EVOLUTION AND DYNAMICS OF TRANSCRIPTIONAL REGULATION IN BACTERIA

by

Shan Li

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2013

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Anthony Fodor

_____
Dr. Jennifer Weller

_____
Dr. Jun-tao Guo

_____
Dr. Shan Yan

ABSTRACT

SHAN LI. Evolution and dynamics of transcriptional regulation in bacteria.
(Under the direction of DR ZHENGCHANG SU)

Although transcription is one of the most important biological functions of cells, our understanding of its regulation is still limited. In this dissertation, we have studied the transcriptional regulation in prokaryotes in three aspects. First, we investigated the extent to which *cis*-regulatory elements are conserved during the course of evolution using the *LexA* regulons in cyanobacteria as an example. We found that in most cyanobacterial genomes analyzed, *LexA* appears to function as the transcriptional regulator of the key SOS response genes. The loss of *lexA* in some genomes might lead to the degradation of its binding sites. Second, directional RNA-seq techniques have recently become the workhorse for transcriptome profiling in prokaryotes, however, it is a challenging task to accurately assemble highly labile prokaryotic transcriptomes for further analyses. To fill this gap, we have developed a hidden Markov model based transcriptome assembler which outperforms the state-of-the-art assemblers. Using our tool, we characterized alternative operon structures in *E. coli* K12 under various growth conditions and growth phases, and found that they are more complex and dynamic than previously anticipated. Lastly, we determined anti-sense and non-coding transcription patterns in *E. coli* K12 under various growth conditions and time points. We found that a large portion of genes have antisense transcription in a condition-dependent manner. Most antisense transcripts are initiated and restricted to the 5'-end of the gene on the sense strand, and their expression levels are correlated with those of the genes on the sense strand, suggesting that these antisense transcripts might play an important role in transcriptional regulation.

DEDICATION

For my parents and teachers who introduced me into science.

ACKNOWLEDGMENT

My greatest appreciation goes to my advisor, Dr. Zhengchang Su for his valuable guidance and patience over the years. He provides me a fertile ground and free academic environment to learn how to do research, and brings me with a mathematics major into the wonderful field of Bioinformatics.

I would also like to thank members in our group for their valuable suggestions and support, both past and current, especially, Dr. Shaoqiang Zhang, who gave me a lot of inspirations and help in my earlier stage of research and study, and Dr. Xia Dong, who performed molecular biology work in this project.

I sincerely thank my committee members Dr. Jun-tao Guo, Dr. Anthony Fodor, Dr. Jennifer Weller, and Dr. Shan Yan, for their time and kindly guidance both on class and research.

I am very grateful to my advisor, the Department of Bioinformatics and Genomics, and UNC Charlotte graduate school for providing me financial support in the past years. I would also like to express my appreciation to all the staff of ISSO and Bioinformatics department for their friendly help.

In addition, I would like to appreciate Dr. Guojun Li, who introduced me to the fascinating field of research.

Finally, I would like to dedicate my deep gratitude to my family for their unconditional love and support all these years. With the care, support and courage of my parents, Mr. Zhongxian Li and Ms. Wei Pan, as well as the love and tolerance of my fiancé, Mr. Wei Song, the journey of my life is more meaningful and wonderful.

INTRODUCTION

Prokaryotic genomes generally consist of two types of sequences: protein- or RNA-specifying *coding sequences* and intergenic *non-coding sequences*, with the former and the latter being about 85% and 15% of the genomes, respectively. While the coding sequences define the molecules that a cell can have, the non-coding sequences contain the regulatory sequences controlling the expression of the coding sequences in the cell under different growth phases and physiological as well as environmental conditions. In prokaryotes, several adjacent genes on the same strand of DNA can be co-transcribed as a polycistronic mRNA, thereby forming a multi-gene transcription unit called an operon. A group of operons and singleton genes regulated by the same transcription factor (TF) is called a regulon. In eubacteria, gene transcription initiation is controlled by the σ-factor of the RNA polymerase (RNAP) and other specific TFs binding to *cis*-regulatory elements in the upstream region of an operon [1]. It is the interactions of the transcriptome and its products in the cell that determine its functions. Therefore, a full understanding of the transcriptional regulation of prokaryotic cells can facilitate the understanding of their physiology and applications in medicine, agriculture and industry.

As transcriptional regulation in prokaryotes plays an important role in controlling their responses to environmental changes, thus it is subject to nature selection during the course of evolution. The structure and complexity of transcriptional regulatory network in prokaryotes have changed, reorganized, enabling them to adapt to almost every environmental niche on earth over millions of years. However, our general understanding of the rules that govern the evolution of the transcriptional regulation is still very limited,

such as how a TF and its binding site co-evolve, and to what extent the changes in transcriptional regulation contribute to the adaptation of prokaryotes to environments. Elucidation of these rules will help to characterize the transcriptional *cis*-regulatory networks in prokaryotes. In this dissertation, we have attempted to derive some of such rules through studying the evolution of the LexA regulon in cyanobacteria.

For a long time it is believed that that transcription largely occurs in the coding region resulting sense mRNAs, rRNAs and tRNAs, and that operons are static structures, when activated, they are transcribed uniformly, and alternative operons are rare. However, recent applications of whole genome directional (strand-specific) tiling array and directional RNA-seq techniques in transcriptome profiling in prokaryotes have completely changed our view of the architecture and complexity of prokaryotic transcriptomes [2-9]. For example, by using a combination of whole genome directional tiling array and RNA-seq techniques, Guell et al. [10] found that the operon utilizations in the reduced parasitic *M. pneumoniae* genome are highly variable and dynamic, with almost half of 139 identified multi-gene operons show varying (dynamic) expression in a staircase-like manner. Furthermore, under different conditions, operons could be divided into smaller sub-operons, resulting in many alternative transcripts, suggesting that the operon structures in *M. pneumoniae* is highly dynamic, more similar to that of alternative splicing in eukaryotes than originally thought. They also identified a large number of ncRNAs and asRNA expressed under various culture conditions, thus a much larger portion of the genome is transcribed than previously anticipated [10]. Similar observations are observed in *H. pylori* [11], *B. subtilis* [12], *Halobacterium salinarum* NRC-1 [13], and *Porphyromonas gingivalis* W83 [14]. However, not all these surprising

observations are noted in some other studies. For instance, pervasive alternative operon usages were not reported in many studies in a broader range of genomes, such as *E. coli* [15], *B. subtilis* [16], *Salmonella enterica serovar* Typhi [17], *Burkholderia cenocepacia* [18], *Caulobacter crescentus* [19], *Staphylococcus aureus* [20], *Vibrio cholera* [21], *Chlamydia trachomatis* [22], *Chlamydia pneumonia* [23], *Clostridium beijerinckii* NCIMB 8052 [24], *Listeria monocytogenes* [25], *Anabaena sp.* strain PCC 7120 [26], *Synechococcus elongatus* PCC 7942 [27], and *Sulfolobus solfataricus* P2 [28], even though multiple transcription starting sites (TSSs) in the upstream intergenic region of genes are frequently reported in most of these studies. Moreover, although most of these studies found extensive anti-senses and non-coding transcription, the levels of their prevalence can vary quite differently from different studies even in the same genome. Contradictory results have also been reported. For instance, although Rasmussen et al. [16] did not note alternative operon utilizations in *B. subtilis*, more recently, Nicolas et al. [12] observed prevalent condition-dependent operon utilizations using similar tiling array techniques. Notwithstanding that these discrepancies can be due to different experimental conditions for different research purposes in these studies, nevertheless, they inevitably raise the following questions needed to be urgently addressed: is the dynamic and alternative operon utilizations a ubiquitous phenomenon in all prokaryotes or only more prevalent in some specific species for their specific genome structures? What are the extent and patterns of anti-sense and non-coding transcriptions in the genomes? What are the molecular mechanisms that lead to their transcription under certain conditions and time points? And what are their biological significances? Furthermore, the existing transcriptome assembly tools are developed for eukaryotes, thus do not work well in

prokaryotes due to the highly labile nature of prokaryotic mRNA [29-32], thus, an tool for accurately assembling full-length prokaryotic transcripts is urgently needed. In this dissertation, we have addressed some of these questions by profiling the transcriptomes in *E. coli* K12 using a directional RNA-seq method. More specifically, we first sequenced the transcriptomes of the *E. coli* cells under a variety of culture conditions and growth phases. We then developed a Hidden Markov Model based algorithm and tool to assemble the full length transcripts from short directional RNA-seq reads. Finally, we analyzed the alternative operon utilizations and antisense transcription patterns and their possible biological functions under these culture conditions and growth phases.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: COMPUTATIONAL ANALYSIS OF LEXA REGULONS IN CYANOBACTERIA

1.1 Abstract

The transcription factor LexA plays an important role in the SOS response in *Escherichia coli* and many other bacterial species studied. Although the *lexA* gene is encoded in almost every bacterial group with a wide range of evolutionary distances, its precise functions in each group/species are largely unknown. More recently, it has been shown that *lexA* genes in two cyanobacterial genomes *Nostoc sp.* PCC 7120 and *Synechocystis sp.* PCC 6803 might have distinct functions other than the regulation of the SOS response. To gain a general understanding of the functions of LexA and its evolution in cyanobacteria, we conducted the current study.

Our analyses indicate that six of 33 sequenced cyanobacterial genomes do not harbor a *lexA* gene although they all encode the key SOS response genes, suggesting that LexA is not an indispensable transcription factor in these cyanobacteria, and that their SOS responses might be regulated by different mechanisms. Our phylogenetic analysis suggests that *lexA* was lost during the course of evolution in these six cyanobacterial genomes. For the 26 cyanobacterial genomes that encode a *lexA* gene, we have predicted their LexA-binding sites and regulons using an efficient binding site/regulon prediction algorithm that we developed previously. Our results show that LexA in most of these 26 genomes might still function as the transcriptional regulator of the SOS response genes as

seen in *E.coli* and other organisms. Interestingly, putative LexA-binding sites were also found in some genomes for some key genes involved in a variety of other biological processes including photosynthesis, drug resistance, etc., suggesting that there is crosstalk between the SOS response and these biological processes. In particular, LexA in both *Synechocystis sp.* PCC6803 and *Gloeobacter violaceus* PCC7421 has largely diverged from those in other cyanobacteria in the sequence level. It is likely that LexA is no longer a regulator of the SOS response in *Synechocystis sp.* PCC6803.

In most cyanobacterial genomes that we analyzed, LexA appears to function as the transcriptional regulator of the key SOS response genes. There are possible couplings between the SOS response and other biological processes. In some cyanobacteria, LexA has adapted distinct functions, and might no longer be a regulator of the SOS response system. In some other cyanobacteria, *lexA* appears to have been lost during the course of evolution. The loss of *lexA* in these genomes might lead to the degradation of its binding sites.

1.2 Background

The LexA protein was first characterized as the transcriptional regulator of the SOS response in *Escherichia coli* [33, 34], and later in several other bacteria, including *Bacillus subtilis* [35, 36] and *Fibrobacter succinogenes* [37]. In fact, the *lexA* gene is found in almost all eubacterial groups examined so far [37, 38]. In *E. coli*, around 30 genes involved in the SOS response are under the regulation of LexA [34]. Under normal growth conditions, LexA represses the SOS response genes by binding to their promoter regions, and thus blocking their transcription. When DNA is damaged, the binding of RecA to the released single-stranded DNA induces the auto-cleavage of the Ala^{84}-Gly^{85}

peptide bond [39, 40] in LexA, thereby inhibiting the dimerization of LexA and preventing its binding to DNA [41-43]. In this manner, SOS response genes are de-repressed and expressed at different time points and different levels in a coordinated way [42].

LexA in *E. coli* consists of an N-terminal DNA-binding domain and a C-terminal dimerization domain [40, 44].The N-terminal contains three α -helices (I, II, III) and an anti-parallel β -sheet [44]. Helices II and III form a helix-turn-helix DNA-binding motif, and all the DNA-contacting residues Ser^{39}, Asn^{41}, Ala^{42}, Glu^{44} and Glu^{45} are located in helix III [45] as revealed by both NMR [44] and X-Ray crystallography analyses [40]. The LexA-binding sites in *E. coli* were found to be a 16-bp palindromic motif with the consensus sequence $CTG(TA)_5CAG$ [46]. It has been shown that two reactive residues Ser^{119} and Lys^{156} in *E. coli* LexA are critical for the auto-hydrolysis of the peptide bond Ala^{84}-Gly^{85} [33, 41]. The core set of the SOS response system consists of *lexA, recA, uvrABCD, umuCD* and *ruvB* [42]. Upon the auto-hydrolysis of LexA, the *uvrABCD* operon is expressed first, whose products are responsible for the nucleotide excision repair (NER). Then *recA* and several other genes for homologous recombination are expressed, retrieving the excised DNA double strands. Next, the cell division inhibitor SfiA is induced to guarantee a sufficient time for the DNA repairing to be completed. In the end, if the DNA is not completely repaired, the operon *umuCD* encoding the mutagenic DNA repair polymerase Pol V will be induced to perform translesion DNA synthesis [42]. Since the *lexA* gene itself is also under the control of LexA, after the damaged DNA is repaired, the activity of RecA declines, the production of LexA surpasses its auto-cleavage. Consequently, the increased concentration of LexA restores

the inhibition of the expression of the SOS response genes.

More recently, LexA homologs were also experimentally studied in a few cyanobacteria [47-53]. These studies suggest that LexA in *Nostoc sp.* PCC 7120 [48] binds to the promoter regions of *lexA* and *recA*; however, LexA in *Synechocystis sp.* PCC 6803 may regulate different genes/systems other than the SOS system. Domain *et al.* concluded from microarray gene profiling analysis [53] that LexA in this species might be involved in carbon metabolism. Later, LexA in *Synechocystis sp.* PCC 6803 was found to regulate the *crhR* gene encoding a RNA helicase [51]. Moreover, it has been shown that the transcription of the bidirectional hydrogenase genes *hoxEFUYH* was regulated by LexA in *Synechocystis sp.* PCC6803[49]. In *Nostoc sp.* PCC 7120, *hoxEFUYH* genes are split into two separate operons, and LexA was found to bind to the upstream regions for both operons [47]. Mazon *et al.* [48] showed that the LexA-binding sites in *Nostoc sp.* PCC 7120 have a 14-bp pseudo-palindromic structure in the form of RGTACNNNDGTWCB, which are similar to those in *B. subtilis*. Additionally, Sjöholm *et al.* [47] found two putative palindromic LexA-binding sites: one in the promoter region of *alr0750-hoxE-hoxF* that resembles Mazon's LexA boxes[48], and another, TTACACTTTAA in the upstream region of *hoxU* in *Nostoc sp.* PCC 7120. Meanwhile, multiple putative LexA boxes were identified in *Synechocystis sp.* PCC6803: a 13-bp pseudo-palindromic segment AGTAACTAGTTCG in the upstream region of *hoxE*, which is similar to Mazon's site but with one base deletion [49]; another direct repeat pattern, CTA-N_9-CTA proposed to be recognized by LexA in vitro [52]; and two putative LexA boxes that resemble none of the putative LexA boxes listed above [50]. Despite this progress, a more extensive study of LexA proteins and their binding sites and regulons in

cyanobacterial genomes is still needed. In this study, we have predicted LexA-binding sites and regulons in all the sequenced cyanobacterial genomes that harbor a *lexA* gene, and analyzed the evolutionary changes in the LexA regulons in cyanobacteria, as well as their relationship with those in proteobacteria and gram-positive bacteria.

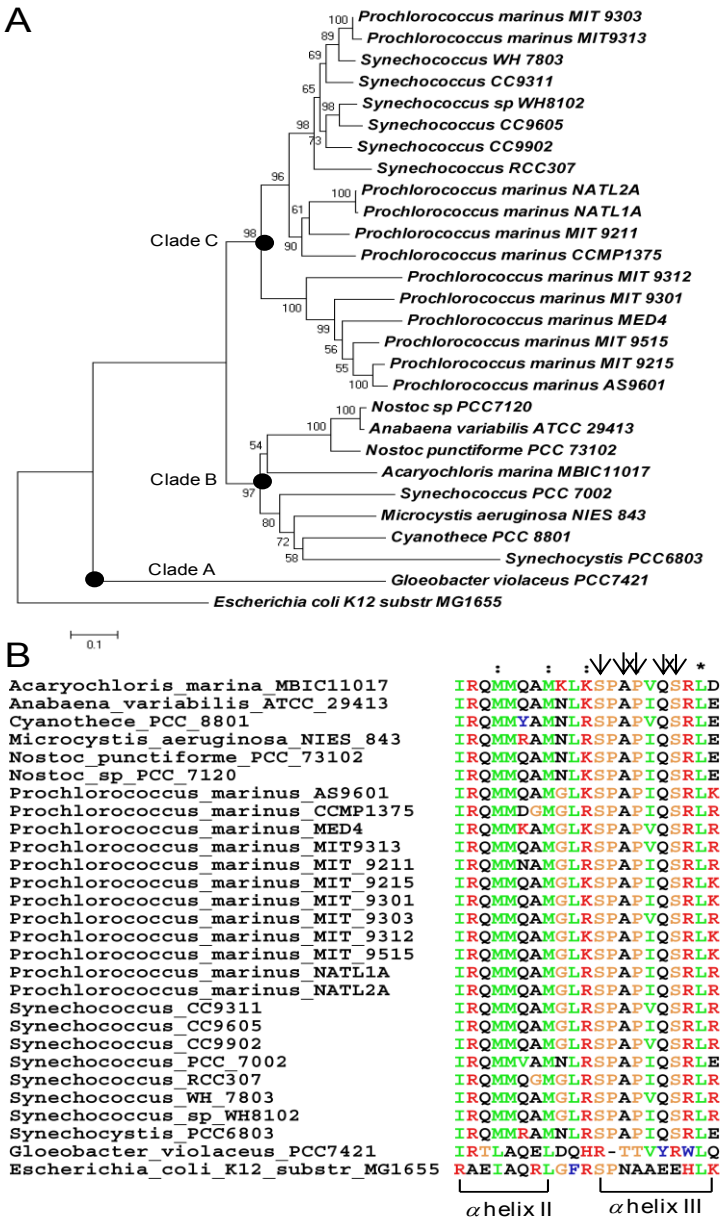


Figure 1.1. Phylogenetic relationships of 27 cyanobacterial LexA proteins and their DNA-binding domains.(A) Phylogenetic relationships of the 27 cyanobacterial LexA proteins. The tree is rooted with the LexA in *E. coli* K12. Bootstrap values are shown on

Figure 1.1 (continued) the nodes. (B) Alignment of the DNA-binding domain (DBD) of the 27 cyanobacterial LexA proteins. The DBD of LexA contains a helix-turn-helix motif, and DNA-contacting residues are located in helix III, and are labelled by vertical arrows.

1.3 Results and Discussion

1.3.1 Conservation of the DNA-binding domain of LexA in cyanobacteria

We identified orthologs of the LexA protein in *Nostoc sp.* PCC7120 (*alr4908*) in 26 of the 33 sequenced cyanobacterial genomes using the bi-directional best hit (BDBH) method based on BLASTP search with an *E*-value cutoff 10^{-10} (see Materials and Methods). Seven genomes appear not to harbor a *lexA* gene under this criterion, namely, *Gloeobacter violaceus* PCC7421, *Synechococcus sp.* JA-3-3Ab A-Prime, *Synechococcus sp.* JA-2-3B'a(2-13) B-Prime, *Synechococcus elongatus* PCC6301, *Synechococcus elongatus* PCC7942, *Trichodesmium erythraeum* IMS101 and *Thermosynechococcus elongatus* BP-1. We removed the *Synechococcus elongatus* PCC7942 genome from our study since *Synechococcus elongatus* PCC6301 is virtually identical to it [54]. However, an ortholog of the *lexA* gene (*Gll0709*) does exist in *Gloeobacter violaceus* PCC7421. The reason we failed to identify this ortholog is that it does not meet our BDBH criterion due to its largely divergent sequence. The phylogenetic tree of these 27 LexA amino acid sequences indicates that they can be clustered into three groups (Figure 1.1A), corresponding to the previously described Clade A (containing *Gloeobacter violaceus* PCC7421), Clade C (containing small marine *Prochlorococcus* and *Synechococcus*), and Clade B (containing most remaining cyanobacteria) [55]. However, aside from *Gloeobacter violaceus* PCC7421, the DNA-binding domains (DBD) of LexA from these cyanobacteria are highly conserved (Figure 1.1B), especially the helix III, where DNA-

contacting residues are located [45]. This result is in agreement with earlier observations [48, 53]. This provides the rationale of our analysis, including the phylogenetic footprinting analysis (next section) and genome-wide scanning for LexA-binding site predictions. On the other hand, since the DBD in *Gloeobacter violaceus* PCC7421 is quite different from those in other cyanobacteria, especially where the DNA-contacting residues locate, thus, we excluded it from our study, leaving 31 species/strains for the putative LexA regulon prediction.

