoto, K., Daiyasu, H., Miyashita, H., Iwai, S., Toh, H., Ishiura, M., & Todo, T. (2000) *Nucleic Acids Res.* **28,** 2353-2362.

180. Yasui, A., Takao, M., Oikawa, A., Kiener, A., Walsh, C. T., & Eker, A. P. (1988) *Nucleic Acids Res.* **16,** 4447-4463.

181. Tamada, T., Kitadokoro, K., Higuchi, Y., Inaka, K., Yasui, A., de Ruiter, P. E., Eker, A. P., & Miki, K. (1997) *Nat. Struct. Biol.* **4,** 887-891.

182. Crosson, S., Rajagopal, S., & Moffat, K. (2003) *Biochemistry* **42,** 2-10.

183. Narikawa, R., Zikihara, K., Okajima, K., Ochiai, Y., Katayama, M., Shichida, Y., Tokutomi, S., & Ikeuchi, M. (2006) *Photochem. Photobiol.* **82,** 1627-1633.

184. Montgomery, B. L. (2007) *Mol. Microbiol.* **64,** 16-27.

185. Palenik, B. & Haselkorn, R. (1992) *Nature* **355,** 265-267.

186. Urbach, E., Robertson, D. L., & Chisholm, S. W. (1992) *Nature* **355,** 267-270.

187. Arita, K., Hashimoto, H., Igari, K., Akaboshi, M., Kutsuna, S., Sato, M., & Shimizu, T. (2007) *J. Biol. Chem.* **282,** 1128-1135.

188. Schelin, J., Lindmark, F., & Clarke, A. K. (2002) *Microbiology* **148,** 2255-2265.

189. Parkinson, J. S. & Kofoid, E. C. (1992) *Annu. Rev. Genet.* **26,** 71-112.

190. Strayer, C., Oyama, T., Schultz, T. F., Raman, R., Somers, D. E., Mas, P., Panda, S., Kreps, J. A., & Kay, S. A. (2000) *Science* **289,** 768-771.

191. Matsushika, A., Makino, S., Kojima, M., & Mizuno, T. (2000) *Plant Cell Physiol.* **41,** 1002-1012.

192. Makino, S., Matsushika, A., Kojima, M., Oda, Y., & Mizuno, T. (2001) *Plant Cell Physiol.* **42,** 334-339.

193. Schwarz, R. & Grossman, A. R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11008-11013.

194. Thomas, C., Andersson, C. R., Canales, S. R., & Golden, S. S. (2004) *Microbiology* **150,** 1031-1040.

195. Ishikawa, J. & Hotta, K. (1999) *FEMS Microbiol. Lett.* **174,** 251-253.

196. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403-410.