1.3.2 LexA-binding sites predicted by phylogenetic footprinting

We considered both an operon and a singleton gene as a transcription unit (TU). As Mazon *et al.* [48] have demonstrated the binding of LexA to the upstream regions of two genes, *lexA* and *recA*, and predicted LexA-binding sites for other four genes (*uvrA*, *ssb*, *alr4905*, and *all4790*) in *Nostoc sp.* PCC7120, we used phylogenetic footprinting to identify possible LexA-binding sites in the pooled 118 inter-TU sequences associated with these six genes in *Nostoc sp.* PCC7120 [48] and their orthologs in the other 25 cyanobacterial genomes (excluding *Gloeobacter violaceus* PCC7421) that harbor a *lexA* gene (see Materials and Methods). We identified 49 high-scoring 14-bp palindromic sequences (Table 1) out of the 118 input sequences by applying the motif finding tools MEME [56] and BioProspector [57] and incorporating the best motifs found by these two programs (See Methods and Materials and Additional file 5). However, the putative LexA box AGTCCTAGAGTCCT (Additional file 5) identified in *Synechocystis sp.* PCC6803 was not identified by Patterson-Fortin *et al.* [52] using DNaseI footprinting assays or by Gutekunst *et al.*[49]. Therefore, we removed this site, leaving 48 putative LexA-binding sites (Table 1.1) for profile construction. The two LexA boxes that have been

Table 1.1: 48 Putative LexA binding sites identified by phylogenetic footprinting analysis

Genome	Transcription Unit	Name	Putative LexA-binding sites	Position[1]
Acaryochloris marina MBIC11017	*AM1_3549 AM1_3550*	*- recA*	**AATAAATCTGTACT**	-97
	AM1_3948	*lexA*	**AGTACAGGTGTTTT**	-132
Anabaena variabilis ATCC 29413	*Ava_2176*	-	**AGTTCTCATGTACT**	-144
	Ava_1462	-	**AGTACTTATGTACT**	-56
	Ava_3591	-	**AGTTCTTCTGTATC**	-112
	Ava_2198	*lexA*	**AGTACTAATGTTCT**	-47
	Ava_2059 Ava_2058	*- -*	**CGTACATTTGTACC**	-71
	Ava_4925	*recA*	**AGTATATCTGTTCT**	-93
Cyanothece PCC 8801	*PCC8801_0945*	-	**AAAACTCTTGTACT**	-78
	PCC8801_2186 PCC8801_2185	*- -*	**AGTACTTATGTTCG**	-101
Microcystis aeruginosa NIES 843	*MAE_39060*	*ssb*	**CATACTATTGTACT**	-59
	MAE_16070	*recA*	**CATACTGCTGTACT**	-68
Nostoc punctiforme PCC 73102	*Npun_F1842*	-	**AGTACACCTGTACT**	-56
	Npun_F2914	*recA*	**AGTATATCTGTTCT**	-102
	Npun_F6100 Npun_F6101 Npun_F6102	*- - -*	**AGTACGATTGTTCT**	-111
	Npun_R5568 Npun_R5567	*- -*	**CGTACATTTGTACT**	-74
Nostoc sp PCC7120	*alr4908*	*lexA*	**AGTACTAATGTTCT**	-35
	all4790 all4789	*- -*	**CGTACATTTGTACC**	-31
	alr4905	-	**AGTTCTCATGTACT**	-100
	alr3716	*uvrA*	**AGTACTATTGTTCT**	-72
	alr0088	*ssb*	**AGTACTTATGTACT**	-16
	all3272	*recA*	**AGTATATCTGTTCT**	-52
Prochlorococcus marinus AS9601	*A9601_17691*	*recA*	**AGTACAGATGTACT**	-126
Prochlorococcus marinus CCMP1375	*Pro1784*	*ssb*	**AAAACATAAGTATT**	-109
Prochlorococcus marinus MED4	*PMM1562*	*recA*	**AGTACACATGTACT**	-123
	PMM1262	*lexA*	**GGTACAAATGTATT**	-57
Prochlorococcus marinus MIT9313	*PMT0380*	-	**GGTACACATGTATT**	-56
Prochlorococcus marinus MIT9211	*P9211_13051 P9211_13041*	*lexA -*	**GGTACATATGTATT**	-69
Prochlorococcus marinus MIT9215	*P9215_18341*	*recA*	**AGTACAGATGTACT**	-126
Prochlorococcus marinus MIT9301	*P9301_17531*	*recA*	**AGTACAGATGTACT**	-125
Prochlorococcus marinus MIT9303	*P9303_19141*	*lexA*	**GGTACACATGTATT**	-81
Prochlorococcus marinus MIT9312	*PMT9312_1654*	*recA*	**AGTACAGATGTACT**	-126
Prochlorococcus marinus MIT9515	*P9515_17441*	*recA*	**AGTACGCATGTACT**	-123
	P9515_18121	-	**AATATATCTATTCT**	-139
Prochlorococcus marinus NATL1A	*NATL1_20071*	*recA*	**CGTACGTCTGTACT**	-132
	NATL1_16801	*lexA*	**AGGACAAATGTACT**	-52
Prochlorococcus marinus NATL2A	*PMN2A_1133*	*recA*	**CGTACGTCTGTACT**	-132
	PMN2A_0828	*lexA*	**AGGACGAATGTACT**	-52
Synechococcus CC9605	*Syncc9605_0929*	*lexA*	**GGTACAAATGTATT**	-61
	Syncc9605_0104	-	**GATACCGCAGTTTA**	-140
Synechococcus CC9902	*Syncc9902_1949*	*recA*	**CGTACGTTTGTACT**	-104
	Syncc9902_1481	*lexA*	**GGTACAAATGTATT**	-59
Synechococcus PCC7002	*SYNPCC7002_A0426 SYNPCC7002_A0425 SYNPCC7002_A0424*	*recA - -*	**AGTACGATTGAACT**	-90
	SYNPCC7002_A0119	*ssb*	**AGAACAGTTGTATG**	-53
Synechococcus RCC307	*SynRCC307_1756*	*lexA*	**GGCACAAATGTATT**	-39
Synechococcus WH7803	*SynWH7803_0171*	*ssb*	**CAACCGTCAGTTCT**	-56
	SynWH7803_0439	*recA*	**CGTACATCTGTACT**	-172
Synechococcus sp WH8102	*SYNW2062*	*recA*	**CGTACGCCTGTACT**	-104

1.Positions of the LexA binding sites relative to the first codon of the operon.

characterized in *Nostoc sp.* PCC 7120 [48] were accurately recovered by the phylogenetic footprinting procedure (Table 1.1), suggesting that most of these high-scoring motifs are likely to be genuine LexA boxes. These putative LexA-binding sites show either a strong

palindromic structure similar to the experimentally characterized LexA boxes in *Nostoc sp.* PCC7120 [48], or a tandem repeat structure with the consensus sequence

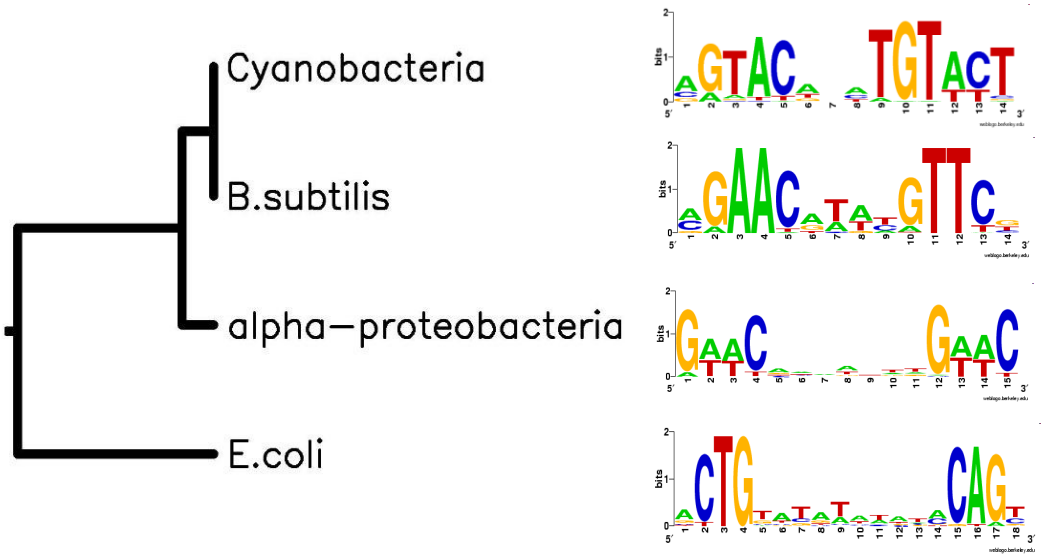


Figure 1.2. Phylogenetic tree of LexA binding sites in cyanobacteria, *B.subtilis*, a-proteobacteria and *E.coli*. Binding sites of cyanobacteria were predicted in this study, those of *B.subtilis* were from DBTBS [58], those of a-proteobacteria were taken from Erill et al [59], and those of *E.coli* were from RegulonDB [60]. Phylogenetic tree was constructed by the STAMP[61] web tool, sequence logos were generated by weblogo [62].

AGTACWNWTGTACT. As demonstrated in Figure 1.2, this pattern is rather similar to the consensus sequence of the LexA-binding sites previously identified in *B. subitlis* ($CGAACN_4GTTCG$) [35], and to a less extent, to that of LexA-binding sites found in α-proteobacteria ($GTTCN_7GTTC$ and $GAACN_7GAAC$) [59], but differs remarkably from that in *E. coli* $CTG(TA)_5CAG$ [46]. These results are consistent with our phylogenetic analysis of the 183 LexA proteins detected in 598 genomes, showing that LexA proteins in cyanobacteria are more closely related to those in gram-positive and α-proteobacteria bacteria than to those in γ-proteobacteria (Figure 1.3). Accordingly, since the LexA-binding sites in *B. subtilis* [35, 48] have a palindromic structure, it is not surprising that the LexA-binding sites in cyanobacterial genomes might have a similar palindromic

structure.

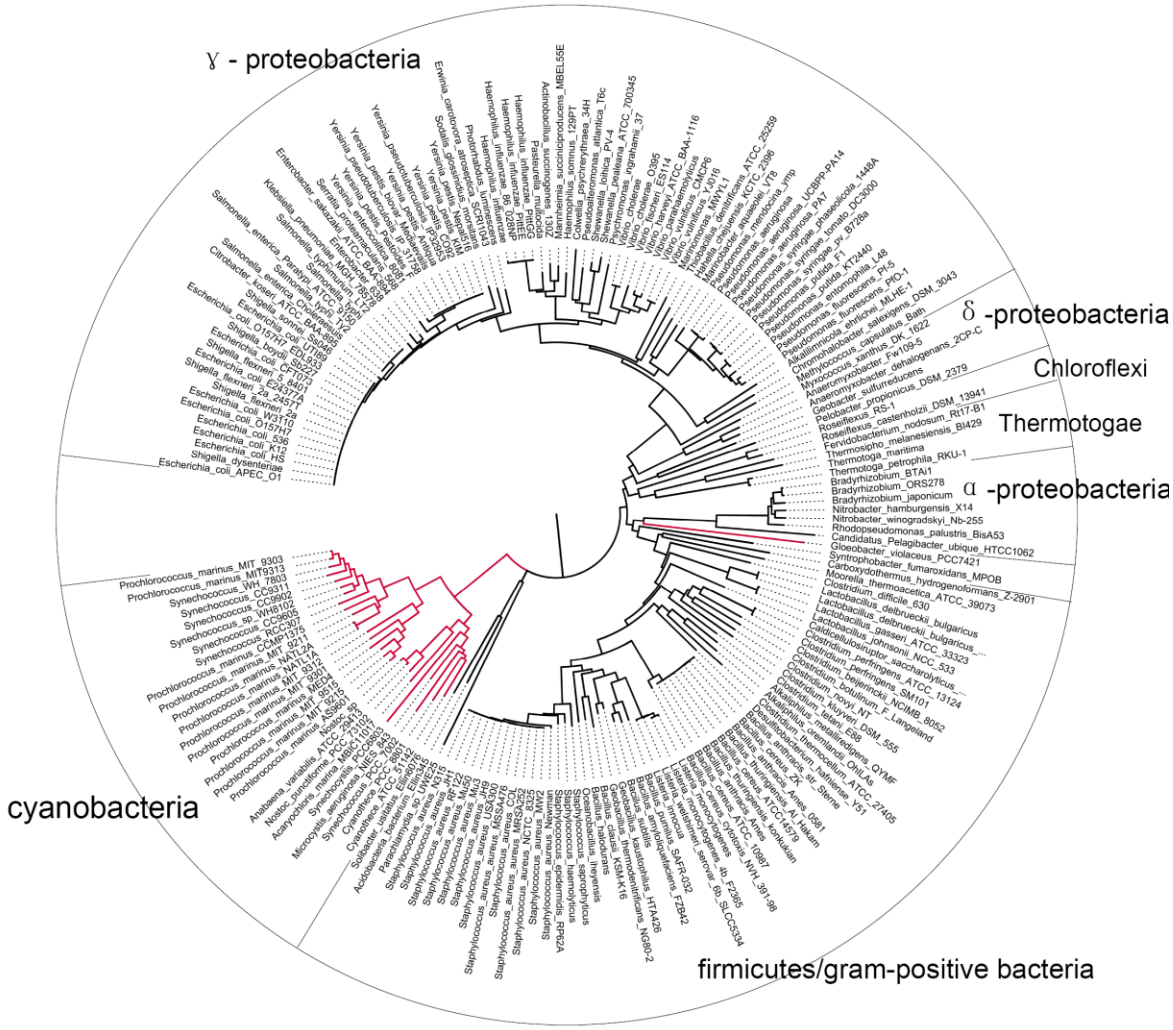


Figure 1.3. Phylogenetic tree of LexA sequences across a total of 183 cyanobacteria, gram-positive bacteria, α-proteobacteria, δ-proteobacteria, γ-proteobacteria and other bacterial species/strains. The tree was constructed in MEGA [63]. Branches of cyanobacteria are colored in red.

1.3.3 Genome-wide prediction of LexA-binding sites and regulons in cyanobacterial genomes

Both consensus sequence and position weight matrix (PWM) have been widely used to represent the pattern of similar sequences. The advantage of PWM (or profile) methods over the consensus sequence methods is that the former can capture more quantitative information about the patterns by using a probabilistic model to represent the sequences. In this way, it can differentiate subtly conserved positions from the non-conserved ones [64]. In our study, we used the profile of these 48 LexA boxes (Table 1.1) to scan the 31 sequenced cyanobacterial genomes to predict additional putative LexA-binding sites and members of LexA regulons, using a scanning algorithm [65-67] that incorporates orthologous information and computes a log-odds ratio (*LOR*) score for evaluating the confidence of predictions in each genome (see Materials and Methods for details). The predicted results with a p-value<0.01 for the 26 genomes harboring a *lexA* gene are listed in Tables S1-26 (Additional file 2), while those for the five genomes without a *lexA* gene are listed in Additional file 6. The predicted results with a *p*-value<0.05 for the 31 cyanobacteria are summarized in Additional file 3.

The score of a detected putative LexA binding site for a TU is the sum of two terms: one evaluates the extent to which the putative LexA binding site resembles the scanning profile; the other evaluates the similarity of this binding site to those identified for the orthologs of genes within the TU in the other genomes. To evaluate the confidence of each motif score *s*, we used randomly selected coding sequences as the null model to test the statistical significance. A false positive rate was used to evaluate this statistical significance, which was defined as the fraction of the randomly selected coding sequences containing binding sites with a score higher than the cutoff *s* in the genome.

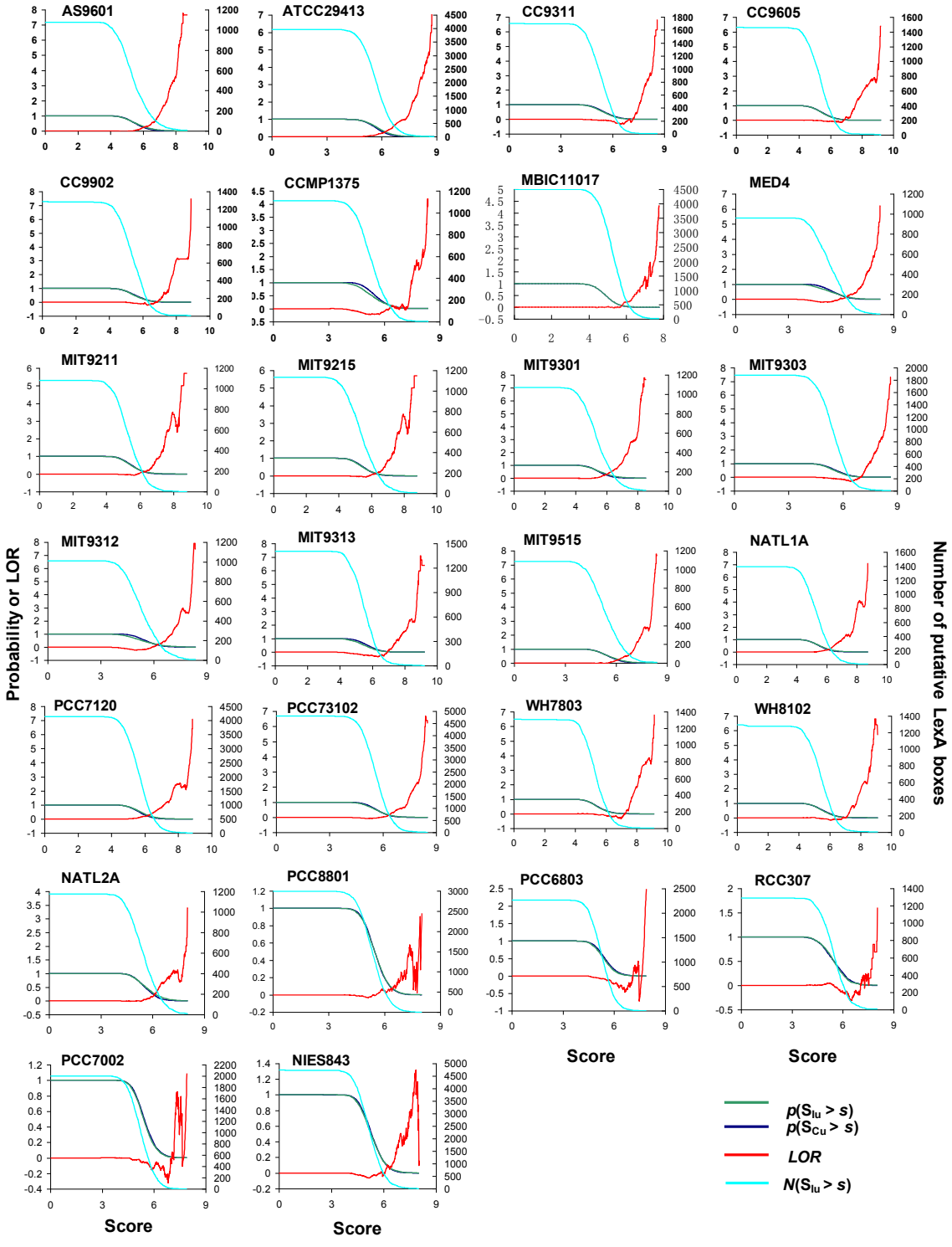


Figure 1.4. Evaluation of the predictions of LexA-binding sites in the 26 cyanobacterial genomes. The green curves represent the probability $p(S_{I_U} > s)$ and the blue curves

Figure 1.4 (continued) $p(S_{C_U} > s)$. The cyan curves are the number of iner-TU regions containing a putative binding site with a score > s, $N(S_{I_U} > s)$. The red curves are the log-odds ratio (*LOR*), defined as $LOR(s) = \ln(p(S_{I_U} > s) / p(S_{C_U} > s))$, (see Methods).

We chose randomly selected coding regions as the null model based on the assumption that a coding sequence is less likely to contain *cis*-regulatory binding sites than an intergenic sequence. Although it might be possible for genuine LexA boxes to occur in coding regions [47, 52], such kind of binding sites should be rare. The *LOR* function for a genome evaluates the ratio of the fraction of the inter-TU sequences containing a binding site with a score higher than *s* to the fraction of the randomly selected coding sequences containing a binding site with a score higher than the same *s* in the genome. Accordingly, positive *LOR* values that increase monotonically with the increase in binding site sores would suggest that an inter-TU sequence is more likely to contain a high-scoring LexA-binding site than does a randomly selected coding sequence in the genome.

As shown in Figure 1.4, when the motif score *s* increases beyond some value, the *LOR* is generally high for most of the 26 cyanobacteria that harbor a *lexA* gene, therefore those genomes with high *LOR* values are likely to contain some true binding sites. Exceptions exist in five genomes, namely, *Cyanothece sp.* PCC 8801, *Synechocystis sp.* PCC6803, *Synechococcus* RCC307, *Synechococcus sp.* PCC 7002, and *Microcystis aeruginosa* NIES-843, in which the *LOR* curves oscillate around zero when binding site score *s* increases. These poor *LOR* values might suggest that there are not more high-scoring LexA-binding sites in the inter-TU regions than in the coding regions in the five genomes. The reason for this could be that our scanning algorithm rewards a binding site that is shared by orthologs in the other genomes. If a true binding site is unique to a genome,

then it will not score high. In this sense, LexA is probably no longer a major SOS response regulator in these genomes. Instead, it might have become a specific local regulator during the course of evolution to adapt to their unique living environments (we will return to this subject later). In the case of *Synechocystis sp.* PCC6803 , it is noted that the LexA-binding sites identified by Patterson-Fortin *et al.*[52] are totally different from those identified by Mazon *et al*. [48], and that the LexA sequence in this genome is largely divergent from those in the other genomes (Figure 1.1A). Accordingly, the LexA binding sites in this genome might differ in some way from those in the other genomes, which can be another reason for its low *LOR* values.

In contrast, as shown in Figure 1.5, the *LOR* values in the five genomes that do not harbor a *lexA* gene (*Synechococcus sp.* JA-3-3Ab A-Prime, *Synechococcus sp.* JA-2-3B'a (2-13) B-Prime, *Synechococcus elongatus* PCC 6301, *Thermosynechococcus elongates* BP-1 and *Trichodesmium erythraeum* IMS101) oscillate around or decrease below zero when the motif score *s* increases beyond a certain value, implying that the chance to find a relative high-scoring putative LexA-binding site in an inter-TU region is not higher than in a randomly chosen coding sequence, suggesting that these genomes are unlikely to contain functional LexA-binding sites. On the other hand, the *LOR* values in the three genomes *Synechococcus sp.* JA-3-3Ab A-Prime, *Synechococcus sp.* JA-2-3B'a (2-13) B-Prime and *Trichodesmium erythraeum* IMS101 are relatively higher than those in the other two genomes (Figure 1.5), or even could be comparable to those of the five poor-LOR-valued cyanobacteria that harbor a *lexA* gene (Figure 1.4). In fact, the numbers of predicted binding sites in the three genomes are not too small (Table S55, S56, S59 in Additional file 6), which suggests that a few putative LexA-like binding sites exist in

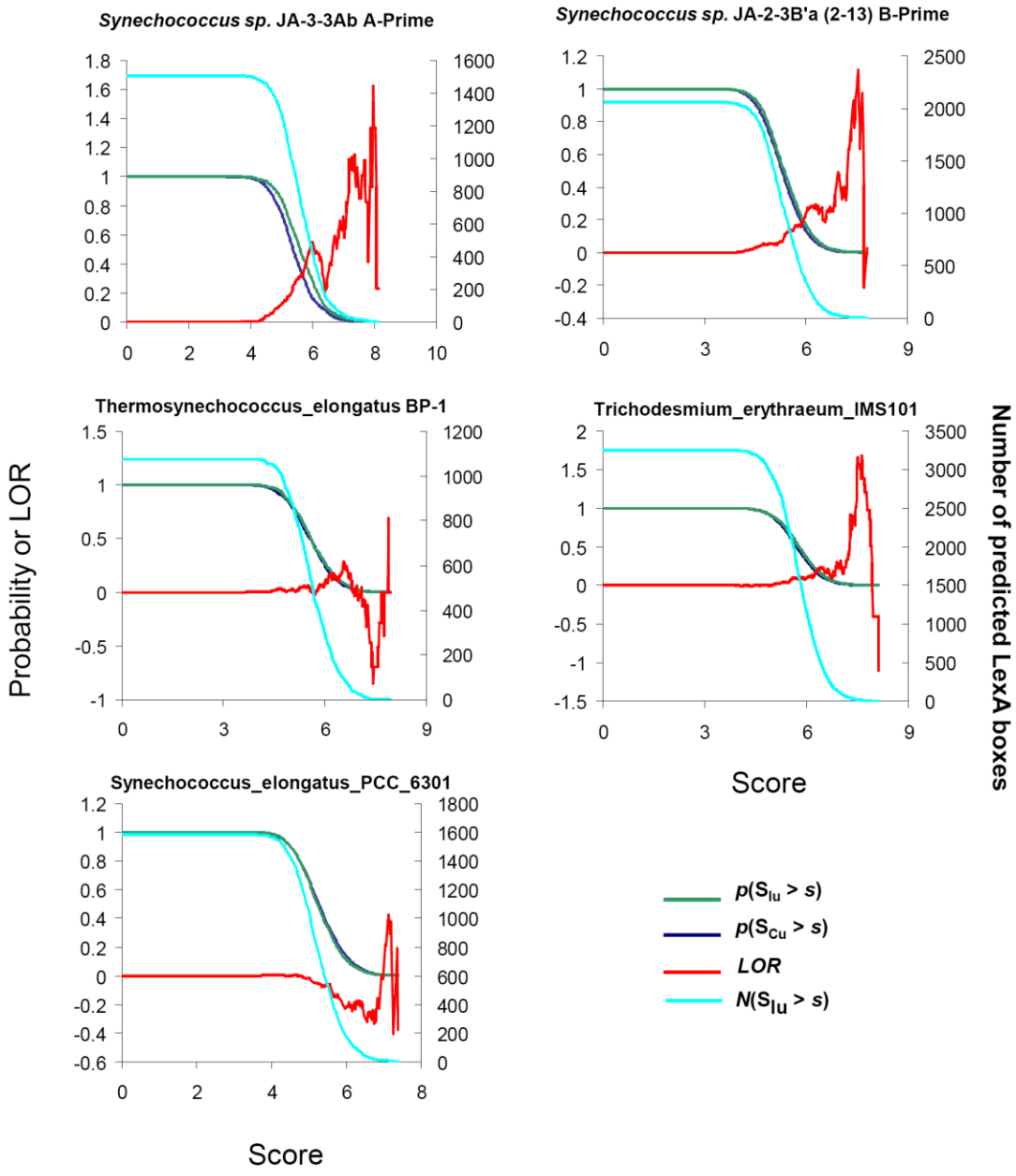


Figure 1.5 Results of genome-wide scanning for LexA-like binding sites in the five genomes that do not encode a LexA protein.

these genomes. A possible explanation for this phenomenon could be that these LexA-like binding sites are recognized by other transcription factors that have similar DNA-binding domains to that of LexA. The predictions of LexA regulons in the 26

cyanobacterial genomes that harbor a *lexA* gene are summarized in Table 1.2.

Table1.2. Summary of genome-wide LexA-binding site predictions in the 26 cyanobacterial genomes

Genome	Number of TUs	Number of genes	Score at p<0.05	LOR at p<0.05	No. of sites at p<0.05	Score at p<0.01	LOR at p<0.01	No. of sites at p<0.01
Acaryochloris_marina_MBIC11017	4507	6254	6.52	-0.143	213	7.02	0.007	48
Anabaena_variabilis_ATCC_29413	3967	5043	6.44	0.549	403	6.96	0.911	107
Cyanothece_PCC_8801	2989	4260	6.24	0.01185	169	6.73	0.22	38
Microcystis_aeruginosa_NIES_843	4736	6312	6.18	0.4941	256	6.70	0.326	73
Nostoc_punctiforme_PCC_73102	4798	6087	6.40	0.323	356	6.88	0.647	89
Nostoc_sp_PCC7120	4136	5366	6.44	0.534	389	6.88	0.995	122
Prochlorococcus_marinus_AS9601	1078	1921	6.34	0.671	107	6.74	1.330	50
Prochlorococcus_marinus_CCMP1375	1110	1883	6.37	0.070	53	6.87	-0.098	9
Prochlorococcus_marinus_MED4	961	1717	6.36	0.454	79	6.79	0.847	29
Prochlorococcus_marinus_MIT9313	1406	2269	6.63	-0.149	61	7.08	0.536	24
Prochlorococcus_marinus_MIT_9211	1081	1855	6.28	0.385	75	6.77	1.036	26
Prochlorococcus_marinus_MIT_9215	1135	1983	6.30	0.668	109	6.79	1.407	42
Prochlorococcus_marinus_MIT_9301	1070	1907	6.30	0.768	117	6.74	1.345	54
Prochlorococcus_marinus_MIT_9303	1881	2997	6.52	-0.0958	83	7.01	0.527	36
Prochlorococcus_marinus_MIT_9312	1013	1810	6.33	0.567	93	6.79	1.339	40
Prochlorococcus_marinus_MIT_9515	1088	1906	6.34	0.564	99	6.78	1.247	43
Prochlorococcus_marinus_NATL1A	1393	2193	6.30	0.495	131	6.81	1.138	43
Prochlorococcus_marinus_NATL2A	1175	1892	6.33	0.678	117	6.89	1.107	36
Synechococcus_CC9311	1700	2892	6.50	-0.153	69	7.14	0.218	19
Synechococcus_CC9605	1466	2645	6.54	-0.135	64	7.12	0.305	18
Synechococcus_CC9902	1288	2307	6.52	-0.103	63	7.00	0.786	22
Synechococcus_PCC_7002	2003	2823	6.31	-0.196	91	6.79	-0.156	20
Synechococcus_RCC307	1303	2535	6.74	-0.0904	35	7.33	0.085	8
Synechococcus_sp_WH8102	1296	2519	6.62	-0.254	59	7.18	0.142	22
Synechococcus_WH_7803	1303	2535	6.56	-0.278	56	7.36	0.623	16
Synechocystis_PCC6803	1312	2533	6.52	-0.388	79	7.00	-0.256	19

1.3.4 Conservation and diversity of the putative LexA regulons in cyanobacteria

To investigate how well the predicted LexA regulons in the 26 cyanobacterial genomes are conserved, we constructed a LexA regulon conservation tree based on the pairwise comparison of the predicted LexA regulons in these genomes (see Materials and Methods). As shown in Figure 1.6, these genomes are divided into two groups. Interestingly, one group is exclusively comprised of marine strains, and the other group

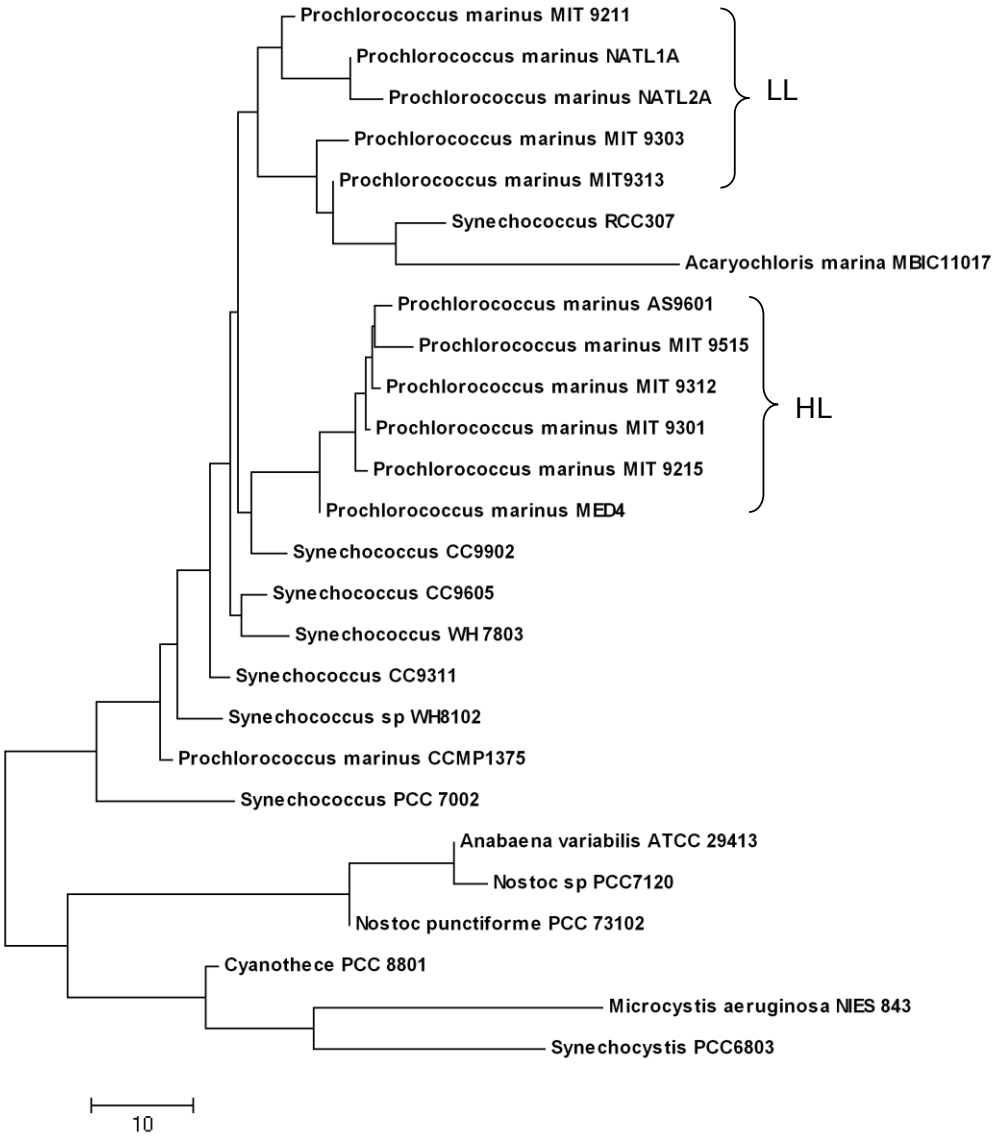


Figure 1.6. Conservation relationships among the predicted LexA regulons in the 26 cyanobacterial genomes. The tree is based on the pairwise conservation of the predicted LexA regulons in the 26 cyanobacterial genomes (see Methods).

contains the remaining genomes isolated from different non-marine habitats. In the former group, high light (HL) adapted and low light (LL) adapted ecotypes are largely grouped into two sub-groups. The results suggest that the composition of LexA regulons is dependent on the habitat of the organisms to a large extent. The general topology of the

tree (Figure 1.6) is basically consistent with both the 16S rRNA gene tree (Figure 1.7) and the LexA protein tree of these genomes (Figure 1.1A). Furthermore, both the HL and

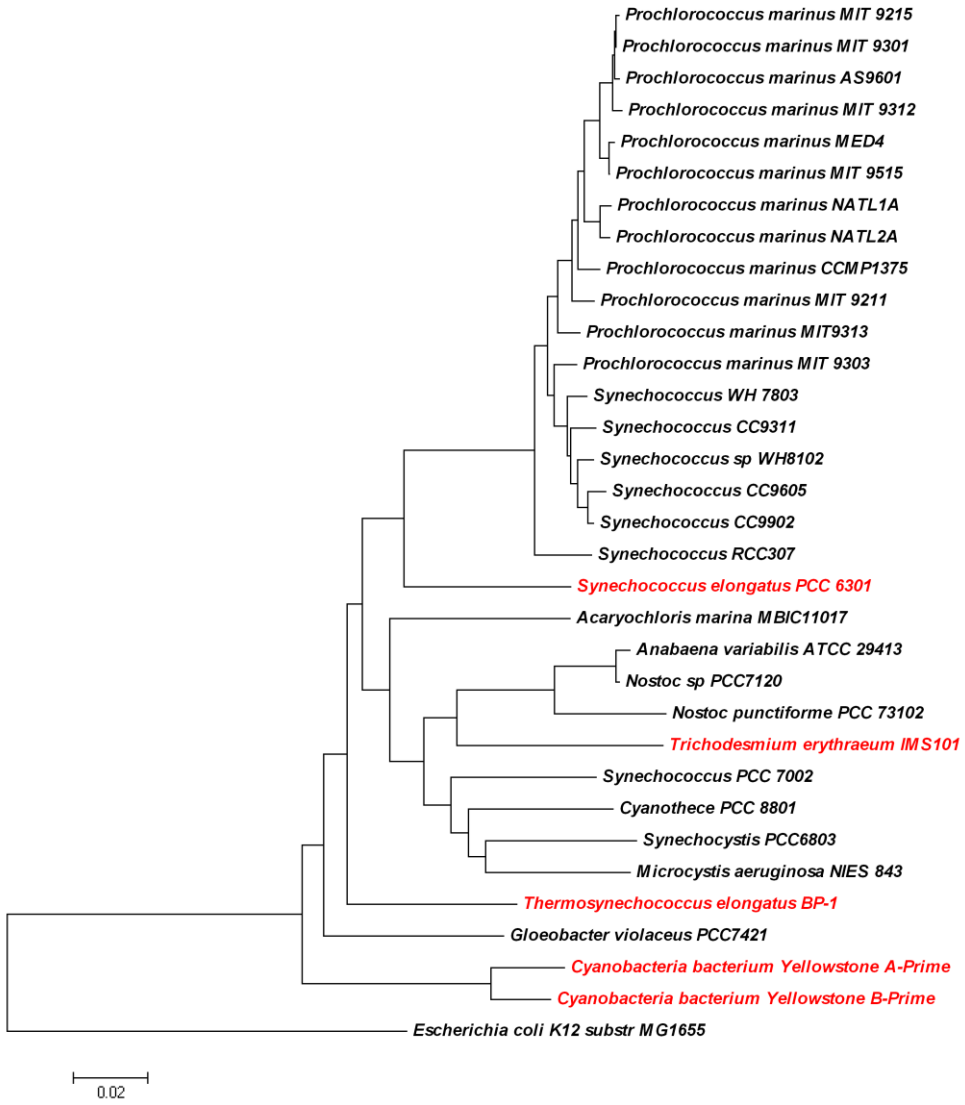


Figure 1.7. Phylogenetic relationships of 32 cyanobacterial genomes based on the 16S rRNA genes. The tree is rooted with the 16S rRNA gene of *E. coli* K12. Bootstrap values are shown on the nodes. Cyanobacterial genomes that do not encode a *lexA* gene are shown in red.

LL adapted marine sub-groups are very compact, indicating that the predicted LexA regulons in both sub-groups are relatively conserved. In contrast, the species in the non-

marine habitats are not so close to one another (Figure 1.6), suggesting that the putative LexA regulons in these genomes share few genes with one another except for the closely related *Anabaena variabilis* ATCC 29413 and *Nostoc sp.* PCC7120. The tree also indicates that *Microcystis aeruginosa* NIES 843 and *Synechocystis sp.* PCC6803 have the most distinct LexA regulons from other cyanobacterial genomes (Figure 1.6).

1.3.5 Functional classification of putative LexA regulons in cyanobacteria

Predicted members of LexA regulons in the 26 cyanobacteria that harbor a *lexA* gene are listed in Tables S1-S26 in Additional file 2, their functions can be summarized as follows.

1). SOS response system

As shown in Table 1.3, all the 33 cyanobacterial genomes included in this study encode a few SOS response genes found in *E. coli*. Several of the SOS genes in some of the 26 genomes that harbor a *lexA* gene bear a high-scoring putative LexA-binding site in their regulatory regions (Table S1-26 in Additional file 1). In particular, two of the core SOS response genes [37, 38], namely, *recA* and *lexA*, are among the most conserved putative LexA targets in the 26 cyanobacterial species/strains (Table S64 in Additional file 6). In addition, the *umuC* and *umuD* genes encoded in 13 genomes are also predicted to bear a putative LexA-binding site in their promoter regions (Table S64 in Additional file 6, Table 3 and Additional file 4). These results suggest that as in *E. coli*, the SOS response in most cyanobacteria might still be regulated by LexA. However, the other SOS response genes were found to bear a putative LexA-binding site only in a few genomes (Table 1.3). For instance, a high-scoring LexA-binding site for the *ruvB* gene

Table 1.3: Putative LexA regulon members involved in various biological processes.

Genomes	SOS	Photo-synthesis	Transporters
Acaryochloris marina_MBIC11017	lexA recA dnaK groEL umuCD		4624
Anabaena_variabilis_ATCC_29413	lexA recA dnaJ sbcC	psbA	4997 4148 4995
Cyanothece_PCC_8801	lexA recA		
Microcystis_aeruginosa_NIES_843	recA ssb	ndhH ycf4	pstB2
Nostoc_punctiforme_PCC_73102	lexA recA sbcC F4123		F3763
Nostoc_sp PCC7120	lexA recA uvrA uvrB dnaKJ		alr5147
Prochlorococcus_marinus_AS9601	recA ruvB umuCD	psbY	11511
Prochlorococcus_marinus_CCMP1375	recA sbcD groES groEL		
Prochlorococcus_marinus_MED4	recA umuCD ruvB	psbY	
Prochlorococcus_marinus_MIT9313	lexA umuCD		
Prochlorococcus_marinus_MIT_9211	recA umuCD		
Prochlorococcus_marinus_MIT_9215	recA umuCD ruvB	psbY	08441
Prochlorococcus_marinus_MIT_9301	recA umuCD	psbY	11521 02331
Prochlorococcus_marinus_MIT_9303	lexA umuCD		21241 15661
Prochlorococcus_marinus_MIT_9312	recA ruvB umuCD uvrD	psaA psbY	0561
Prochlorococcus_marinus_MIT_9515	recA ruvB dnaK	psbY	06251
Prochlorococcus_marinus_NATL1A	lexA recA		
Prochlorococcus_marinus_NATL2A	lexA recA	psaM	
Synechococcus_CC9311	recA umuCD		2443
Synechococcus_CC9605	recA umuCD		2635
Synechococcus_CC9902	recA umuCD		0850
Synechococcus_PCC7002	recA	psaF	
Synechococcus_RCC307	lexA		
Synechococcus_sp_WH8102	recA umuCD ruvC		2111 0959
Synechococcus_WH7803	recA	ndhH	
Synechocystis_PCC6803		psbB	0467

encoding Holliday junction DNA helicase B was found only in HL adapted *Prochlorococcus* ecotypes MIT9312, MIT9515, MIT9215, MED4 and AS9601. Moreover, in the case of the nucleotide excision repair (NER) genes *uvrA, B, C and D,* which are under the regulation of LexA in *E. coli* [42], we were able to identify putative LexA-binding sites only in the promoter regions of the *uvrA* and *uvrB* in *Nostoc sp.* PCC7120 and the promoter region of *uvrD* in *Prochlorococcus marinus* MIT 9312 (Table 1.3). Thus, it is likely that the NER process in the remaining genomes is regulated by some transcription factor other than LexA, given that *uvr* genes are present in all the 32 cyanobacterial genomes analyzed in this study, including those that do not encode a *lexA*

gene (Table S63 in Additional file 6). These results are consistent with the earlier observation that LexA target genes in bacteria are highly diversified in order for them to adapt to different ecological niches [37, 38].

On the other hand, in *Synechococcus* PCC7002, *Synechococcus* RCC307 and *Synechococcus* WH7803, LexA boxes were only detected for one of the core SOS response genes, i.e., SYNPCC7002_A0426 (*recA*) in *Synechococcus* PCC7002, SynRCC307_1756 (*lexA*) in *Synechococcus* RCC307 and SynWH7803_0439 (*recA*) in *Synechococcus* WH7803, although these genomes all encode a *lexA* gene and other core SOS response genes, such as *recA* (SYNPCC7002_A0426, SynRCC307_2111 and SynWH7803_0439,) and *ruvB* (SYNPCC7002_A1390, SynRCC307_1756 and SynWH7803_0185), *umuC* (SynRCC307_0043 and SynWH7803_1080,) and *umuD* (SynRCC307_0042 and SynWH7803_1081). Since only one single SOS response gene bears a putative LexA box in these genomes, it is likely that the role of LexA in the regulation of the SOS response in these genomes might have been attenuated. The case of *Synechocystis sp.* PCC6803 seems to go even further in this direction as detailed below.

As indicated previously [48, 50], the LexA protein of *Synechocystis sp.* PCC6803 is unusual in two aspects compared to those in the other genomes analyzed in this study. First, as shown in Figure 1.8, the Ala-Gly dyad in the N-terminus of LexA responsible for auto-cleavage of the protein in all other cyanobacteria as well as in *E. coli* and *B. subtilis* [48] are replaced by Gly-Gly in *Synechocystis sp.* PCC6803. Second, the reactive residue Ser (Ser^{119} of LexA in *E. coli*) that attacks the Ala-Gly peptide bond is replaced by Asp of LexA in *Synechocystis sp.* PCC6803 [48, 50]. It has been shown that SOS induction cannot be initiated by a non-cleavable LexA repressor [42, 48]. Therefore, it is highly

likely that LexA in *Synechocystis sp.* PCC6803 cannot undergo the auto-cleavage

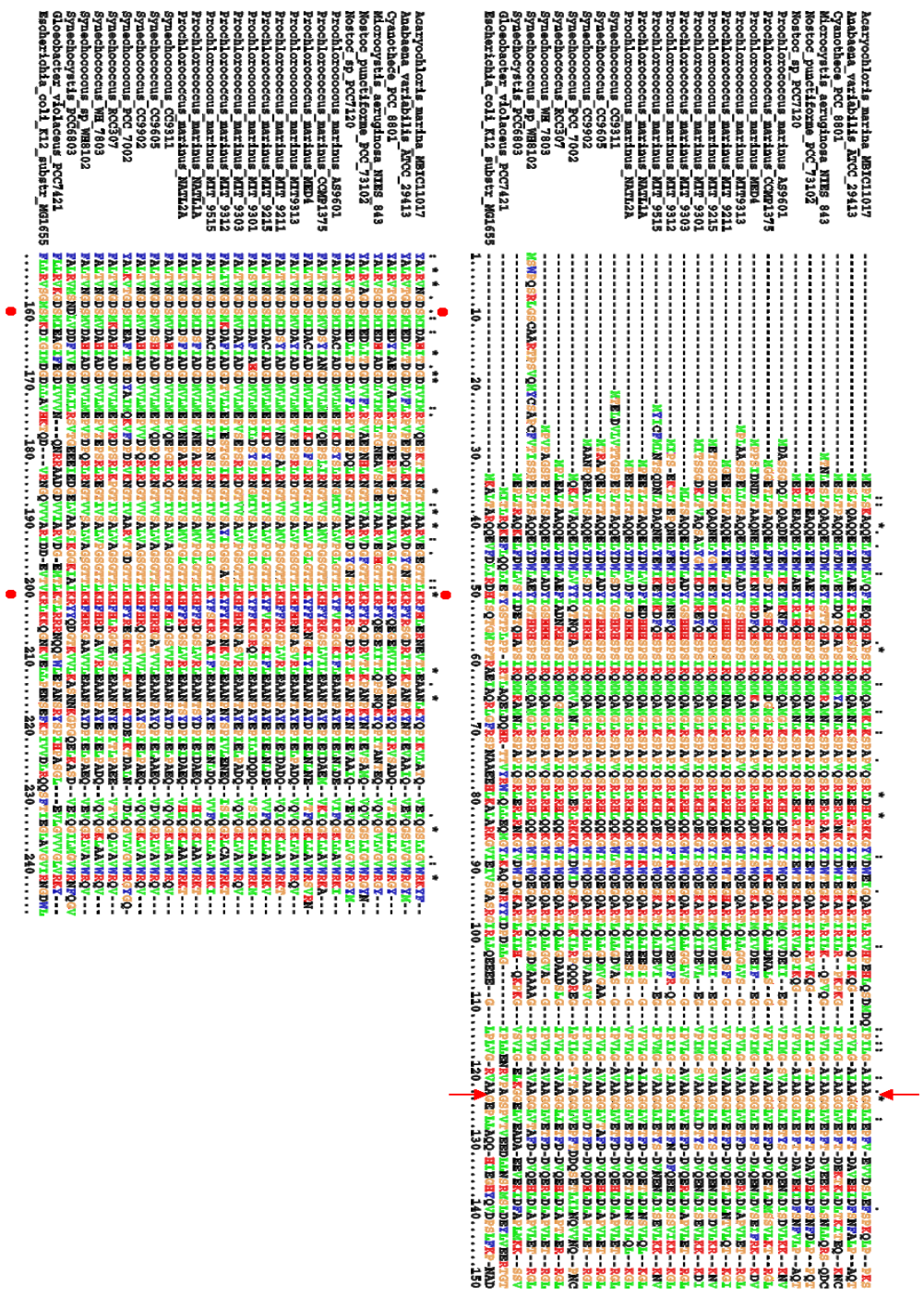


Figure 1.8. Multiple sequence alignments of the full-length LexA in the 27 cyanobacterial genomes. The sequence from *E.coli* LexA is also included for comparison. The auto-cleavage sites are indicated by a red arrow, and the reactive residues are indicated by red dots.

reaction in response to DNA damage, and it might have adopted a different function other than the canonical SOS response regulator seen in *E. coli* and *B. subtilis* [68]. This argument is consistent with the observation that *Synechocystis sp.* PCC6803 has a notably larger branch length in the 27 LexA protein tree (Figure1.1A), but this is not seen in the 16S rRNA gene tree (Figure 1.7).

Although the *Synechocystis sp.* PCC6803 genome harbors some core SOS response genes including *lexA* and *recA* (Table S63 in Additional file 6), none of them belongs to our predicted LexA regulon at a *p*-value < 0.01 (Table S26 in Additional files 2 and Table 1.3). The *mutS* (sll1772) gene is the only gene that is likely to be in involved in DNA mismatch repair, while bearing a putative LexA binding site in the genome. However, the orthologs of *mutS* is not under the regulation of LexA in *E. coli* [34, 69] or within the putative LexA regulon of any other cyanobacteria (Table S1-26 in Additional files 2). These results suggest that at least most of SOS response genes are not under the regulation of LexA in *Synechocystis sp.* PCC6803. Indeed, using microarray gene expression profiling in response to *lexA* depletion, Domain *et al.*[53] concluded that LexA in *Synechocystis sp.* PCC6803 might be involved in carbon metabolism or controlled by carbon availability rather than the regulation of SOS response. However, our predicted LexA regulon in *Synechocystissp.* PCC6803 (Table S26 in Additional files 2) has no intersection with the LexA-responsive genes identified by Domain *et al.*[53]. Since the LexA-binding sites that were experimentally characterized [49, 52] in *Synechocystis sp.* PCC6803 are different from the sequences in our scanning profile, and considering the distinct nature of the LexA protein in *Synechocystis sp.* PCC6803 indicated above, it would be particular interesting to determine by experiment the

function of the predicted sites in this genome.