197. Gordon, D., Desmarais, C., & Green, P. (2001) *Genome Res.* **11,** 614-625.

198. Gordon, D., Abajian, C., & Green, P. (1998) *Genome Res.* **8,** 195-202.

199. Mizuuchi, K. & Craigie, R. (1986) *Annu. Rev. Genet.* **20,** 385-429.

200. Reznikoff, W. S. (1993) *Annu. Rev. Microbiol.* **47,** 945-963.

201. Clerico, E. M., Ditty, J. L., & Golden, S. S. (2007) *Methods Mol. Biol.* **362,** 155-171.

202. Mackey, S. R., Ditty, J. L., Clerico, E. M., & Golden, S. S. (2007) *Methods Mol. Biol.* **362,** 115-129.

203. Golden, S. S., Brusslan, J., & Haselkorn, R. (1986) *EMBO J.* **5,** 2789-2798.

204. Nair, U., Thomas, C., & Golden, S. S. (2001) *J. Bacteriol.* **183,** 1740-1747.

205. Liu, Y., Tsinoremas, N. F., Golden, S. S., Kondo, T., & Johnson, C. H. (1996) *Mol. Microbiol.* **20,** 1071-1081.

206. Maurizi, M. R., Clark, W. P., Kim, S. H., & Gottesman, S. (1990) *J. Biol. Chem.* **265,** 12546-12552.

207. Muller, E. G. & Buchanan, B. B. (1989) *J. Biol. Chem.* **264,** 4008-4014.

208. Golden, S. S. & Sherman, L. A. (1984) *J. Bacteriol.* **158,** 36-42.

209. Chen, I. & Dubnau, D. (2004) *Nat. Rev. Microbiol.* **2,** 241-249.

210. Dubnau, D. (1999) *Annu. Rev. Microbiol.* **53,** 217-244.

211. McBride, M. J. (2001) *Annu. Rev. Microbiol.* **55,** 49-75.

212. Nudleman, E. & Kaiser, D. (2004) *J. Mol. Microbiol. Biotechnol.* **7,** 52-62.

213. Shi, W. & Sun, H. (2002) *Infect. Immun.* **70,** 1-4.

214. Bhaya, D. (2004) *Mol. Microbiol.* **53,** 745-754.

215. Bhaya, D., Bianco, N. R., Bryant, D., & Grossman, A. (2000) *Mol. Microbiol.* **37,** 941-951.

216. Yoshihara, S., Geng, X., Okamoto, S., Yura, K., Murata, T., Go, M., Ohmori, M., & Ikeuchi, M. (2001) *Plant Cell Physiol.* **42,** 63-73.

217. Nakasugi, K. & Neilan, B. A. (2005) *Appl. Environ. Microbiol.* **71,** 7621-7625.

218. Duggan, P. S., Gottardello, P., & Adams, D. G. (2007) *J. Bacteriol.* **189**, 4547-4751.

219. Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzel, H., & Hess, W. R. (2005) *Genome Biol.* **6,** R73.

220. Valentin-Hansen, P., Eriksen, M., & Udesen, C. (2004) *Mol. Microbiol.* **51,** 1525-1533.

221. Guillier, M., Gottesman, S., & Storz, G. (2006) *Genes Dev.* **20,** 2338-2348.

222. Valentin-Hansen, P., Johansen, J., & Rasmussen, A. A. (2007) *Curr. Opin. Microbiol.* **10**, 152-155.

223. Rippka, R., Deruelles, J., Waterbury, J., Herdman, M., & Stanier, R. (1979) *J. Gen. Microbiol.* **111,** 1-61.

224. Mori, T., Saveliev, S. V., Xu, Y., Stafford, W. F., Cox, M. M., Inman, R. B., & Johnson, C. H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 17203-17208.

225. Taniguchi, Y., Yamaguchi, A., Hijikata, A., Iwasaki, H., Kamagata, K., Ishiura, M., Go, M., & Kondo, T. (2001) *FEBS Lett.* **496,** 86-90.

226. Vakonakis, I., Sun, J., Wu, T., Holzenburg, A., Golden, S. S., & LiWang, A. C. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 1479-1484.

227. Iwasaki, H. & Kondo, T. (2004) *J. Biol. Rhythms.* **19,** 436-444.

228. Aschoff, J. (1960) *Cold Spring Harb. Symp. Quant. Biol.* **25,** 11-28.

229. Braman, J., Papworth, C., & Greener, A. (1996) *Methods Mol. Biol.* **57,** 31-44.

230. Potts, M. (1984) *FEMS Microbiology Letters* **24** 351–354.

231. Rebière, M.-C., Anne-Marie-Castets, Houmard, J., & Marsac, N. T. d. (1986) *FEMS Microbiology Letters* **37** 269–275.