Thus, although *Synechocystis sp.* PCC6803 clearly harbors the components of a basic SOS response system (Table 1.3), it is probably no longer under the regulation of LexA. Accordingly, LexA in this genome might have adopted a different function. Thus, the loss of the original function of LexA in *Synechocystis sp.* PCC6803 is coupled with the loss of the sequence constraint, thereby accelerating its divergence from other cyanobacterial LexA proteins, at both the sequence and functional levels. On the other hand, given the importance of the SOS response in cell survival, it is highly likely that the transcriptional regulator of the SOS response system in *Synechocystissp.* PCC6803 has been replaced by another protein.

2). Other cellular processes

Interestingly, we also found putative LexA-binding sites in the regulatory regions of genes that participate in various cellular processes in these 26 cyanobacterial genomes (Table 1.3). The major cellular processes that are likely under the regulation of LexA are summarized below.

2.1) Photosynthesis

Putative LexA-binding sites were predicted for the following photosynthetic genes in the 26 cyanobacteria that harbor a *lexA* gene with $p<0.01$(Table 1.3, Table S1-26 in Additional file 4): Ava_3553, A9601_12231, PMM1117, P9215_12531, P9301_12241, PMT9312_1128, and P9515_12081, coding for a photosystem II reaction center protein PsbY; slr0906, coding for the photosystem II CP47 protein; and MAE_44810, PMT9312_1615, PMN2A_1682a and SYNPCC7002_A1008, coding for a protein involved in photosystem I. These results suggest that the SOS response system might

have cross-talk with photosynthesis in those genomes.

2.2). Transporters

Around 20 genes encoding transporters were predicted to bear a putative LexA box (Table 1.3). Most of them belong to the ABC transporter proteins, including AM1_4624 in *Acaryochloris marina* MBIC11017, Ava_4995 in *Anabaena variabilis* ATCC29413, MAE_18340 in *Microcystis aeruginosa* NIES843, Npun_F3763 in *Nostoc punctiforme* PCC73102, alr5147 in *Nostoc sp* PCC7120, P9215_08441 in *Prochlorococcus marinus* MIT9215, P9303_15661 in *Prochlorococcus marinus* MIT9303, P9515_06251 in *Prochlorococcus marinus* MIT9515, SYNW2111 in *Synechococcus sp* WH8102 and slr0467 in *Synechocystis sp.* PCC6803. In addition, several toxin and antibiotics exporters were identified to have a putative LexA-binding site in their regulatory regions, including cadmium resistance transporter Ava_4997 in *Anabaena variabilis* ATCC29413; MFS (major facilitator superfamily) multidrug efflux transporter P9301_11521 in *Prochlorococcus marinus* MIT9301 and A9601_11511 in *Prochlorococcus marinus* AS9601; multidrug efflux ABC transporter P9515_06251 in *Prochlorococcus marinus* MIT9515 and SYNW0959 in *Synechococcus sp* WH8102; putative ABC transporter/multidrug efflux family protein SYNW2111 in *Synechococcus sp* WH8102; drug exporter-1 ABC transporter ATPase subunit AM1_4624 in *Acaryochloris marina* MBIC11017. These findings are interesting since it has been shown that the SOS response system is related to drug resistance in *E. coli* [70, 71] and *Staphylococcus aureus* [71-73] by mechanisms that are not fully understood. It was reported that the *vP2449* gene encoding a toxin exporter responsible for xenobiotic resistance in *Vibrionales parahaemolyticus* was under the direct control of LexA[74]. Therefore, it is

likely that these drug resistance genes are regulated by LexA, thereby coupling the SOS response to drug resistance in these cyanobacteria.

1.3.6 The origin of the *lexA* gene in cyanobacteria

As indicated earlier, 27 of the 32 cyanobacterial genomes analyzed evidently harbor a *lexA* ortholog, while the remaining five genomes do not, even when being scrutinized by more sensitive sequence search methods such as PSI-BLAST (data not shown). The five cyanobacteria lacking a LexA are *Synechococcus sp.* JA-3-3Ab A-Prime, *Synechococcus sp.* JA-2-3B'a(2-13) B-Prime, *Synechococcus elongatus* PCC6301, *Trichodesmium erythraeum* IMS101 and *Thermosynechococcus elongatus* BP-1. However, the core SOS response genes remain in these five genomes (Table 1.3). In the tree of 183 detected LexA proteins in 598 sequenced genomes (Figure 1.3), the 26 cyanobacterial LexA proteins that are detected by BDBH (see Materials and Methods) form a monophyletic group while LexA in *Gloeobacter violaceus* PCC7421 is clustered with the group of α-proteobacteria. Furthermore, the topology of the 16S rRNA gene tree (Figure 1.7) and the LexA tree/subtree of 27 cyanobacterial genomes (Figure 1.1A and Figure 1.3) are quite similar. This result suggests that *lexA* in the 26 cyanobacterial genomes (excluding *Gloeobacter violaceus* PCC7421) is likely to be vertically inherited from the last common ancestor of cyanobacteria. However, *Gloeobacter violaceus* PCC7421 might have lost its LexA protein during evolution and obtained an ortholog later through horizontal transfer from an α-proteobacterium. The five genomes that lack a *lexA* gene do not form a monophyletic group in the 16S rRNA gene-based phylogenetic tree of these 32 cyanobacteria (Figure 1.7). In particular, *Synechococcus elongatus* PCC6301, and *Trichodesmium erythraeum* IMS101 are spread in a clade whose members except these

two genomes harbor a *lexA* gene. The most parsimonious explanation of this distribution would be that these two genomes *Synechococcus elongatus* PCC6301 and *Trichodesmium erythraeum* IMS101 lost their *lexA* genes through two independent events (one for each genome) to adapt to their corresponding environments during the course of evolution. Furthermore, the remaining three genomes, *Thermosynechococcus elongatus* BP-1, *Synechococcus sp. JA-3-3Ab* A-prime and *Synechococcus sp. JA-2-3B'a (2-13)* B-prime, which do not possess a *lexA* gene, branch earlier from the others (Figure 1.7). A plausible explanation of this distribution would be that these genomes lost their *lexA* genes inherited from the last common ancestor of cyanobacteria during the course of evolution. Interestingly, all these three genomes are thermophilic, their extreme ecological niches might facilitate the loss of the *lexA* gene. Since the core SOS response genes remain in these five genomes (Table S63 in Additional file 6) after *lexA* was lost, they might have been hijacked by another transcription factor given the importance of the regulation of the SOS response genes for cell survival. The genomes that lost their *lexA* gene appear to have lost LexA-binding sites (Figure 1.5). Alternatively, these five cyanobacteria might still harbor a *lexA* gene that has largely diverged from the others' during evolution to such a level that our method could not detect them

In addition, it has been suggested that the *lexA* gene was derived from gram-positive bacteria, which then spread into cyanobacteria and fibrobacteres. Then α-proteobacteria acquired *lexA* from cyanobacteria [37, 38, 48]. Our phylogenetic analysis of the LexA proteins and their binding sites supports such an argument. As mentioned before, cyanobacterial LexA proteins are more closely-related to those in gram-positive bacteria and α-proteobacteria than those in the other groups (Figure 1.3), and the predicted LexA-

binding sites in cyanobacteria are clustered together with those in the gram-positive bacterium *B. subtilis* and in α-proteobacteria, but are far away from those in *E.coli* (Figure 1.2).

Moreover, Erill *et al*.[59] have suggested that there is a common set of genes in the LexA regulon of proteobacteria and gram-positive bacteria: *recA, uvrA, ssb,* and *ruvC*. However, our predicted LexA regulons in cyanobacteria do not always include this set of genes. Thus, the concept of a common set of SOS response gene in its more general form warrants further scrutinization.

1.4 Methods

1.4.1 Materials

The sequences and annotation files of 33 sequenced cyanobacterial and the other genomes were downloaded from NCBI at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. The cyanbacterial genomes used in this study include: *Acaryochloris marina* MBIC11017 (MBIC11017), *Anabaena variabilis* ATCC 29413 (ATCC29413), *Synechococcus sp.* JA-3-3Ab (A-Prime), *Synechococcus sp.* JA-2-3B'a (2-13) (B-Prime), *Cyanothece sp.* PCC 8801(PCC8801), *Gloeobacter violaceus* PCC7421 (PCC7421), *Microcystis aeruginosa* NIES 843 (NIES843), *Nostoc punctiforme* PCC 73102 (PCC73102), *Nostoc sp.* (PCC7120), *Prochlorococcus marinus* AS9601 (AS9601), *Prochlorococcus marinus* CCMP1375 (CCMP1375), *Prochlorococcus marinus* MED4 (MED4), *Prochlorococcus marinus* MIT9313 (MIT9313), *Prochlorococcus marinus* MIT 9211 (MIT9211), *Prochlorococcus marinus* MIT 9215 (MIT9215), *Prochlorococcus marinus* MIT 9301 (MIT9301), *Prochlorococcus marinus* MIT 9303 (MIT9303), *Prochlorococcus marinus* MIT 9312 (MIT9312), *Prochlorococcus marinus* MIT 9515 (MIT9515), *Prochlorococcus*

marinus NATL1A (NATL1A), *Prochlorococcus marinus* NATL2A (NATL2A), *Synechococcus sp.* CC9311 (CC9311), *Synechococcus sp.* CC9605 (CC9605), *Synechococcus sp.* CC9902 (CC9902), *Synechococcus sp.* PCC 7002 (PCC7002), *Synechococcus sp.* RCC307 (RCC307), *Synechococcus* WH 7803 (WH7803), *Synechococcus elongatus* PCC 6301 (PCC6301), *Synechococcus sp.* WH8102 (WH8102), *Synechocystis sp.* PCC6803 (PCC6803), *Synechocystis sp.* PCC7942 (PCC7942), *Thermosynechococcus elongates* BP-1 (BP-1) and *Trichodesmium erythraeum* IMS101 (IMS101).

1.4.2 Prediction of transcription units

We predicted the operon structures in cyanobacterial genomes using the operon prediction algorithm developed by Dam et al. [75]. The algorithm is based on the integration of both genome-specific and comparative genomic information. In this work, both the multi-gene operon and singleton operon (containing one gene) are considered as a transcription unit (TU), and the upstream intergenic sequence of the first open reading frame is not considered as a part of the operon.

1.4.3 Prediction of orthologs

We used the bi-directional best hit (BDBH) method based on BLASTP searches with an *E*-value cut-off of 10^{-10} for both directions to predict orthologous protein pairs between any two proteomes. The BDBH method assumes that a cross-species protein pair are orthologous if each protein returns the other as the best hit in the whole proteome comparison [76].

1.4.4 Phylogenetic analysis

To construct the phylogenetic tree of LexA in cyanobacteria, multiple sequence

alignment of the LexA amino acid sequences from 27 cyanobacterial genomes and the *E.coli* K12 genome were made using ClustalW implemented in MEGA [63] with default settings. The phylogenetic tree was then constructed using the neighbor-joining method with Poisson correction model in MEGA. *E.coli* LexA was placed as the outgroup of the tree. To construct the species tree, the DNA sequences of 16S ribosomal RNA genes from the 32 cyanobacteria and *E.coli* were aligned using ClustalW with manual adjustment by removing the unalignable regions. A neighbor-joining tree was then constructed with *E.coli* K12 being the outgroup using the Kimura 2-parameter model. Statistical significance at each node in the trees was evaluated using 500 bootstrap resamplings.

To construct the LexA protein tree across cyanobacteria, gram-positive bacteria, α-proteobacteria, δ-proteobacteria and γ-proteobacteria and some other bacteria strains/species (Figure 1.3), we first downloaded 598 sequenced microbial genome sequences from NCBI, and then identified LexA orthologs in them by the BDBH method described above. Multiple sequence alignments of these LexA sequences were made using ClustalW implemented in MEGA[63] with default settings. The phylogenetic tree was then constructed in the same way as the 27 LexA protein tree.

The phylogenetic tree (Figure 1.2) of LexA-binding sites in cyanobacteria, *B.subtilis*, α-proteobacteria and *E.coli* K12 was generated by the STAMP web tool [61] with the default alignment parameters: Pearson correlation coefficient for column comparison metric; ungapped Smith-Waterman for pair-wise alignment. The phylogenetic tree was constructed using the UPGMA method implemented in STAMP[61].

1.4.5 Phylogenetic footprinting and construction of LexA-binding sites in cyanobacteria

The previous study by Mazon *et al*[48] characterized the LexA boxes associated with

two genes: *alr4908* (*lexA*) and *all3272* (*recA*). Four putative LexA boxes were also identified in the promoter regions of *alr3716* (*uvrA*), *alr0088* (*ssb*), *alr4905*, and *all4790* in *Nostoc sp* PCC7120 in that study. The orthologs (if they exist) of these six genes in PCC7120 were identified in the other 25 cyanobacterial genomes which harbor a *lexA* gene. We pooled the entire upstream inter-TU regions of these six genes in the target genome *Nostoc sp.* PCC7120 as well as those of the TUs containing at least one of the orthologs of these six genes in other cyanobacteria. If the length of the inter-TU region is longer than 800 bases, then only the immediate upstream 800 bases were extracted. Two motif finding programs, MEME [56, 77] and BioProspector [57], were then applied to these pooled inter-TU regions to identify palindromic 14-mers as putative LexA-binding sites in these sequences according to previous studies [48]. MEME applies an expectation maximization method to fit a two-component finite mixture model and returns the identified motifs with an E-value[77], while BioProspector employs a Gibbs sampling strategy and estimates the significance of the identified motif by a Monte Carlo method [78]. These two programs were selected as they are widely used and often have complementary predictions [79, 80]. MEME identified 45 putative LexA-binding sites with an overall E-value of 1.4e-026 for its most significant predicted motif, while BioProspector detected 39 putative LexA boxes in its most significant predicted motif (see Additional file 1 for details). High-scoring putative LexA-binding sites from either program were selected to build the LexA-binding sites profile (Table 1.1) in cyanobacteria. Sequence logos of binding sites were created using the Weblogo server [62].

1.4.6 Genome wide prediction of LexA-binding sites

We used the profile constructed above to scan the inter-TU regions of the genomes to predict all putative LexA-binding sites using a scanning algorithm that we previously developed [65-67]. This algorithm is briefly described as follows.

For each predicted TU $U(g_1,g_2,...,g_n)$ composed of genes $g_1,g_2,...,g_n$ in genome G, we extracted its upstream inter-TU regions and the first 40 bases of coding region(if its length is longer than 800 bases, then only the immediate upstream 800 bases were extracted), denoted as $I_{U(g1,g2,...,gn)}$. The set of all the $I_{U(g1,g2,...,gn)}$ in this genome is denoted as I_U. To find the best matching substring in a sequence t in I_U ($t \in I_U$) when scanned by profile M, we use the following scoring function:

$$s_M(t) = \max_{h \subset t} \sum_{i=1}^{l} I_i \ln \frac{p(i,h(i))}{q(h(i))}, \qquad (1.1)$$

$$I_i = \Big(\sum_{b \in \{A,C,G,T\}} p(i,b) \ln \frac{p(i,b)}{q(b)} \Big) / a, \qquad (1.2)$$

$$a = \frac{n+1}{n+4} \ln(n+1) - \ln(n+4) - \frac{1}{n+4} \sum_{b \in \{A,C,G,T\}} \ln q(b) - \frac{n}{n+4} \ln \min_{b \in \{A,C,G,T\}} q(b), \qquad (1.3)$$

where l is the length of the binding sites of profile M, h any substring of sequence t with length l (i.e. each l-mer of the sequence t), $h(i)$ the base at the i-th position of h, $p(i,b)$ the frequency of base b at position i in M, $q(b)$ is the frequency of base b in the aggregated inter-TU regions for the organism, I_i is basically the information content or the relative entropy of the column [64, 81] divided by a normalization factor a, a is the upper limit of the information content I_i for this column to keep $I_i \in [0, 1]$, n the number of binding sites for constructing the profile M. To avoid zero value of the numerator $p(i,b)$, a pseudo

count 1 is added to the counts of the each base {A, C, G, T} in column *i*.

To show the derivation of the normalization factor *a*, we considered the extreme case: for a column *i* of profile M containing *n* binding sites, the more conserved the column is, the higher its information content I_i will be, and the maximum information content for column *i* occurs when this column is completely homogeneous. That is, all sequences have the same nucleotide, say, A at that position, and this nucleotide has the smallest background frequency, *q(A)*, noted as q_0. Thus, after adding one pseudocount to the counts of each of the four nucleotides to column *i*, the frequency of base A of column *i* in the profile will therefore be (n+1)/(n+4), and 1/(n+4) for the other three nucleotides. Then the upper limit *a* of the prenormalized I_i as shown by formula (1.3) can be derived as follows.

$$I_i = \sum_{b\in\{A,C,G,T\}} p(i,b)\ln\frac{p(i,b)}{q(b)}$$

$$\le \frac{n+1}{n+4}\ln\frac{n+1}{(n+4)q_0}$$ The term for the most conserved nucleotide A

$$+\sum_{b\in\{A,C,G,T\}}\frac{1}{n+4}\ln\frac{1}{(n+4)\cdot q(b)}$$ The terms for the other three nucleotides and a duplicated term for A

$$-\frac{1}{n+4}\ln\frac{1}{(n+4)q_0}$$ Subtracting out the duplicated term for A

$$=\frac{n+1}{n+4}\ln(n+1)-\frac{n+1}{n+4}\ln(n+4)-\frac{n+1}{n+4}\ln q_0$$

$$+\frac{1}{n+4}(-4\ln(n+4)-\sum_{b=A}^{T}\ln q(b))$$

$$+\frac{\ln(n+4)}{n+4}+\frac{\ln q_0}{n+4}$$

$$= -\frac{4}{n+4}\ln(n+4) - \frac{1}{n+4}\sum_{b=A}^{T}\ln q(b) - \frac{n}{n+4}\ln(n+4) - \frac{n}{n+4}\ln q_0 + \frac{n+1}{n+4}\ln(n+1)$$

$$= -\ln(n+4) - \frac{1}{n+4}\sum_{b=A}^{T}\ln q(b) - \frac{n}{n+4}\ln q_0 + \frac{n+1}{n+4}\ln(n+1) = a \qquad (1.4)$$

where $q_0 = q(A) = \min_{b\in\{A,C,G,T\}} q(b)$.

Intuitively, we slide a window of length *l* across sequence *t* with the profile *M*, and return the substring *h* with the highest score defined by (1.1).

Since true regulatory binding sites are likely to be more conserved than other inter-TU sequences and thus tend to be shared by closely related orthologous genes. For each genome (considered as a target genome), we reward its putative binding sites appeared to be conserved in regions upstream from orthologous genes in other genomes. To do this, we assume a transcription unit $U(g_1,g_2,...,g_n)$ in the target genome *G* is composed of n genes. Gene g_i ($i = 1\ldots n$) has orthologs in m_i genomes G_1, G_2, …, G_{m_i}, and $o_k(g_i)$ is the upstream inter-TU sequence associated with the orthologous gene g_i in genome G_k (for a graphic explanation, see Figure 1.9). Then the $s_M(t)$ score for the inter-TU sequence *t* upstream from $U(g_1,g_2,...,g_n)$ in genome G can be increased by a term $A_{max}(g_i)$:

$$s(t) = s_M(t) + A_{\max}(g_i) \qquad (1.5)$$

where $A_{max}(g_i)$ is the value calculated for gene g_i whose orthologs across other genomes have the maximum average of the product of two terms:

$A_{\max}(g_i) = \max_{1\le i\le n}$ {average[(similarity between the two sites)*(score of this orthologous site)]} $= \max_{1\le i\le n}\{average[\frac{l-d_{i,k}}{l}s_M(o_k(g_i))]\}$ (1.6)

where $d_{i,k}$ is the Hamming distance between the sequence h detected by the profile M in sequence t and the corresponding sequence in $o_k(g_i)$, and l is the length of the binding sites in profile M.

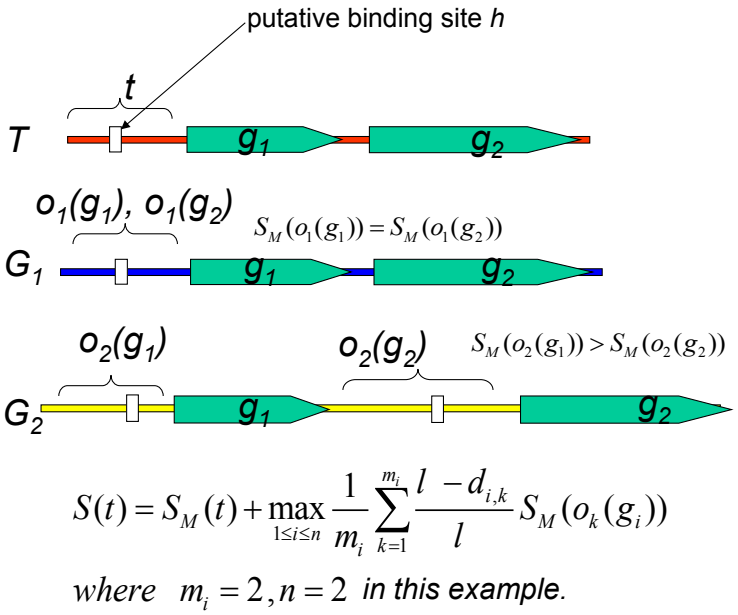


Figure 1.9. An example for explaining the algorithm. For $o_1(g_1)$, $o_2(g_1)$ and $o_2(g_2)$, assume they have the same value of sequence similarity to h,i.e., $(l - d_{i,k})/l$. Then we should select orthologs of g_1 in this case, i.e., i =1, and add its average score across genome G_1 and G_2 to $S_M(t)$.

Since the orthologs of genes of a transcription unit in one organism may not comprise a single transcription unit in another. For $U(g_1,g_2,...,g_n)$ in target genome G, orthologs of $g_1,g_2,...,g_n$ may be separated into different TUs in other genomes, therefore we evaluated the orthologous inter-TU sequences for each gene in $(g_1,g_2,...,g_n)$, and chose the gene g_i whose orthologs across other genomes have the maximum average of the product of two terms indicated above. Then by combining formula (1.5) and (1.6), the refined score of the best putative binding site in sequence t can be defined as:

$$s(t) = s_M(t) + \max_{1 \le i \le n} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{l - d_{i,k}}{l} s_M(o_k(g_i)), \qquad (1.7)$$

1.4.7 Statistical significance of predicted binding sites

To evaluate the extent to which a putative binding site with a score *s* or higher can be found purely by chance, we randomly extracted coding sequence with the same length as $I_{U(g1,g2,...,gn)}$, denoted as $C_{U(g1,g2,...,gn)}$. All the $C_{U(g1,g2,...,gn)}$ extracted in genome G form the set C_U. The score of an extracted sequence t ($t \in C_U$) scanned by a profile M is also defined by formula (1). Note that each randomly chosen $C_{U(g1,g2,...,gn)}$ has nothing to do with $U(g_1,g_2,...,g_n)$. Therefore, when incorporating the additional score from reference genomes (formula (1.7)), the coding sequence $o_k(g_i)$ is unlikely the coding sequence associated with the orthologous genes of $g_1,g_2,...,g_n$ in a reference genome G_k as it is a randomly chosen one. To avoid possible biased sampling of coding sequences for each $I_{U(g1,g2,...,gn)}$ in I_U, we randomly extracted 300 coding sequences $C_{U(g1,g2,...,gn)}$ sharing the same length as $I_{U(g1,g2,...,gn)}$. These randomly chosen coding regions for all the $U(g_1,g_2,...,g_n)$ in genome G form a sequence set C_U, then each sequence in the set C_U was scanned using formula (1.7). Let $S(I_U)$ and $S(C_U)$ be the set of scores of binding sites found in I_U and C_U, respectively, and $P(S(t) > s)$ be the cumulative probability of finding a binding site in a sequence t ($t \in I_U$ or $t \in C_U$) with a score $S(t) > s$ as defined by equation (1.7). Next, the false positive rate, $p(S_{C_U} > s)$ can be used to evaluate the statistical significance of the motif score *s* of a inter-TU sequence. $p(S_{C_U} > s)$ is actually the fraction of coding sequences bearing a putative binding site with a score higher than *s* in the coding sequences set C_U in genome G. In other words, it describes the extent to

which one can find a motif with a score higher than *s* by chance. Thus, it can be considered as an empirical p-value for a binding site score *s*. A cut-off score *s* corresponding to a *p*-value < 0.01 is used for the LexA-binding site and regulon prediction in each genome in this study.

To evaluate the confidence of our overall predictions in inter-TU regions in one genome, we used a log odds ratio (*LOR*) to compare the probability of finding a putative binding site in an inter-TU region with the probability of finding a putative binding site in a randomly extracted coding region by considering all the extracted $I_U s$ and $C_U s$ in a genome. We estimated the statistical significance of the predictions using the *LOR* function defined as

$$LOR(s) = \ln \frac{p(S_{I_U} > s)}{p(S_{C_U} > s)}. \qquad (1.8)$$

The *LOR* function for a genome is the log-odds ratio of the fraction of the inter-TU sequences containing a binding site with a score higher than s to the fraction of the randomly selected coding sequences containing a binding site with a score higher than the same *s* in the genome. Accordingly, a monotonic increase in positive *LOR* with the increase in the motif score in a genome would suggest that this genome is likely to contain some high-scoring LexA-binding sites.

1.4.8 Analysis of the conservation of LexA regulons in cyanobacteria

We defined the conservation (c_{ij}) between two regulons R_i and R_j from genome *i* and *j*, respectively, as,

$$c_{ij} = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} = \frac{|R_i \cap R_j|}{|R_i| + |R_j| - |R_i \cap R_j|} \qquad (1.9)$$

Where $|R_i \cap R_j|$ is the number of orthologous genes shared by both regulons R_i and R_j. We took the reciprocal of c_{ij}, $\frac{1}{c_{ij}}$ as the distance d_{ij} between the two regulons. A neighbor joining tree (Figure 1.6) based on a distance matrix such defined was constructed using PHYLIP [82] and displayed by MEGA [63].

1.5 Conclusions

In this study we have predicted LexA-binding sites and analyzed the putative LexA regulons in 26 cyanobacterial genomes that harbor a *lexA* gene using a highly efficient motif scanning and regulon prediction algorithm. In most *lexA*-containing cyanobacterial genomes, some SOS response genes bear a putative LexA box. Some genes involved in various cellular processes such as photosynthesis, drug resistance, etc. are also predicted to bear a putative LexA box in their promoter regions. However, in *Synechocystis sp.* PCC6803, LexA might have adopted a new function and no longer be in charge of the SOS response genes. In some genomes, *lexA* was likely lost during the course of evolution accompanied by the loss of its binding sites. The SOS response genes in these genomes that appear to lack a *lexA* gene might be regulated by another or multiple transcription factors. Moreover, we conclude that cyanobacteria inherited the *lexA* gene from their last common ancestor; however, substantial genome-wide turnover seems to have led to the high degree of variation of the LexA regulons in some species during evolution.

CHAPTER 2: RECONSTRUCTION OF OPERON STRUCTURES IN PROKARYOTES USING DIRECTIONAL RNA-SEQ SHORT READS BY A HIDDEN MARKOV MODEL

2.1 Abstract

Although prokaryotic gene transcription has been studied over decades, many aspects of the process remain poorly understood. Particularly, recent studies using tiling array and RNA-seq have revealed that prokaryotic transcriptomes are far more complex and dynamic than previously thought. Genes in an operon are often alternatively and dynamically transcribed under different conditions, revolutionizing the classic operon definition. With continuous drop in costs, RNA-seq becomes the major method for profiling prokaryotic transcriptomes. However, it is a challenging task to accurately assemble full length transcripts/operons using short reads because of the highly labile nature of prokaryotic RNAs and the read bias of current RNA-seq techniques, leading to many uncovered parts in transcripts. To address this missing-read problem, we have developed a Hidden Markov Model based algorithm, TruHmm, for reconstructing full length transcripts/operons using directional RNA-seq reads. When tested on a dataset of *Escherichia coli* K12 under a variety of culture conditions and growth phases, TruHmm has achieved rather high specificity and sensitivity for assembling multi-gene operons. As RNA-seq becomes a routine for probing transcriptomes in prokaryotes, TruHmm can be a useful tool for understanding the complexity of transcriptomes and the underlying mechanisms in prokaryotic cells.

2.2 Introduction

In prokaryotes, several adjacent genes on the same strand of DNA are often co-transcribed as a polycistronic mRNA, forming a multi-gene transcription unit called an operon. More recently, it is found that in addition to genes, some parts of non-coding sequences and the opposite strands of coding sequences can be also transcribed under certain conditions, generating non-coding RNAs (ncRNAs) [1] and anti-sense RNAs (asRNA) [83, 84], respectively. Accumulating evidences suggest that ncRNAs [1, 85] and asRNAs [83, 84] may have important roles in the physiology of prokaryotes. Therefore, a full understanding of the transcriptomes of prokaryotic cells is necessary to annotate the functional elements in their genomes and reconstruct the gene transcriptional networks in their cells. However, experimental determination of operon structures, ncRNAs and asRNAs by traditional molecular biology methods is time consuming and labour-intensive. As a result, no single prokaryote has so far had all of its operon structures, ncRNA and asRNAs characterized by such methods. For instance, even for the most well-studied model bacteria *E. coli* K12 and *B. subtilis*, only 3,409 [86] and 736 [87] operons have been determined in their genomes using these methods, respectively, after decades of research. On the other hand, although great progresses have been made in computational prediction of operons [75, 88-94] and small RNA genes [95-98], the accuracy of these predictors are still low [93, 99], and they can only predict the longest possible operons without considering possible alternative operons [75, 88-94].

In the past few years, increasing applications of whole genome directional (strand-specific) tiling array and directional RNA-seq techniques to prokaryotes have completely changed our view of the architecture and complexity of prokaryotic transcriptomes [2-9].

For example, using a combination of whole genome directional tiling array and RNA-seq techniques, Guell *et al*. [10] found that operon utilizations in the reduced parasitic *M. pneumoniae* genome are highly variable and dynamic, almost half of 139 identified multi-gene operons show varying (dynamic) expression in a staircase-like manner. Under different conditions, operons could be divided into smaller sub-operons, resulting in many alternative transcripts, suggesting that the operon structures in *M. pneumoniae* is highly dynamic, more similar to that of alternative splicing in eukaryotes than originally thought [10]. They also identified a large number of ncRNAs and asRNA expressed under various culture conditions, thus a much larger portion of the genome is transcribed than originally anticipated [10]. Similar results were observed in many other species [11-14].

Compared to the tiling array technique, the RNA-seq method is more suitable for understanding the complexity of the prokaryotic transcriptomes due to its single-nucleotide resolution, higher dynamic range, and lower noise natures, thus it has gained increasing popularity [100]. One important step in RNA-seq data analysis is to accurately assemble all meaningful transcripts in their full-length, so that correct conclusions can be drawn from typically tens of thousands of short RNA-seq reads. However, as has been shown earlier [10-14, 101, 102] and we will indicate later in this paper, the coverage of reads on transcribed regions in these studies are highly non-uniform, and there are even numerous zero coverage positions in transcribed regions [103-105], leading to gaps in otherwise an overlapping mapping of reads to a transcribed region [106-108]. These gaps make the transcriptome assembly a highly challenging task [10, 29-32, 101, 109]. Several technical problems in current library construction protocols and sequencing technologies

have been recently identified responsible for the non-uniform coverage and gaps. First, chemical fragmentation of RNA employed in many protocols may have bias to break or degrade some sequences [110]. Second, random primer based reverse transcription may preferentially transcribe some sequences than others [107, 111]. Third, ligases may preferentially link the adaptors to some sequences [112-114]. Fourth, PCR amplification is well-known for introducing GC contents-dependent bias in libraries [115-118]. Fifth, it was also recently found that sequencing errors were biased to some specific sequences, making such sequences missing the reads [119]. Moreover, prokaryotic RNAs are more labile than their counterparts in eukaryotes, thus segments of some RNAs can be more easily lost during the library preparation. Although some of these problems can be avoided by new technical development, such as using FRET-seq for amplification-free sequencing to avoid GC content-dependent PCR bias [120], or using single RNA molecular sequencing for longer reads to ease assembly problem [121, 122], no routine effective technique has been developed to avoid all these problems, however.

Although several transcriptome assemblers using short RNA-seq reads have been developed in the past few years, they are mainly for reconstructing alternative isoforms in eukaryotes [31]. These assemblers can be classified in two basic categories: the reference-based assemblers when a reference genome sequence is available, and the de novo assemblers when a reference genome is not available [31]. The reference-based assemblers usually involve two steps: RNA-seq reads are first mapped to the reference genome using an aligner, such as BLAT [123], TopHat [124] or Bowtie [125], and then a graph representing all possible isoforms from overlapping reads is constructed, and isoforms are resolved by traversing the graph. Examples of this strategy include Cufflinks

[32] and Scripture [126]. On the other hand, the de novo assemblers such as Trinity [127], Oases [128], TransAByss [129], Rnnotator [130], and Multiple-k [131], generally assemble isoforms based on a De Bruijn graph constructed using overlapping reads. The advantage of de novo strategy is that it can assemble the transcripts when a reference genome is not available and can recover transcripts that are missing in the genome assembly. However, de novo transcriptome assembly is very sensitive to sequencing errors, missing reads and chimerical reads in the dataset, and their accuracy is generally lower than the reference-based approaches [31].

With thousands of sequenced prokaryotic genomes available, transcriptome assembly in prokaryotes can often be done using the reference-based approaches. The only reference-based transcriptome assembler for prokaryotes that we are aware of is a Hidden Markov Model (HMM)-based method for reconstructing operons in *Bacillus anthracis* [132]. However, the aforementioned gap problem in transcript assembly was not fully addressed in the algorithm, and no tool was delivered from this research [132]. On the other hand, *de novo* assembly can be even more challenging in prokaryotes owing to prevalence of zero-coverage gaps caused by the aforementioned problems. Because of the lack of a good prokaryotic assembler that sufficiently addresses the gap-problem, currently prokaryotic transcripts were assembled by either simply stitching the two covered segments if the gap between them is shorter than a cutoff [5], or determining 5' and 3' ends of transcripts via a probability-based approach [27], or relying on an additional source of information for the assembly, such as tiling array data that tend to have a more even and consecutive coverage along transcribed regions albeit at a lower resolution level [10, 12]. Therefore, as RNA-seq become a routine technique for probing

transcriptomes in prokaryotes, an efficient and more accurate assembly algorithm and tool that are tailored to prokaryotic transcriptomes are urgently needed in the research community.

To meet this need, here we present a reference-based prokaryotic transcriptome assembly algorithm and tool, TruHmm (TRancription Unit assembly by a Hidden Markov Model), which specifically addresses the gap problem in assembling prokaryotic transcriptomes using a HMM. When evaluated on a directional RNA-seq dataset collected in *Escherichia coli* K12 str. MG1655 (*E. coli* K12) under different culture conditions and growth phases, TruHmm is able to reconstruct known operons with very high sensitivity and specificity. Since other reference-based assemblers were designed to reconstruct eukaryotic isoforms, we compared TruHmm with the state-of-the-art *de novo* transcriptome assembler, Trinity [127]. Our method outperforms Trinity in most accuracy metrics, especially in sensitivity.

2.3 Material and methods

2.3.1 Bacteria culture

A frozen stock of *Escherichia coli* K12 strain MG1655 (a gift from Dr. Todd Steck, Department of Biology, the University of North Carolina at Charlotte) was thawed, inoculated in LB medium in a test tube by 1:100 dilution and cultured overnight at 37 ^{o}C and 250 rpm. The cells were then transferred to fresh LB medium in a flask by 1:100 dilutions, and cultured at 37 ^{o}C and 250 rpm. When the cells grew to the log phase with an optical density at 610 nm [OD_{610}] of 0.87, they were spun down at 3,200g for 25 min. For heat shock treatment (HS), the cell pellets were resuspended in the same volume of MOPS medium (100 ml of 10X MOPS mixture, 880ml of sterile H_2O, 10ml (0.132M)

KH_2PO4 and 10ml of 20% glucose, Teknova, Hollister, CA), and incubated at 48^oC and 250 rpm. For phosphorus-starvation treatment (M-P), the cell pellets were resuspended in the MOPS medium without KH_2PO4. Three milliliter cell suspension were collected in a tube containing 1.5ml RNA Later (Invitrogen) immediately after the cell pellets were resuspended in the indicated medium (0 min) and at the indicated time points thereafter (HS:15min, 30min and 60min; M-P: 0hrs, 2hrs, 4hrs). Cells were spun down at 6,000g, 8 min and -4 ^oC, and the pellets were resuspended in 1.5 ml of *RNAlater*. The samples were stored at -800 ^oC until use.

2.3.2 Isolation and enrichment of mRNA

RNA was isolated using RiboPure^{TM} -Bacteria Kit (Ambion) following the manufacturer's instructions. Once isolated, ~10g total RNA was treated with 8 units DNase (Invitrogen) twice to remove genomic DNA, and the complete removal of DNA was confirmed by 35 cycles PCR amplification of a 196 bps fragment of the *crp* gene (5'-primer:AGCATATTTCGGCAATCCAG;3'-primer:TACAGCGTTTCCGCTTTTTC). rRNAs were removed from the total RNA using a MICROBExpress kit (Ambion) to enrich mRNAs.

2.3.3 Construction of directional RNA-seq libraries

In our earlier experiments, sequencing was done on an Illumina GAII platform at the sequencing core facility of the University of North Carolina at Chapel Hill, and the directional RNA-seq libraries were constructed by following an Illumina's instruction using their Small RNA Sample Prep Kit with some modifications. Briefly, after the purified mRNA was fragmented using a RNA fragmentation kit (Ambion), the fragmented RNA was treated with Antarctic phosphatase (NEB) to remove the 5'-tri-

phosphate groups of RNAs with an intact 5'-end. A mono-phosphate group was then added back to the 5'-end of RNAs by polynucleotide kinase (PNK, NEB) in the presence of 10 mM ATP. The v1.5 sRNA 3' Adaptor (5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTG /3ddC/) was ligated to the 3'-end of fragmented RNAs using truncated T4 ligase 2 (NEB), and the SRA 5' RNA adaptor (5'GUUCAGAGUUCUACAGUCCGACGAUC) was ligated to the 5'-end of fragmented RNAs using T4 ligase. To preserve short inserts from small RNAs we omitted the size selection step after PCR application of inserts. For our later experiments, sequencing was done on an Illumina HiSeq 2000 platform at David H. Murdock Research Institute of the North Carolina Research Campus (Kannapolis, NC), and we constructed the directional RNA-seq libraries using Illumina's TruSeq Small RNA Sample Prep Kit, so that multiplex sequencing can be achieved by using the barcoded PCR primers. The details of the method will be described elsewhere (Dong, Li and Su). Briefly, after similar treatments as described above, the 5' Adapter (RA5: 5' GUUCAGAGUUCUACAGUCCGACGAUC), and 3' Adapter (RA3: 5' TGGAATTCTCGGGTGCCAAGG) were ligated to 5'- and 3'-end of fragmented RNAs, respectively. Reverse transcription-PCR (RT-PCR) was performed using SuperScript II Reverse Transcriptase Kit using the SRA RT primer, followed by 16 cycles of PCR amplification. Again, the size selection was omitted on PCR products to preserve short inserts from possible small RNAs. Single-end sequencing on the Illumina GA II platform was done with 76 cycles, while that on the HiSeq 2000 platform was done with 100 cycles. Some samples (HS15min and M-P4h) were sequenced on both platforms.

2.3.4 Mapping and filtering RNA-seq reads

The genome sequence of *E. coli* K12 substr. MG1655 was obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/). The gene annotation file and the experimentally verified operons in the bacterium were downloaded from RegulonDB [60] (http://regulondb.ccg.unam.mx/). A total of 4501 annotated genes (also including pseudo genes and non-coding small RNAs) are included in this analysis. As the reads were not size-selected during the library construction, we trimmed the 3' adapters attached to some short insertions. Adapter-free reads with lengths of <10 nucleotides (nts) were discarded; the remaining reads were mapped to the *E. coli* K12 genome using Bowtie [125]. For the reads of length 10-14, 15-29 and ≥30 nts, up to 1, 2, and 3 mismatches were allowed, respectively. Only uniquely mapped reads were used for further analysis. The alignment of mapped reads to the reference genome was visualized by Integrated Genome Browser (IGB) [133].

As for comparison to the de novo assembler Trinity [127], we first reconstructed the transcript using the following parameters:

Trinity.pl --seqType fq --kmer_method meryl --SS_lib_type F --single *.fastq --CPU 12

Then the alignment of the reconstructed transcript to the reference genome was generated using the BLAT program [123], using the built-in command of Trinity:

alignReads.pl --single Trinity.fasta --target NC_000913.fa --seqType fa --aligner BLAT --SS_lib_type F

2.3.5 Normalization of the mapped counts

Normalization of the mapped counts is crucial for differential expression detection using RNA-seq [134], as different samples may have different total read counts, i.e. sequencing depths, as well as various bias mentioned earlier. The most commonly used methods include reads per kilobase of exon model (or open reading frame) per million mapped reads (RPKM) [103], fragments per kilobase of transcript per million fragments mapped (FPKM) [32], the hypergeometric model [135] and the more recent sophisticated model-based methods [104, 105, 107, 108, 115, 116, 136, 137]. However, it has been found that these kinds of global normalizations are strongly affected by a small proportion of highly expressed genes in the published datasets, leading to biased estimation of gene expression level across different conditions [134]. As shown in Figure 2.1, our datasets are no exception to the problem as around 10% of gene with highest

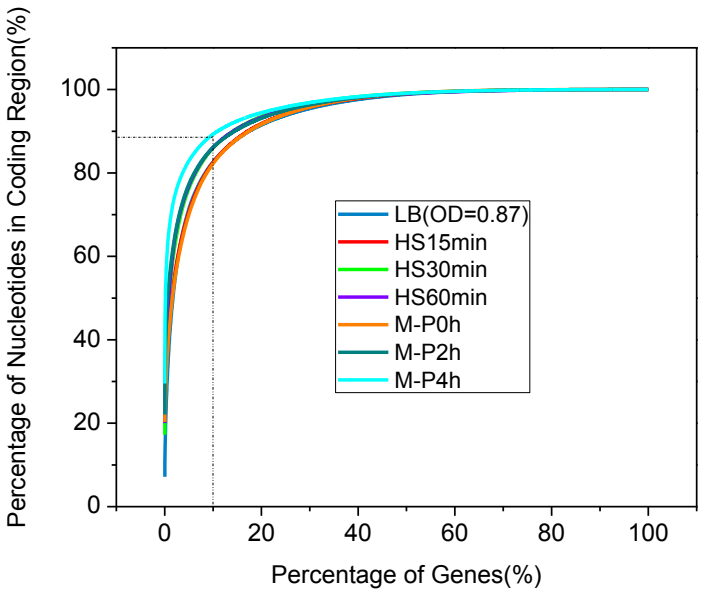


Figure 2.1. Impact of highly expressed genes on the mapped nucleotides in coding regions. Genes are sorted in the descending order of their number of mapped nucleotides in reads. The top 10 percent of genes with highest read counts contribute to around 80% ~90% mapped nucleotides in the coding regions.

number of mapped nucleotides contribute up to 80%~90% of mapped nucleotides in the gene-coding regions across all the seven samples. Inspired by [134] and also for computational efficiency, in this study we used N^* defined as the total nucleotide counts minus the counts of the top 10% of genes with highest counts to scale the gene expression level in each sample, instead of using the total counts of mapped nucleotides in each sample.

Furthermore, because our mapped reads have different lengths (see RESULTS AND DISSCUSION), instead of using the mapped read counts per gene, we used the mapped nucleotide counts per gene to measure the gene expression level defined as "Nucleotides Per Kilo base of transcript per Billion nucleotides mapped" (NPKB):

$$NPKB = \frac{n}{\frac{N^*}{10^9} \times \frac{L}{10^3}}, \qquad (2.1)$$

where n is the number of nucleotides of the reads mapped to the transcript, N^* our normalization factor defined above, and L the length of the transcript. Clearly, when all reads have the same length, NPKB and RPKM differ by a constant scaling factor. A similar method has been used earlier [101], except that our NPKB is further normalized by the global scaling factor N* in each sample.

2.3.6 Training the HMM and reconstruction of full length transcripts/operons

A HMM is a machine-learning algorithm that can be used to decode the path of hidden states that generate a sequence. In this paper, we use a HMM to infer whether or not a segment of a strand of DNA is consecutively transcribed given the expression values obtained from the mapped reads. The model consists of two states: the expression state E and non-expression state N (Figure 2.2).

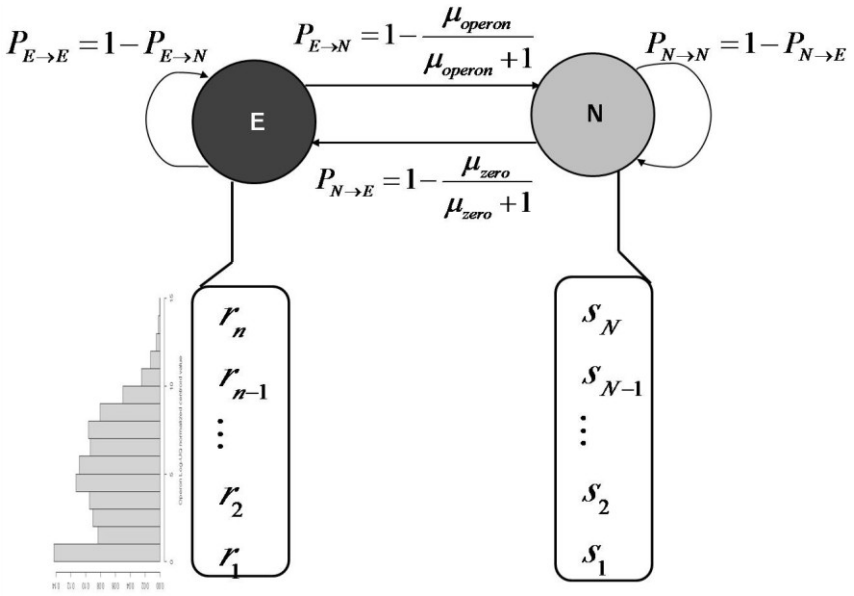


Figure 2.2. Structure of the HMM for transcript assembly using RNA-seq reads. E represents the expression state and N the non-expression state, Letters $r_1, r_2,\ldots,r_n$ are the emission values of E; and $s_1, s_2,\ldots, s_N$ are the emission values of N.