232. Gendel, S. M. (1988) *Curr. Microbiol.* **V17,** 23-26.

233. Felkner, R. H. & Barnum, S. R. (1988) *Curr. Microbiol.* **V17,** 37-41.

234. Schwabe, W., Weihe, A., Börner, T., Henning, M., & Kohl, J.-G. (1988) *Current Microbiology* **V17,** 133-137.

235. Bose, S. G. & Carmichael, W. W. (1990) *Journal of Applied Phycology* **V2,** 131-136.

236. Soper, B. W. & Reddy, K. J. (1995) *Curr. Microbiol.* **V31,** 169-173.

237. Xu, W. & McFadden, B. A. (1997) *Plasmid* **37,** 95-104.

238. Yang, X. & McFadden, B. A. (1994) *Plasmid* **31,** 131-137.

239. Yang, X. & McFadden, B. A. (1993) *J. Bacteriol.* **175,** 3981-3991.

240. Kaneko, T., Nakamura, Y., Sasamoto, S., Watanabe, A., Kohara, M., Matsumoto, M., Shimpo, S., Yamada, M., & Tabata, S. (2003) *DNA Res.* **10,** 221-228.

241. Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T.*, et al.* (2001) *DNA Res.* **8,** 205-213; 227-253.

242. Lau, R. H. & Doolittle, W. F. (1979) *J. Bacteriol.* **137,** 648-652.

243. Lau, R. H., Sapienza, C., & Doolittle, W. F. (1980) *Molecular and General Genetics MGG* **V178,** 203-211.

244. van den Hondel, C. A., Verbeek, S., van der Ende, A., Weisbeek, P. J., Borrias, W. E., & van Arkel, G. A. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 1570-1574.

245. Van der Plas, J., Oosterhoff-Teertstra, R., Borrias, M., & Weisbeek, P. (1992) *Mol. Microbiol.* **6,** 653-664.

246. Stanier, R. Y., Kunisawa, R., Mandel, M., & Cohen-Bazire, G. (1971) *Bacteriol. Rev.* **35,** 171-205.

247. van den Hondel, C. A., Keegstra, W., Borrias, W. E., & van Arkel, G. A. (1979) *Plasmid* **2,** 323-333.

248. Sherman, L. A. & van de Putte, P. (1982) *J. Bacteriol.* **150,** 410-413.

249. Golden, S. S. & Sherman, L. A. (1983) *J. Bacteriol.* **155,** 966-972.

250. Laudenbach, D. E., Straus, N. A., Gendel, S., & Williams, J. P. (1983) *Molecular and General Genetics MGG* **V192,** 402-407.

251. Laudenbach, D. E., Straus, N. A., & Williams, J. P. (1985) *Molecular and General Genetics MGG* **V199,** 300-305.

252. Nicholson, M. L., Gaasenbeek, M., & Laudenbach, D. E. (1995) *Molecular and General Genetics MGG* **V247,** 623-632.

253. Nicholson, M. L. & Laudenbach, D. E. (1995) *J. Bacteriol.* **177,** 2143-2150.

254. Tatusova, T. A. & Madden, T. L. (1999) *FEMS Microbiol. Lett.* **174,** 247-250.

255. Paithoonrangsarid, K., Shoumskaya, M. A., Kanesaki, Y., Satoh, S., Tabata, S., Los, D. A., Zinchenko, V. V., Hayashi, H., Tanticharoen, M., Suzuki, I.*, et al.* (2004) *J. Biol. Chem.* **279,** 53078-53086.

256. Nelson, D. O. (2007) in *Water Encyclopedia*. http://www.waterencyclopedia.com/En-Ge/Fresh-Water-Natural-Composition-of.html

257. Collier, J. K., Herbert, S. K., Fork, D. C., & Grossman, A. R. (1994) *Photosynthesis Research* **V42,** 173-183.

258. Schwarz, R. & Forchhammer, K. (2005) *Microbiology* **151,** 2503-2514.

259. Hayes, F. (2003) *Science* **301,** 1496-1499.

260. Ried, J. L. & Collmer, A. (1987) *Gene* **57,** 239-246.

261. Cai, Y. P. & Wolk, C. P. (1990) *J. Bacteriol.* **172,** 3138-3145.

262. Marchler-Bauer, A., Anderson, J. B., Derbyshire, M. K., DeWeese-Scott, C., Gonzales, N. R., Gwadz, M., Hao, L., He, S., Hurwitz, D. I., Jackson, J. D.*, et al.* (2007) *Nucleic Acids Res.* **35,** D237-240.