2.3.6.1 Selection of Expressed Adjacent Operon Pairs.

A gene is considered sufficiently expressed if over 50% of its length is covered by at least one read and at least 20nts of both of its termini are covered by at least one read. We used the 476 experimentally verified operons in RegulonDB (supplementary file 2) to train the parameters of the HMM, and evaluate the performance of our algorithm. Since these operons were not necessarily expressed in our samples, and alternative operon utilization could be very prevalent, as the first step to construct a positive operon set in a sample, we selected a pair of adjacent genes in a known operon (adjacent operon pair) if they met the following two criteria: 1) both genes were sufficiently expressed and over 50% of the length of their intergenic region were covered by at least one read in the sample; and 2) the correlation between the expression levels of the two genes and their intergenic region was greater than a cutoff. To compute the correlation between the expression

levels of the two genes and their intergenic region, we extended the two ends of the intergenic region into the two flanking genes to double its length or extended until the other end of either gene was reached (Figure 2.3A). We equally divided the extended intergenic region as well as the intergenic region into *n* bins, and thus the expression levels (NPKB) over these bins formed two *n*-element vectors (Figure 2.3B). Pearson Correlation Coefficient (PCC) between the two vectors was used to quantify the correlation between the expression levels of the two genes and their intergenic region. To find an appropriate cutoff, we similarly divided a sufficiently expressed gene as well as

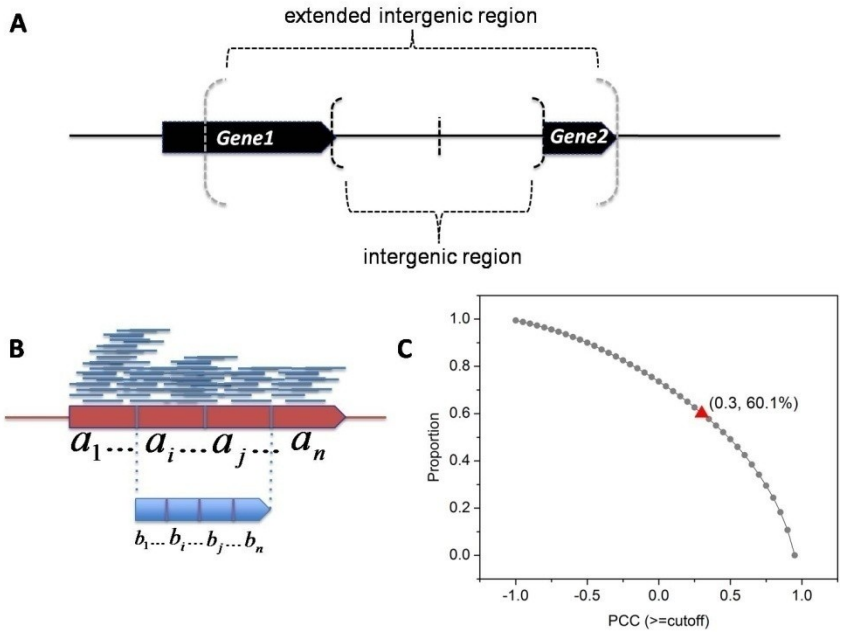


Figure 2.3. Selection of known operon pairs for training and evaluation. A: The intergenic region between two adjacent genes in an operon is doubled by extending its two ends in the two flanking genes. B: A sufficiently expressed gene is equally divided into n bins, and its central half is further equally divided into n bins. The NPKB values for each bin of a gene and its central portion are $a_1,\ldots, a_i,\ldots, a_j, \ldots, a_n$ and $b_1,\ldots, b_i,\ldots, b_j,\ldots, b_n$, respectively. An extended intergenic region is similarly divided by treating it as a "gene" with the intergenic region being the central portion of the "gene". C: Distribution of PCC values between these two vectors for the sufficiently expressed genes with the bin size n=4. We choose 0.3 as the cutoff of PCC value since 60.1% of sufficiently expressed genes can be included.

its central half into *n* equal bins, and computed the correlation of the expression levels between the whole gene and its central half. We reason that for an expressed adjacent operon pair, the PCC value between the intergenic region and the extended intergenic region should follow the same distribution of the PCC value between the central half of an expressed gene and the whole gene, since an adjacent operon pair and their intergenic region should be expressed in a similar way as the different parts of a gene. The distribution of the PCC value between the central half and the whole gene (n = 4) is shown in Figure 2.3C. We chose 0.3 as the cutoff for our second criterion to select positive adjacent operon pair since this would allow us to include over 60% of sufficiently expressed genes.

2.3.6.2 Positive and Negative Training Sets.

To train the HMM, we constructed a positive training set in a sample by simply stitching the known adjacent operon pairs that met the two criteria described above to form a large operon if it was a subset of a known operon according to RegulonDB. These positive training subsets in the seven samples are listed in supplementary file 2. To construct a relatively large negative training set in a sample, we included all the zero-coverage regions excluding the ones inside the sufficiently expressed genes in the sample.

2.3.6.3 Positive and Negative Testing Sets.

We evaluated the operon prediction accuracy using two methods: one was based on adjacent operon pairs, and the other on the entire operon structure using all the gene pairs of in a known operon. For the first method, we constructed a positive testing set in a sample, consisting of sufficiently expressed adjacent gene pairs, and a negative testing set consisting of known adjacent non-operon pairs that were both sufficiently expressed in

the sample. A known adjacent non-operon pair was made of either the first gene of a known operon and its immediate upstream gene, or the last gene in a known operon and its immediate downstream gene, as long as the intergenic region of the gene pair had at least one zero-coverage region, regardless of its length. For the second method, we constructed a positive testing set in a sample, consisting of all pair-wise combinations of the genes in a sufficiently expressed operon, and a negative testing set consisting of the gene pairs between the genes of the operon and the immediate upstream or immediate downstream gene, given that the known adjacent non-operon pairs with non-overlapping (UnTranslated Region) UTR as well as all these relevant genes were sufficiently expressed.

2.3.6.4 Leave-one-out Cross Validation.

We employed a leave-one-out cross validation strategy to evaluate the performance of our algorithm. Specifically, we used the positive training sets and negative training sets in (n-1) samples to train the emission and transition probabilities of the HMM, and used the positive testing set and the negative testing set in the remaining sample to test the trained model.

2.3.6.5 Training Emission Probabilities.

We used "coverage" to designate the number of reads mapped to a specific position (nucleotide) in the reference genome. To deal with the zero-coverage problem, we used a sliding window to compute a centroid coverage of each position on DNA, assuming that if the flanking regions of a position are transcribed, it is very likely that the position itself is also transcribed. Specifically, given a window size L (L is an odd number), the centroid coverage of the nucleotide i in the middle of the window is defined as:

$$Centroid(i) = \log(\frac{10^9}{N^*}(\frac{1}{L}\sum_{k=i-(L-1)/2}^{i+(L-1)/2} Coverage(k)+1)), \qquad (2.2)$$

where i is the *i-th* position (nucleotide) on the chromosome. N^* the normalization factor defined in equation (2.1), L the window size and Coverage (k) the coverage at position k in the reference genome. Note that a pseudo count of 1 is added to the average value of each window. The optimal window size is determined by balancing two goals with opposite effects: to cover as many gaps as possible and to exclude as many interoperonic regions as possible. See 2.4 Results and discussion for details of window size selection.

The emission signals of the states E (r_1, r_2, …,) and N (s_1, s_2, …) are the centroid coverage values of nucleotides in the reference genome. We used the positive training sets to estimate the emission probabilities of the signals of E. The distribution of centroid coverage values of the positive training set from all samples except LB is shown in Figure 2.4. The QQ plot indicates that the centroid coverage values of the positive training set approximately follow a Poisson distribution, which is consistent with the earlier results [134]. Thus, the emission probability of the centroid coverage values in the state E could be computed by the Poisson distribution, whose parameters were estimated with the maximum likelihood method. Since our negative training set were virtually not covered by reads, the signal that the state N emits should be the centroid coverage values with zero coverage,

$$\log(\frac{10^9}{N^*}(\frac{1}{L}\sum_{k=i-(L-1)/2}^{i+(L-1)/2} 0+1)) \qquad (2.3)$$

Thus we arbitrarily assigned a high probability $1\text{-}10^{-20}$ for N to emit this value, and a low probability 10^{-20} for *N* to emit any other values. The value 10^{-20} is also a pseudo probability to avoid zero probability for decoding the HMM later.

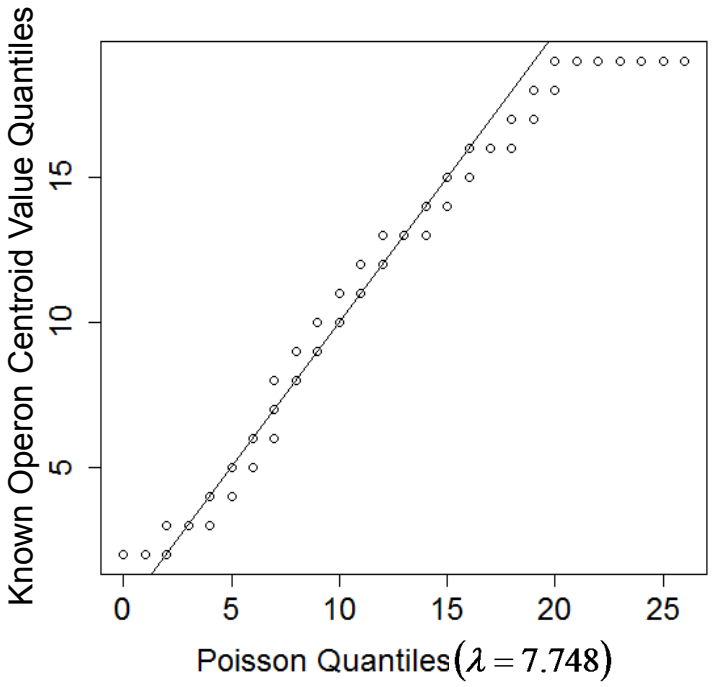


Figure 2.4 QQ-plot comparing the distribution of centroid coverage values of the positive training set in all the samples but LB with the fitted Poisson distribution. Deviation of a data point from the line y=x indicates its deviation from the theoretical Poisson distribution. Parameters of the Poisson distribution are estimated using the maximum likelihood method.

2.3.6.6 Training Transition Probabilities.

Let P_{ij} be the transition probability from state *i* to *j*. To determine the transition probabilities P_{EE} and P_{EN} i.e., the probability to stay in the state *E* and to transit from the state *E* to the state *N*, respectively, let *X* be the length of a consecutively expressed segment of genome DNA. Under the Markov assumption, we assume that *X* follows a geometric distribution,

$$P(X = n) = P_{EE}^{n} \cdot (1 - P_{EE}). \qquad (2.4)$$

Similarly, let Y be the length of a consecutively non-expressed segment of genome DNA. Then Y also follows a geometric distribution,

$$P(Y = n) = P_{NN}^{n} \cdot (1 - P_{NN}). \qquad (2.5)$$

We use the lengths of sufficiently expressed known operons in the positive training set to estimate the probability of staying in the state E as $P_{EE} = \mu_{operon}/(\mu_{operon}+1)$, where μ_{operon} is the mean value of the lengths of sufficiently expressed regions and can be determined from the raw coverage data. For example, using the positive training set from the all samples excluding LB, we obtained $\mu_{operon} = 1990.695$ nts (Figure 2.5A). Similarly, we use the lengths of non-expressed regions in the negative training set to estimate the probability of remaining in the state N as $P_{NN} = \mu_{zero}/(\mu_{zero}+1)$, where μ_{zero} also can be determined from raw coverage data, for example, $\mu_{zero} = 172.22$ nts for all the negative training set from all samples except LB (Figure 2.5B). The derivation of transition probabilities estimations is given in Figure 2.6. The QQ plot indicates that the lengths of expressed regions indeed can be nicely modelled as a geometric distribution (Figure 2.5C), but that of non-expression regions is not that good (Figure 2.5D), probably because of the gaps introduced in the expressed regions, constituting more short false non-expressed regions. However, we found that this deviation had little effects on the performance of the algorithm (see 2.4 Results and disscussion).

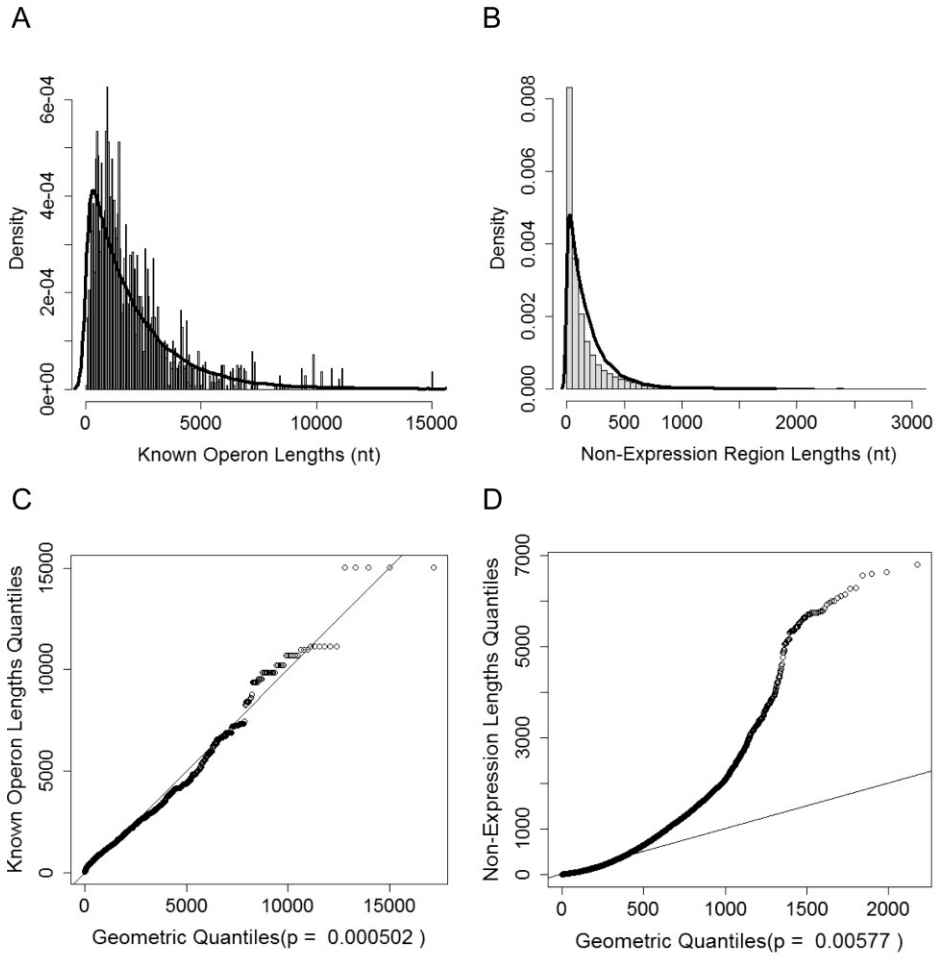


Figure 2.5. Distributions of the lengths of positive and negative training sets in all samples except LB. A: Histogram of the length of the positive training set (bin size =50 nt). The curve is the fitted geometric distribution with success probability p = 0.000502 estimated by the maximum likelihood method. B: Histogram of the length of negative training set (bin size =50 nt). The curve is the fitted geometric distribution with p = 0.00577 estimated by the maximum likelihood method. C: QQ-plot of the length of positive training set against the fitted geometric distribution. D: QQ-plot for the length of negative training set against the fitted geometric distribution.

2.3.6.7 Reconstruction of Operons.

We used the Viterbi algorithm [138] to decode the path of states that best explains the centroid coverage values of a region of DNA. If both genes in a neighbouring pair are at

least partly covered by a consecutive sequence of expressed states, the two genes are predicted as an adjacent operon pair. We stitched any two adjacent candidate operon pairs A-B and B-C to obtain the full length transcripts/operons. The transcription start site and transcription terminate site of predicted operon were determined by the locations of the 5'-end and the 3'-end of the stitched transcript, respectively. If over half of the length of a terminal gene is predicted to be expressed, this gene is considered as a member of the predicted operon, otherwise the terminal gene is only considered as the UTR of the operon. When comparing our algorithm with Trinity for their performance of reconstructing full length transcripts/operons, we applied the same stitching strategy to the transcripts assembled by Trinity.

$$
\begin{aligned}
&\Pr(Y = k) = (1-p)^k p \\
&E(Y) = \sum_{k=0}^{\infty} (1-p)^k p \cdot k \\
&\quad = p \sum_{k=0}^{\infty} (1-p)^k k \\
&\quad = p[\frac{d}{dp}(-\sum_{k=0}^{\infty}(1-p)^k)] \cdot (1-p) \\
&\quad = -p(1-p)\frac{d}{dp}\frac{1}{p} \\
&\quad = \frac{1-p}{p} \\
&\therefore \quad p = \frac{1}{1+E(Y)} \\
&\therefore \quad 1-p = \frac{E(Y)}{E(Y)+1}
\end{aligned}
$$

Figure 2.6. Derivation of transition probabilities P_{EE} and P_{NN}. The geometric distribution $\Pr(Y = k) = (1-p)^k p$ is used to model the number of failures before the first success. The length for the consecutive expression state *E* or non-expression state *N* should follow a geometric distribution. Therefore the probability of staying in the expression state

$P_{EE} = 1 - P_{EN} = \frac{E(Y_{EE})}{E(Y_{EE}) + 1} = \frac{\mu_{operon}}{\mu_{opeorn} + 1}$, and the probability to transit from the expression state to Figure 2.6 (continued) the non-expression state is $P_{EN} = \frac{1}{E(Y_{EE}) + 1} = \frac{1}{\mu_{opeorn} + 1}$. Similar results can be derived for P_{NN} and P_{NE}.

The algorithm was encoded in C++ and perl. The software package is open-source, and can be downloaded from http://bioinfolab.uncc.edu/TruHmm_package/. We provide the users the option to train their model if enough known operons are available in their genomes of interest and if more than two samples are available. Otherwise users can apply our algorithm using the default settings without the need of any training.

2.3.7 Performance Metrics

To evaluate the performances of our algorithm, we use the following metrics.

$$Sensitivity = Recall = TPR = \frac{TP}{TP + FN}$$

$$Specificity = 1 - FPR = \frac{TN}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - factor = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Where, TP (true positive) = Number of known operon pairs accurately classified as operon pairs by the model.

FP (False Positive) = Number of non-operon pairs falsely classified as operon pairs by the model.

FN (False Negative) = Number of known operon pairs falsely classified as non-operon pairs by the model.

TN (True Negative) = Number of non-operon pairs accurately classified as non-operon pairs by the model.

Sensitivity, i.e. TPR (True Positive Rate or recall) is the proportion of known operon pairs that can be correctly identified as operon pairs by the model. Specificity, i.e. 1-FPR (False Positive Rate) is the proportion of non-operon pairs that are correctly classified as non-operon pairs. Accuracy combines the two metrics to quantify the overall performance of the model. A high Accuracy value represents a low total error rate. Precision denotes the proportion of predicted positives that are true positives. F-factor combines Recall and Precision and normalized them to an idealized value.

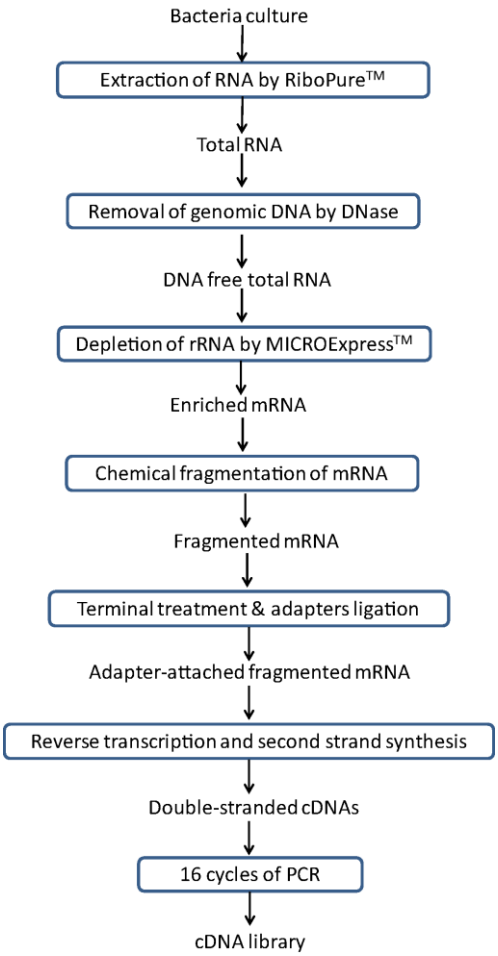


Figure 2.7. Flowchart of directional RNA-seq library constructions.

2.4 Results and discussion

2.4.1 RNA-seq Reads Quality

We prepared the directional RNA-seq libraries from seven *E. coli* K12 samples collected at the log phase growth in LB, and different time points under heat shock (HS) or phosphorus starvation (M-P) treatments, denoted as LB, HS15min, HS30min, HS60min, M-P0h, M-P2h, and M-P4h to reflect the treatment and sampling time points. The experimental procedure of our work is listed in Figure 2.7. The libraries were sequenced on either Illumina GA II or HiSeq 2000 platforms. Specifically, sample LB was sequenced using the GAII platform, samples HS30min, HS60min, M-P0h, and M-P2h were sequenced using HiSeq 2000 platform, whereas samples HS15min and M-P4h

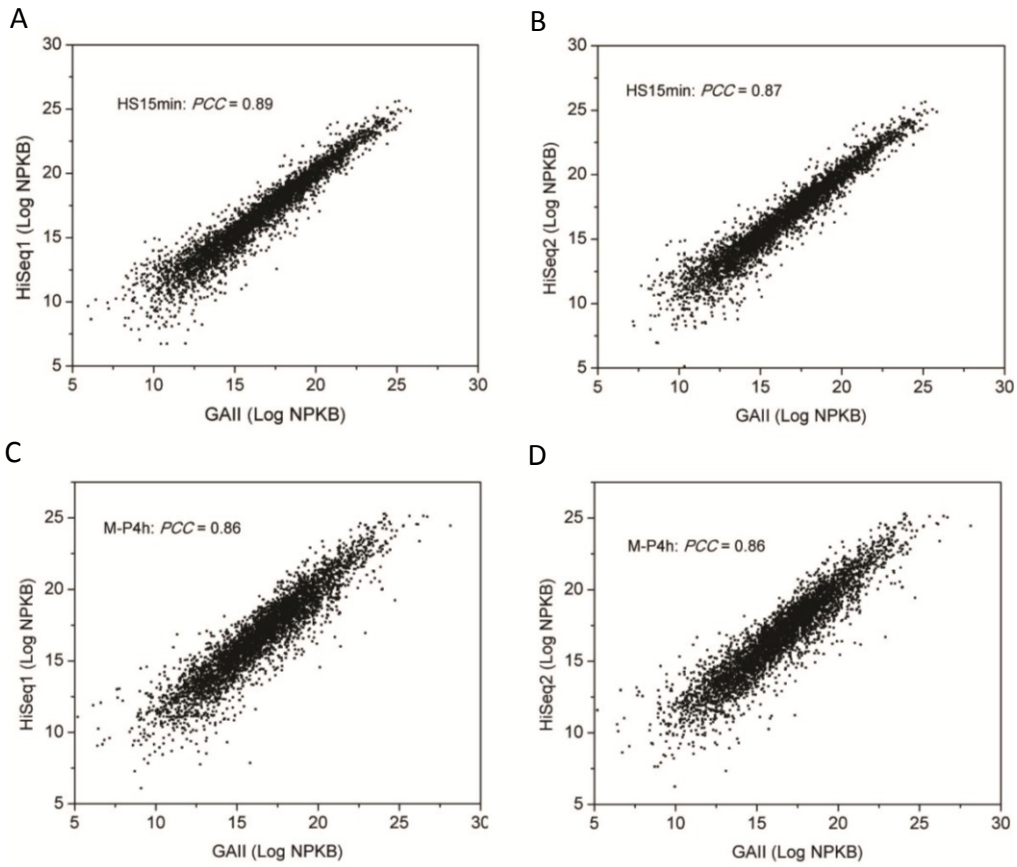


Figure 2.8. Correlation of expression levels of all the genes between two platforms: GAII and HiSeq. Each dot represents a gene. The expression level is evaluated using log of the NPKB values. A) PCC of expression levels for HS15min between GAII reads and HiSeq reads. B) PCC of expression levels for HS15min between GAII reads and 2nd HiSeq reads. C) PCC of expression levels for M-P4h between GAII reads and HiSeq reads. D) PCC of expression levels for M-P4h between GAII reads and 2nd HiSeq reads. The duplicates for each sample are from the same biological samples sequenced twice using platform HiSeq 2000.

were sequenced using both platforms. Each sample sequenced using HiSeq 2000 was repeated twice (technical replicates). The reads obtained from different platforms for the same sample are highly correlated (Figure 2.8), thus the data for the same sample were combined for the analysis. A total of 330,611,663 reads were generated from the seven samples. The mapping statistics of the samples are summarized in Table 2.1 showing that

Table 2.1. Summary of mapping results

Sample	Platform	Total reads	% Reads having adapter	Total reads after trimming	Uniquely mapped reads	Multiple mapped reads	Reads failed to map	% Unique	% Multiple	% Failed
LB	GAII	32,129,789	16.70%	31,767,554	12,856,757	14,956,218	3,954,579	40.47%	47.08%	12.45%
HS15min	GAII+HiSeq	72,868,580	60.29%	72,586,098	16,743,042	45,758,784	10,084,272	23.07%	63.04%	13.89%
HS30min	HiSeq	35,042,119	84.22%	34,979,745	13,034,034	19,411,877	2,533,834	37.26%	55.49%	7.24%
HS60min	HiSeq	25,930,027	80.37%	25,905,637	7,735,369	15,403,470	2,766,798	29.86%	59.46%	10.68%
M-P0h	HiSeq	46,464,309	81.87%	46,342,018	14,129,411	27,602,193	4,610,414	30.49%	59.56%	9.95%
M-P2h	HiSeq	67,034,479	76.30%	66,962,875	29,581,761	31,717,549	5,663,565	44.18%	47.37%	8.46%
M-P4h	GAII+HiSeq	86,184,479	51.16%	85,795,131	29,183,476	44,847,797	11,763,858	34.02%	52.27%	13.71%

23.07~44.18% of reads could be uniquely mapped to the genome, resulting in 7,735,369 ~ 29,581,761 uniquely mapped reads in each sample, corresponding to 93~355 time coverage of the genome. Of the 47.08~63.04% multiple mapped reads in each sample, over 99.6% were from duplicated tRNA/rRNA genes (data not shown). Furthermore, as shown in Figure 2.9, in all the samples over 90% and less than 10% of the total mapped nucleotides were mapped to the sense strand and intergenic regions, respectively, with only 0.35~0.95% of the total mapped nucleotides mapped to the antisense strand. These results indicate that most of our reads were from the sense strand, and thus our libraries were highly strand specific, which is consistent with an earlier result using a similar

library construction protocol [102].

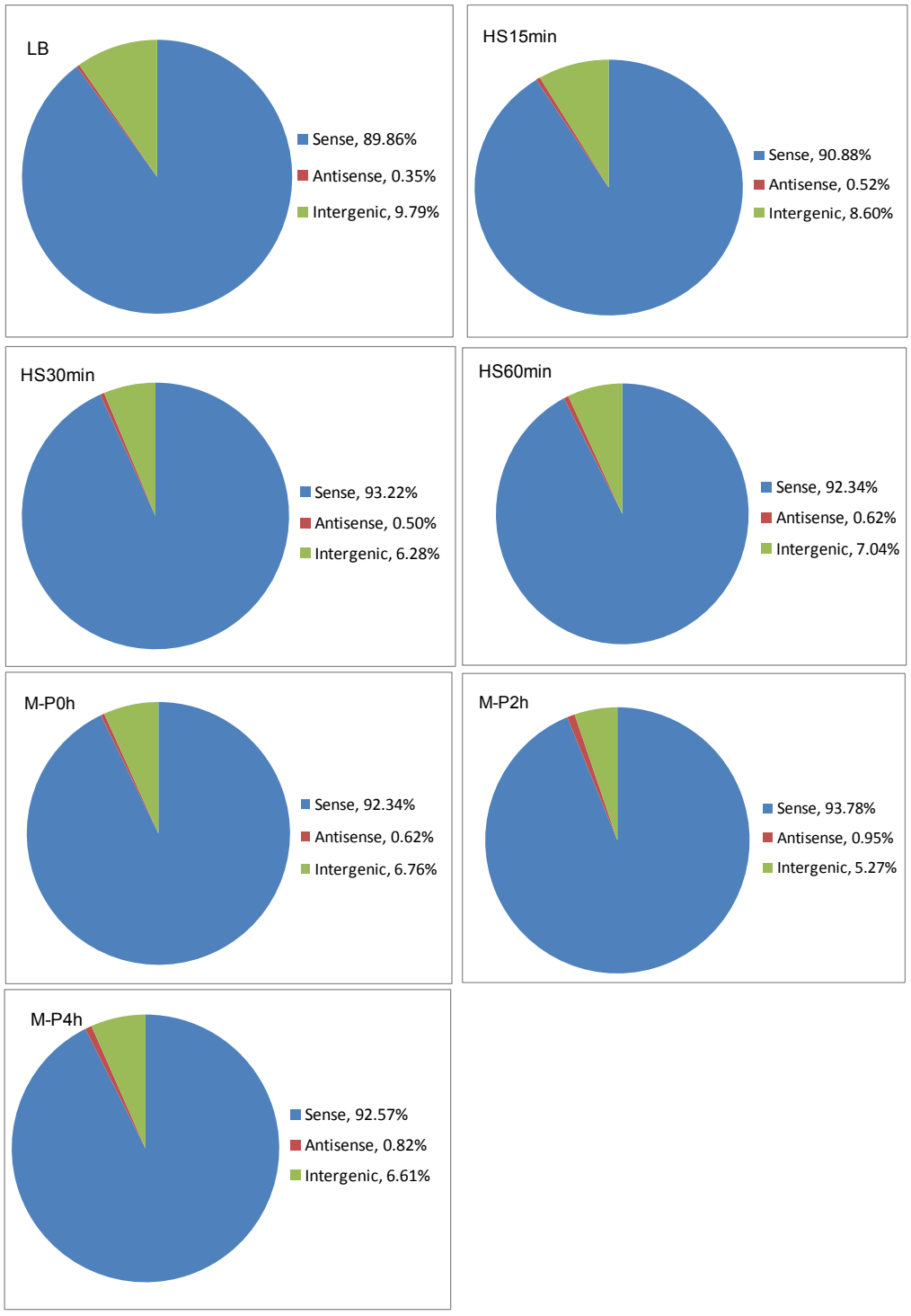
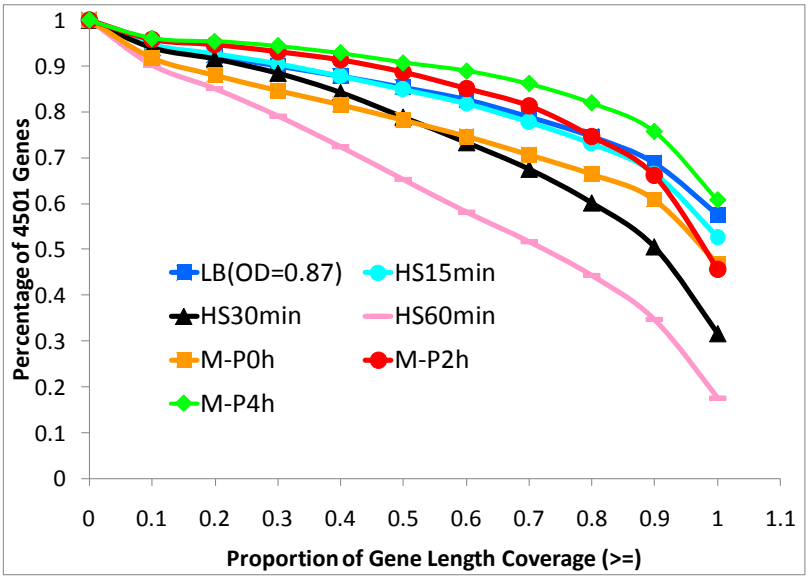


Figure 2.9. Strand specificity of the directional RNA-seq libraries. The percentage of total nucleotides mapped to sense strand, antisense strand and intergenic regions is shown for

Figure 2.9 (continued) the seven samples.

A



B

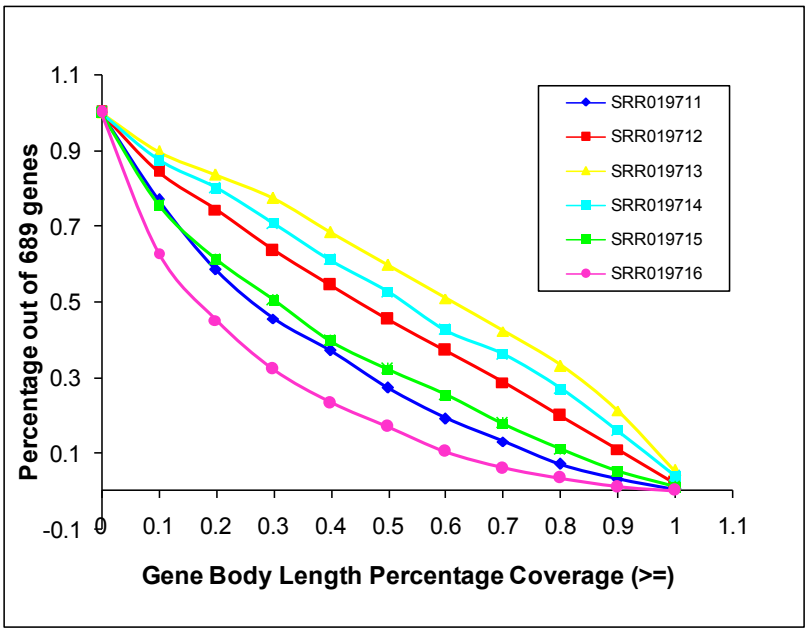


Figure 2.10. Distribution of the genes with more than the indicated percentage of their length covered by at least one read in the samples. A) Distribution generated by our data: Less than 60% of genes have their length completely covered by at least one read. Over 80% genes have over 50% of their length covered by at least one read except for sample HS60min. B) Distribution generated by Vivancos *et. al*[101]. Only less than 60% of genes have over 50% of their length covered by at least one read.

However, as shown in Figure 2.10A, even with such high levels of coverage, less than 60% genes in the genome had their length completely covered by at least one read, while only less than 90% genes in the genome had at least 10% of their length covered by at least one read, suggesting that some expressed transcripts were not completely covered by the reads. The very same problem has been widely noted in both eukaryotes [102-104, 107, 108, 139] and prokaryotes [11, 101] due to the aforementioned technical artifacts of the current RNA-seq techniques [106, 110, 111, 113, 119]. In fact, we found that the problem were even more serious in all published prokaryotic datasets we have reanalyzed, a typical example from [101] is shown in Figure 2.10B. These prevalent zero-coverage gaps may be also partially caused by the loss of some RNA fragments during the library preparation due to the highly labile nature of prokaryotic RNAs as mentioned earlier, in addition to the aforementioned technical artifacts. Our data seems to support this hypothesis, as the percentage of gene body coverage in our samples collected under heat shock treatment were generally lower than that in other treatments, in particular, after 30 and 60 min heat shock (Figure 2.10A). It is well known that RNAs have a shorter living time at a higher temperature. It is because of this gap problem that we define a gene with ≥50% of the length covered by at least one read to be sufficiently expressed. Also, this 50% cutoff was chosen, as all the samples except HS60min had over 80% of genes with at least 50% length being covered (Figure 2.10A). Moreover, as shown in Figure 2.11, our uniquely mapped reads consisted of well-balanced different sizes of RNA fragments, indicating that our library preparation protocol could potentially capture small RNA species such as as-RNA and ncRNA, which were otherwise left out by a typical size

selection step in the library preparation process. Additionally, as shown in Figure 2.12, in consistence with the earlier results [11, 109, 140], our libraries were also biased to the 5'-end of transcription units. The data have been submitted to NCBI SRA database with accession number XXX.

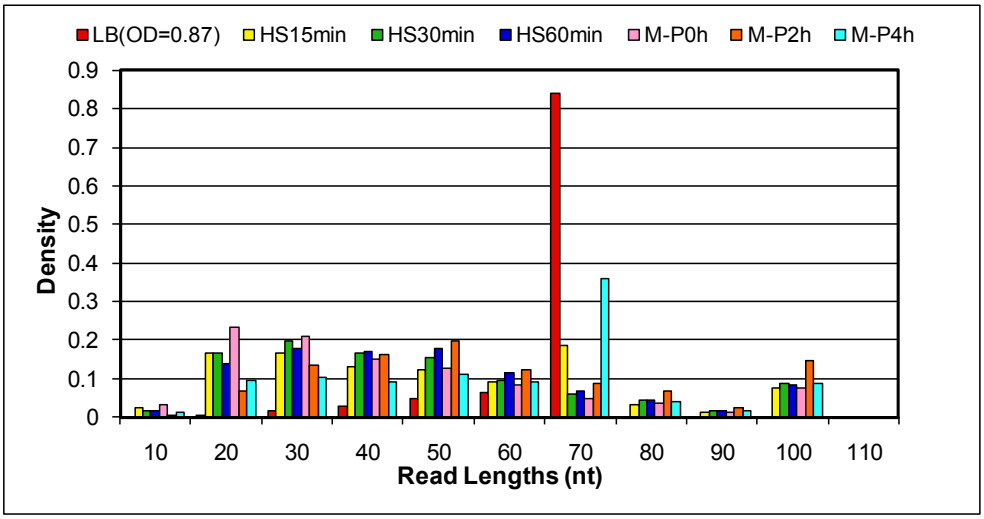


Figure 2.11. Distribution of the length of the uniquely mapped reads in the samples.

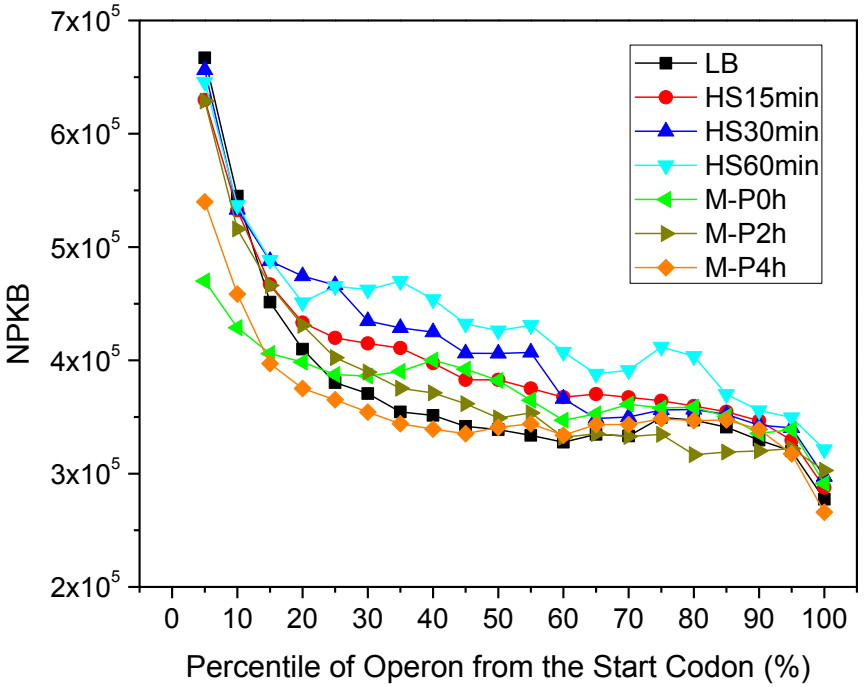


Figure 2.12. Average percentile coverage of known operons in each sample. The

Figure 2.12 (continued) sufficiently expressed known multiple-gene operons (supplementary file 2) and singleton operons are equally divided into 20 bins, and the expression values in each bin were computed and averaged in each sample. The top 10% highest expressed genes were excluded from the calculation.

2.4.2 Operon Prediction Accuracy

To compensate for the negative effect of zero-coverage gaps in the expressed regions on assembling, we used a centroid coverage value in a sliding window to represent the reads coverage for each nucleotide of DNA (see 2.3 Materials and methods). Meanwhile, we do not want to increase false positives by mistakenly bridging irrelevant reads using such a strategy. To find an appropriate widow size for this purpose, we plotted the distributions of interoperonic and gap lengths shown in Figure 2.13, which suggest that

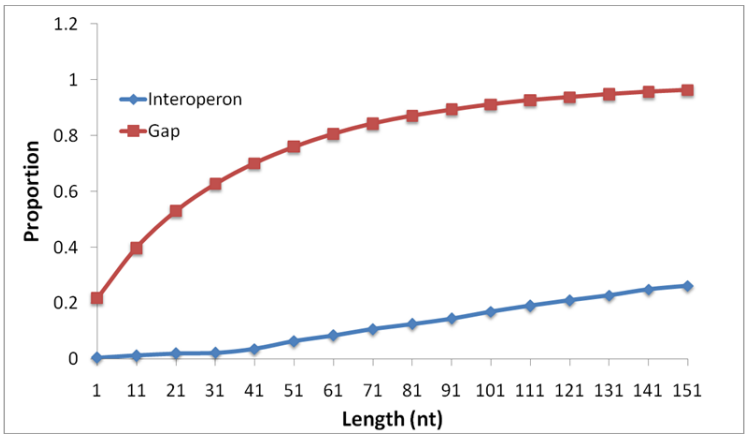


Figure 2.13. Distributions of the length of interoperonic regions and the length of gaps in sufficiently expressed regions.

the optimal window size might be shorter than 41nts. Therefore, we evaluated the performances of window size ranging from 1 to 41nts with an increment of 10nts on all the seven samples using the leave-one-out validation strategy as detailed in 2.3 Materials and methods. As shown in Figure 2.14A, when evaluated using the adjacent operon pairs,

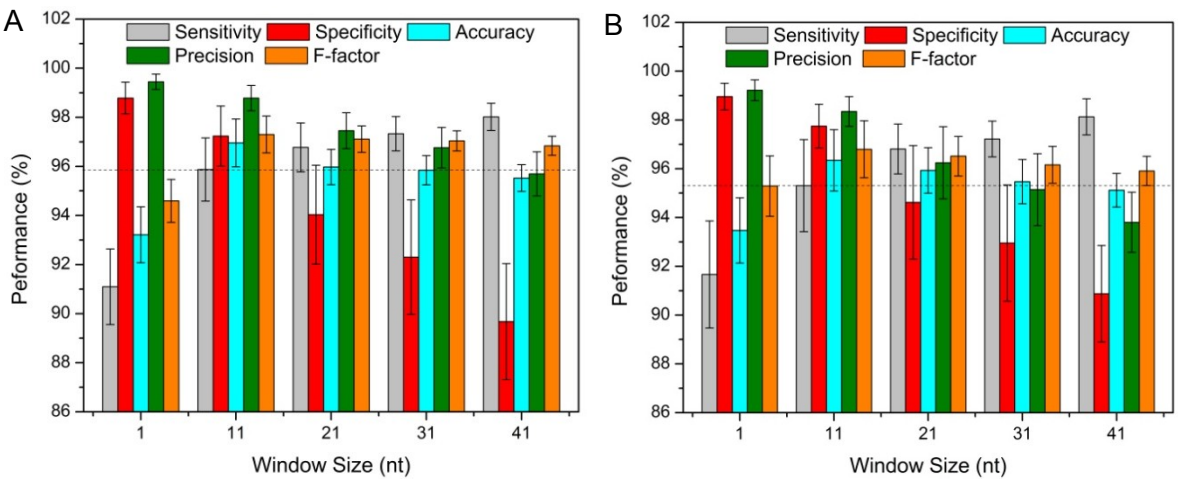


Figure 2.14. Evaluation of the algorithm on the seven samples by the five metrics. A) Using operon pairs. The dashed horizontal line is at the 95.87% level. The vertical bars indicate standard errors. B) Using entire operon structure. The dashed horizontal line is at the 95.3% level and the vertical bars indicate standard errors.

our algorithm was very robust for the choice of the window size in the range of 11~21nts (the mean values for each metric are ≥ 94%). Particularly, when the window size *L*=11nts, the algorithm achieved probably the best-balanced performance (the mean values for each metric are ≥ 95.87%), especially in terms of the three most important measures: sensitivity, specificity and accuracy. When evaluated using the entire operon structure, our algorithm still achieved very good performance with all the five metrics being over 94.6% for window size of 11~21nts (Figure 2.14B), and the best performance (the mean values for each metric are ≥ 95.3%) was also obtained when L=11nts. Therefore, we chose L=11nts for our further analysis. We also evaluated the effect of sequencing depth on the performance of our algorithms. As shown in Table 2.2 using M-P4h as an example, when the sequencing depth is over 153 times of genome size, our algorithm was very robust to the sequencing depth.

Table 2.2. Effect of sequencing depth on the performance of TruHmm using sample M-P4h as an example

Portion of total reads	Total unique mapped nt	Sequencing depth	Sensitvity	Specificity	Accuracy	Precision	F-factor	Total operons	Zero-coverage positions (nt)
10%	177,394,644	38.23	86.60%	96.00%	89.00%	98.00%	92.00%	2705	5,705,298
20%	354,676,352	76.44	89.60%	91.00%	90.00%	96.80%	93.00%	2513	4,983,472
40%	710,081,281	153.05	94.50%	93.20%	93.50%	95.80%	95.20%	2345	4,220,369
80%	1,419,721,627	306.00	96.60%	97.10%	96.50%	97.30%	96.95%	2133	3,438,913
100%	1,780,472,931	383.75	97.20%	98.90%	97.70%	99.00%	98.10%	2091	3,193,336

2.4.3 Prediction of Transcription Start Sites and Terminate Sites

We next evaluated the ability of TruHmm to define operon boundaries, i.e., the transcription starting sites (TSSs) and ending sites (TTSs) of transcripts. However, accurate evaluation of predicted operon boundaries is complicated by the recently discovered fact that alternative TSSs and TTSs are far more prevalent than previously thought [10-12, 109, 140], and we lack a real gold standard TSS and TTS datasets even though some different TSSs and/or TTSs are documented in RegulonDB. Thus we evaluated our reconstructed TSSs and TTSs by the following two ways. First, we wanted to know how many experimentally verified TSS and TTS in RegulonDB could be recovered by the boundaries of our assembled operons in any of the samples. If two known TSSs were within 10nt from each other, we considered them as the same one in our evaluation. Thus, there are 1,387 TSSs (supplementary file 3) associated with the genes expressed in the seven samples. We considered a known TSS was recovered by our predicted TSS if they were at most 20 nts from each other. Using this criterion, 587 out of 1,387 (~43.2%) known TSS were recovered by a total of our 6,424 predicted TSSs (supplementary file 3). Second, as for the remaining 5,837 predicted TSS with no match to a known TSS, 2,465 of which appeared in at least two samples, suggesting that they were likely to be true TSSs, The remaining 3,372 predicted TSSs that did not match the

known ones are summarized in Table 2.3. As for the TTS predictions, our algorithm

Table 2.3. Summary of predicted TSS not accordant with known TSS in regulonDB

Predicted TSS without matches	Upstream of known TSS	Downstream of known TSS	Associated with genes without known TSSs
Multiple appearance (2406)	598	92	1,775
Single appearance (3661)	771	189	2,412

recovered 148 out of 221 (~67%) known TTSs associated with expressed genes (supplementary file 3), which is higher than the recovery rate of known TSSs, even though the mapped reads are strongly biased to the 5'-ends (Figure 2.12). The lower recovery rates of known 5' ends (TSS) compared to 3' ends (TTS) might indicate that operons utilize more alternative TSSs than TTSs under different conditions. In other words, the predicted TSSs without a match with the known TSSs in RegulonDB are highly likely to be novel alternative TSS of the transcripts in different conditions. Furthermore, we computed the distribution of the distance in term of number of nucleotides between a predicted TSS and the first codon of the most upstream gene as well as the distance between a predicted TTS and the stop codon of the most downstream gene of a reconstructed operon. As shown in Figure 2.15A, the vast majority of the predicted TSSs are at least 25nt (average 55 nt) upstream of the start codon of first gene in a reconstructed operon, in agreement with the length of known 5'UTR in *E. coli* K12. Furthermore, as shown in Figure 2.15B, the vast majority of the predicted TTSs are at least 10 nt (average 45nt) downstream of the stop codon of the last gene in an operon, in consistence with the length of 3'-UTR of the known operons. Taken together, all these results strongly suggest that most of the predicted TSSs and TTSs are likely to be true

transcription boundaries. The assembled operons and their expression levels in each sample are listed in supplementary file 5. Nonetheless, as demonstrated in earlier studies [11, 109, 140], to more accurately detect TSSs and TTSs of transcripts/operons, specific libraries targeted to the 5' and 3'-ends of transcripts might be needed, in particular, a library targeted to the newly discovered transcription starting site RNAs (tssRNAs) [141].