263. Fraser, C. M., Eisen, J., Fleischmann, R. D., Ketchum, K. A., & Peterson, S. (2000) *Emerg. Infect. Dis.* **6,** 505-512.

264. Hall, A. E., Fiebig, A., & Preuss, D. (2002) *Plant Physiol.* **129,** 1439-1447.

265. Dufresne, A., Garczarek, L., & Partensky, F. (2005) *Genome Biol.* **6,** R14.

266. Hess, W. R. (2004) *Curr. Opin. Biotechnol.* **15,** 191-198.

267. Martin, K. A., Siefert, J. L., Yerrapragada, S., Lu, Y., McNeill, T. Z., Moreno, P. A., Weinstock, G. M., Widger, W. R., & Fox, G. E. (2003) *Photosynth. Res.* **75,** 211-221.

268. Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D*., et al.* (2006) *Proc. Natl. Acad. Sci. USA* **103,** 13126-13131.

269. Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J*., et al.* (2003) *Nature* **424,** 1037-1042.

270. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., & Blankenship, R. E. (2002) *Science* **298,** 1616-1620.

271. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R*., et al.* (2003) *Nature* **424,** 1042-1047.

272. Sato, N. (2002) *Genome Inform.* **13,** 173-182.

273. Palenik, B., Ren, Q., Dupont, C. L., Myers, G. S., Heidelberg, J. F., Badger, J. H., Madupu, R., Nelson, W. C., Brinkac, L. M., Dodson, R. J*., et al.* (2006) *Proc. Natl. Acad. Sci. USA* **103,** 13555-13559.

274. Honda, D., Yokota, A., & Sugiyama, J. (1999) *J. Mol. Evol.* **48,** 723-739.

275. Robertson, B. R., Tezuka, N., & Watanabe, M. M. (2001) *Int. J. Syst. Evol. Microbiol.* **51,** 861-871.

276. Herranen, M., Aro, E. M., & Tyystjarvi, T. (2001) *Physiol. Plant* **112,** 531-539.

277. Salih, G. F. & Jansson, C. (1997) *Plant Cell* **9,** 869-878.

278. Tauber, E., Last, K. S., Olive, P. J., & Kyriacou, C. P. (2004) *J. Biol. Rhythms* **19,** 445-458.

279. Dvornyk, V. & Knudsen, B. (2005) *Genetica* **124,** 247-254.

280. Dvornyk, V., Vinogradova, O., & Nevo, E. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2082-2087.

281. Dvornyk, V. & Nevo, E. (2004) *J. Mol. Evol.* **58,** 341-347.

282. Min, H., Guo, H., & Xiong, J. (2005) *FEBS Lett.* **579,** 808-812.

283. Mori, T., Binder, B., & Johnson, C. H. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10183-10188.

284. Parrott, L. M. & Slater, J. H. (1980) *Arch. Microbiol.* **127,** 53-58.

285. Tsinoremas, N. F., Kutach, A. K., Strayer, C. A., & Golden, S. S. (1994) *J. Bacteriol.* **176,** 6764-6768.

286. Mori, T. & Johnson, C. H. (2001) *J. Bacteriol.* **183,** 2439-2444.

287. Huang, T.-C., Chen, H.-M., Pen, S.-Y., & Chen, T.-H. (1994) *Planta* **193,** 131-136.

VITA

Name: You Chen

Address: Texas A&M University

Department of Biology

BSBE 320; Mail Stop 3258

College Station, TX 77843-3258

Email Address: ychen@mail.bio.tamu.edu

Education: B.S., Botany, Nanjing University, China, 1994

M.S., Genetics, Chinese Academy of Sciences, Shanghai, 2000