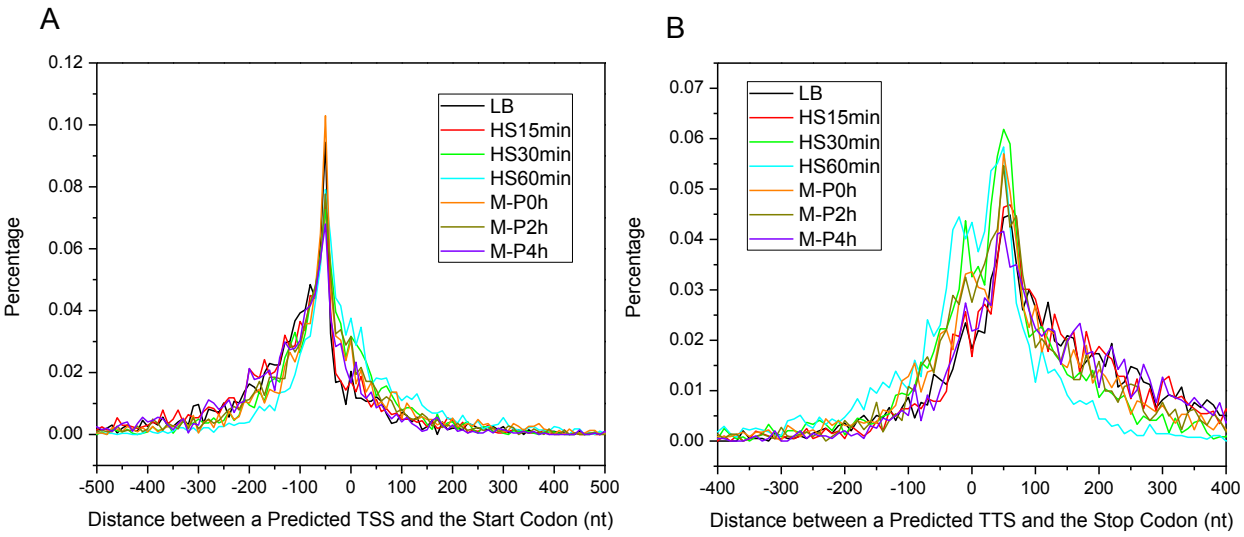


Figure 2.15. A) Distribution of the distance between the predicted TSS and the start codon of an assembled transcript. B) Distribution of distance between the predicted TTSs and the stop codon of an assembled transcript. A negative distance values means upstream, and a positive value downstream.

2.4.4 Comparison between TruHmm and Trinity

Since the existing reference-based transcriptome assemblers such as Cufflinks [32] and Scripture [126] were designed to reconstruct isoforms of genes in eukaryotes instead of the full length transcript/operon structures of prokaryotes, they perform very poorly on prokaryotic datasets (data not shown), we compared TruHmm with a state-of-the-art *de novo* assembler, Trinity [127], for recall of the sufficiently expressed known adjacent

operon pairs in each sample. To make the comparison relative even, we applied the same stitching strategy to Trinity. The original transcripts assembled by Trinity are listed in supplementary file 6, and the further stitched operons are provided in supplementary file 7. When evaluated on both adjacent operon pairs (Figure 2.16A) and entire operon structure composed of all the gene pairs inside/outside operons (Figure 2.16B), TruHmm

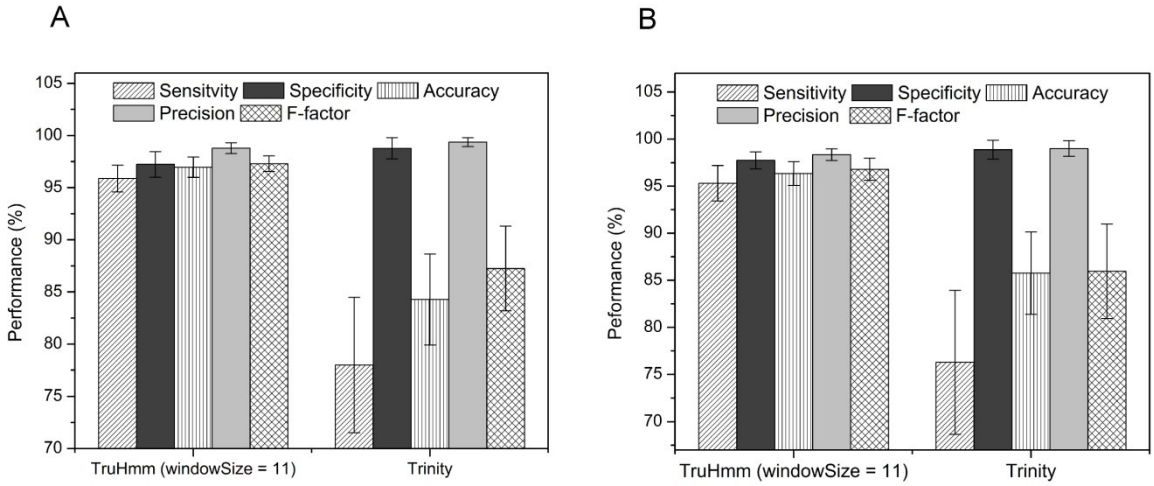


Figure 2.16. Comparison of performances between TruHmm and Trinity. A) Comparison of the performance of TruHmm (with window size 11) and Trinity on the seven samples based on the positive and negative sets composed of neighboring gene pairs. The vertical bars indicate standard errors. B) Comparison of the performance of TruHmm (with window size 11) and Trinity on the seven samples based on the positive and negative sets composed of all the gene pairs within/outside operons. The vertical bars indicate standard errors.

significantly outperforms Trinity in sensitivity, accuracy and F-factor for all the samples, although Trinity does have a little higher specificity and precision. The poor performance of Trinity in sensitivity might be caused by the lack of sufficiently overlapping reads of the datasets (Figure 2.10), based on which Trinity and many other *de novo* assemblers assemble a transcript.

2.4.5 Prediction of Alternative Operons

As summarized in Table 2.4, our algorithm detected more than 2,000 operons involving

Table 2.4. Summary of assembled operons in the samples

	LB(OD=0.87)	HS15min	HS30min	HS60min	M-P0h	M-P2h	M-P4h
# Genes expressed	4,314	4,366	4,340	4,222	4,296	4,395	4,420
# Hypothetical Proteins	29	29	31	27	28	30	32
# Operons	2,131	2,247	2,635	2,865	2,452	2,339	2,091
# Multi-gene operons	875	915	853	732	825	933	933
# Consistent operons	1,064	1,086	1,081	1,049	1,055	1,098	1,105
# Consistent multi-gene operons	207	207	207	206	207	206	207
# Alternative operons	1,065	1,160	1,552	1,815	1,396	1,239	981

more than 4200 genes in each sample. There were 1121 consistent operons appearing in at least two of the seven samples, 207 of which were multiple-gene operons (supplementary file 8). Of these 207 consistent multiple-gene operons, 206 were expressed in all the seven samples except the operon istR-1 istR-2/b4616, which was not expressed in samples HS60min and M-P2h (supplementary file 8). Figure 2.17 shows an example of a consistent operon *hemCDXY* encoding enzymes involved in tetrapyrrole synthesis. Although all the four genes were consistently expressed and continuously covered by the reads under different cultures and growth phases, they had similar position-dependent varying level of the read coverage along the genes and operon, indicating the non-uniform coverage of our libraries. As we indicated earlier, this phenomenon has been widely noted for different variants of current RNA-seq techniques [10, 11, 13, 14, 101-108].

Furthermore, we consider a non-consistent operon as an alternative operon if it shares a portion of genes with another operon in other samples. As shown in Table 2.4, from 981 to 1,815 alternative operons were detected in each sample. Thus, around 50% of the reconstructed operons in each sample had at least an alternative form, a number

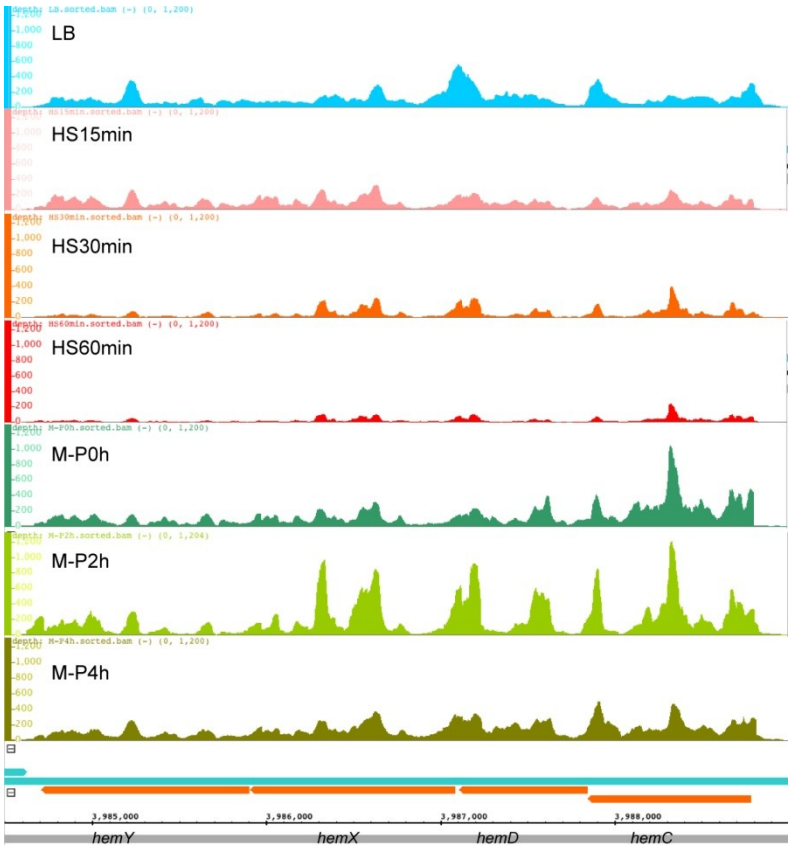


Figure 2.17. Reads coverage of the genes in the hem operon *hemCDXY*. The vertical axis is the number of reads covered at the position. The orange and dark green bars at the button of the graph represent the reverse and forward strands, respectively. Segments with arrows represent genes. Genes from the right to left are *hemC*, *hemD*, *hemX* and *hemY*. The graphs were generated using IGB. To make the expression levels for the four genes comparable in different samples, the same scale (1,200) of the vertical axis is used for all the samples. Although this four-gene operon is fully covered by the reads under different cultures and growth phases, note the similar position-dependent non-uniform coverage of the reads along the operon.

comparable to that found in *M. Pneumonia* [10] and other prokaryotes [11-14] demonstrating that like many other prokaryotes [2, 8, 10-12], *E. coli* K12 also expresses enormous alternative operons under different culture conditions and growth phases, and it is far more prevalent than previously expected. An interesting example is the 14-gene operon *phnCDEFGHIJKLMNOP* coding for proteins responsible for the assimilation of

C-P bond-containing phosphonates under phosphorus starvation conditions [142]. In LB, and heat shock samples (HS15min, HS30min and HS60min), this operon was split into several short suboperons (Table 2.5 and supplementary file 5) with low expression levels,

Table 2.5. Reconstruction of the alternative *phn* operons

Sample	Operons/suboperons
LB (OD=0.87)	*phnC, phnD, phnH, phnK, phnL, phnM, phnNOP*
HS15min	*phnCD, phnE, phnGHIJK, phnMNOP*
HS30min	*phnC, phnDE, phnGH, phnI, phnJ,phnK, phnL, phnM, phnNOP*
HS60min	*phnC, phnD, phnG, phnH, phnK, phnNOP*
M-P0h	*phnCDE, phnFGH, phnI, phnJK, phnLMNOP*
M-P2h	*phnCDEFGHIJKLMNOP*
M-P4h	*phnCDEFGHIJKLMNOP*

whereas under phosphorus starvation (samples M-P2h and M-P4h), the *phn* genes were transcribed to form the large operon *phnCDEFGHIJKLMNOP* with high expression levels (Figure 2.18 and supplementary file 5), which is consistent with previous observations [142]. In fact, this 14-gene operon and its suboperons have been studied previously by several groups [142-145]. It is now known that *phnCDE* encodes a phosphonate transport system, which was detected in sample M-P0h, and *phnF* works as a repressor for this suboperon [146]. Moreover, *phnGHIJKLM* is essential for the C-P bond cleaving activity [147]. More recently, Jochimsen *et.al* have shown that *phnGHIJK* encodes a protein complex essential for organophosphonate utilization [145], this suboperon was detected in sample HS15min. Furthermore, genes *phnNP* function as downstream processing enzymes [148], whereas the *phnO* gene is unnecessary for transport or catalysis, and may therefore have a regulatory role [147]. Finally, as shown in Figure 2.18, the *phnCDEFGHIJKLMNOP* operon displayed varying/decreasing expression levels along the operon, another form of the complexity of prokaryotic

transcriptomes in addition to alternative operon utilization [10]. However, further investigation of this phenomenon is out of the scope of this work.

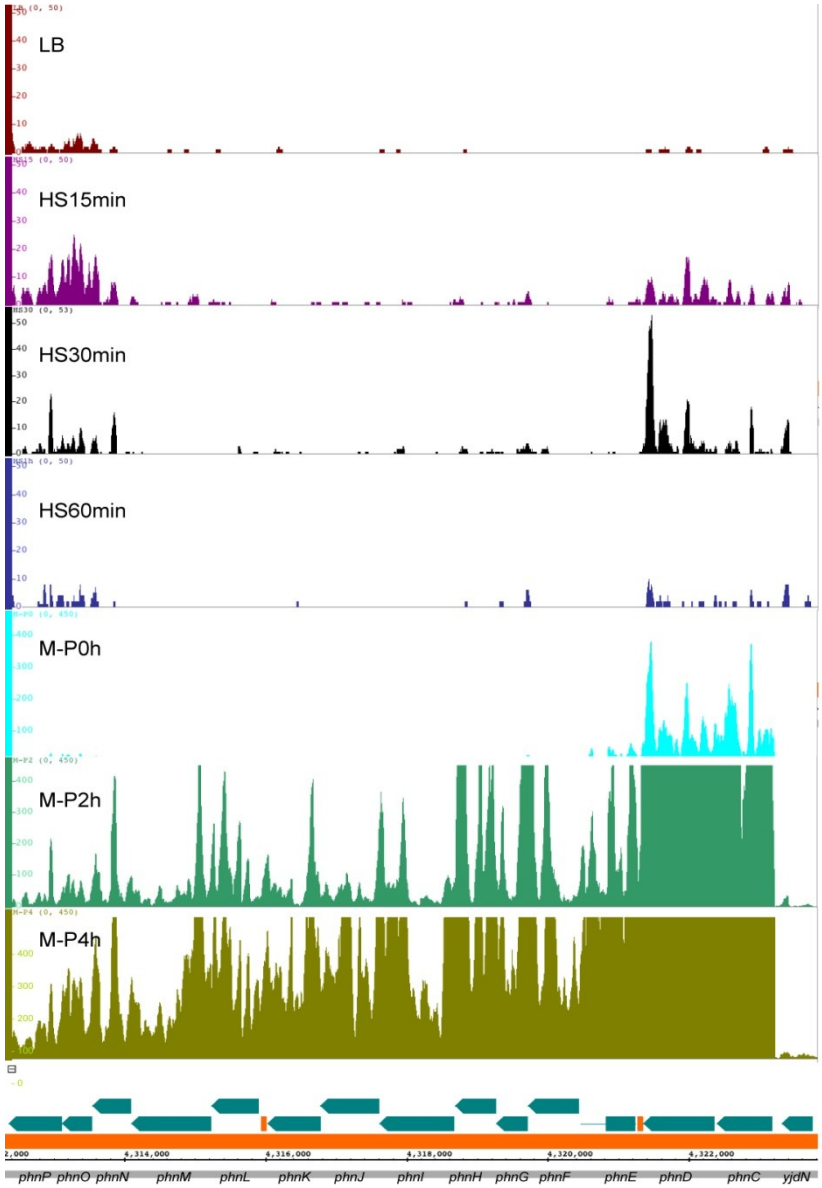


Figure 2.18. Reads coverage of the genes in the phn operon. The vertical axis is the number of reads covered at the position. The orange and dark green bars represent the forward and reverse strands, respectively. Segments with arrows represent genes. Genes from the right to left are *yjdN, phnC, phnD, phnE, phnF, phnG, phnH, phnI, phnJ, phnK, phnL, phnM, phnN, phnO and phnP*. The graphs were generated using IGB. To make the expression levels for the 14 genes in different samples visible and comparable, the same vertical axis scale (50) is used for the LB and HS treatments, and the same vertical axis

Figure 2.18 (continued) scale (450) is used for M-P treatments. Some positions with low read coverage cannot be shown while some other positions with high coverage are truncated. Note the varying levels of coverage and gaps along the operon under different cultures and growth phases, and again the similar position-dependent non-uniform coverage of the reads along the operon.

Another interesting example in our samples is the alternative utilization of the 13-gene operon *fliFGHIJKLMNOPQR* encoding proteins in the flagella of *E. coli K12* (Table 2.6 and supplementary file 5). Although the *fli* operon was expressed as a 13-gene polycistron in sample LB, it was split into short suboperons under the treatments of heat shock or phosphorus starvation in a time dependent manner (Table 2.6). For example, at

Table 2.6. Reconstruction of alternative *fli* operons

Sample	Operons/suboperons
LB (OD=0.87)	*fliFGHIJKLMNOPQR*
HS15min	*fliFGHIJK, fliLMN, fliOPQ, fliR*
HS30min	*fliFGH, fliI, fliJKL, fliMN, fliO, fliQ, fliR*
HS60min	*fliGH, fliI, fliK, fliL, fliM, fliQ*
M-P0h	*fliFGHIJKLMN, fliOPQR*
M-P2h	*fliFGHIJKL, fliMN, fliOP, fliQR*
M-P4h	*fliFGHIJKLMNO, fliPQR*

the beginning of heat shock (sample HS15min), the *fli* operon was divided into four suboperons, it then was further split into six to seven suboperons (samples HS30min and HS60min). Interestingly, it has been shown that heat shock reduces bacterial mobility possibly through the regulatory interactions between the heat shock system and the flagellum/chemotaxis system [149]. Moreover, it has been indicated that inorganic phosphorus is necessary for the motility of bacteria [150]. However, the underlying mechanisms of these observations are largely unknown. Therefore, our results might provide a possible molecular explanation of these earlier observations: the extreme

conditions (heat shock/phosphorus starvation) alter the expression of flagella proteins by changing the patterns of alternative usages of the *fli* operon, thus influence the motility of the bacterial cells.

2.4.6 Verification of Hypothetical Genes

Although *E. coli* K12 is probably the best studied and understood model organism, researchers have not completely defined even its coding genes. For instance, there are still 36 sequences labelled as hypothetical protein genes as of this writing in the RegulonDB [60]. Interestingly, we found that all these 36 hypothetical genes were transcribed in at least one of our seven samples (Supplementary file 9), and 21 (b0050, b0137, b1356, b1382, b1419, b1446, b1457, b1607, b1952, b1998, b3471, b3638, b3937, b4325, b4335, b4336, b4593, b4596, b4610, b4615 and b4620) of them were expressed in all the samples, suggesting they are highly likely to be true protein coding genes. Furthermore, 20 of them formed multi-gene operons with other known genes (Supplementary file 9). The functions of these known genes might provide hints to possible functions of the associated hypothetical genes for "guilt by association".

2.5 Conclusion

In this chapter we present a HMM-based assembly algorithm, TruHmm, for assembling prokaryotic transcriptomes. TruHmm is specifically designed to address the missing reads problem in the current RNA-seq library preparation procedures and short reads NGS technologies. By using a sliding window, TruHmm relieves the negative effect of gaps in the transcribed region on assemble, and therefore enhances its power to stitch broken segments. When tested on 7 RNA-seq datasets from *E. coli* K12, TruHmm has achieved very high performance measured by five metrics for assembling known

operons in the bacterium. Furthermore, TruHmm is able to detect alternative operon utilizations under different conditions. We hope that our programs will be useful for decoding the dynamics and complexity of transcriptomes in prokaryotes.

CHAPTER 3: PREVALENT ANTISENSE TRANSCRIPTS MODULATE GENE TRANSCRIPTION IN PROKARYOTES IN A CONDITION-DEPENDENT WAY

3.1. Abstract

It has been recently shown that antisense transcription in prokaryotes is more prevalent than originally anticipated, and the resulting antisense transcripts may play critical roles in the regulation of the activity of the cognate genes. However, little is known about the molecular mechanisms and patterns of antisense transcriptions of different genes under different growth phases and environmental conditions. In this chapter, we determined the transcriptome of *E. coli* K12 at different growth phases and four different culture conditions using directional RNA-seq technique and the TruHmm tool developed in Chapter 2. We found that 0.5~29% of the genome had transcripts from both the forward and reverse strands, and 13~87% of transcribed ORFs had at least one antisense transcript, dependent on growth phases and culture conditions. ORFs could have six different modes of transcription in a growth phase and culture condition dependent manner: sense only, sense dominant, equal transcription, antisense dominant, antisense only and silent modes. Almost all transcribed genes in our dataset changed their transcription modes between different growth phases and culture conditions, except for some housekeeping genes. Moreover, we found that antisense transcriptions can be initiated anywhere along an ORF, but strongly biased and restricted to the 5'end the ORF, giving hints to the possible mechanisms of regulation by antisense transcripts. Therefore, antisense transcription is

very prevalent in *E. coli* K12, and may play important roles in various aspects of the bacterium's physiology.

3.2 Introduction

Bacterial transcriptomes have long been considered to be composed of mRNAs, rRNAs, tRNAs and some small RNAs. In the past few years, applications of high density directional tiling array and in particular, directional RNA-seq techniques to prokaryotic transcriptome profiling have revealed prevalent transcription from the reverse complementary strand of protein coding genes, resulting in antisense RNAs (asRNAs) with a length from tens of nucleotides to thousands of nucleotides [10, 11, 16, 109, 132, 151-154]. asRNAs can overlap the 5' end, the 3' end, the middle, or the entire gene on the opposite strand. It has been proposed that asRNA can affect the gene expression by several different mechanisms [83]: 1) An asRNA can induce transcriptional interference, as the transcription from the antisense promoter blocks the transcription from the promoter of the gene. 2) An asRNA can induce transcriptional attenuation of the gene encoded on the opposite strand by changing the target mRNA's structure, resulting in premature transcription termination. 3) The duplex substrate generated by base pairing between the sense and antisense transcripts can either promote or block RNA degradation by directly generating or blocking a ribonuclease target site, or indirectly change the binding position of the ribonulease. 4) An asRNA can block translation of mRNA either by directly binding to ribosome binding position or indirectly block ribosome binding by impacting the target mRNA structure. All these regulatory mechanisms were proposed based on specific sense-antisense partners in the specific cases in different prokaryotic

species and strains, however, the prevalent antisense transcription strongly suggest that asRNAs might play important roles in the physiology of prokaryotes.

E. coli K12 is probably the best known free living model organism [155, 156], where novel biological hypotheses and computational algorithms can be tested. Indeed, it is mainly through the studies in *E. coli* K12 that we have understood many fundamental biological processes, including the mechanisms of gene transcription regulation [157-159]. Extensive anti-senses and non-coding transcription have been experimentally identified in the *E. coli* [15, 154, 160], however, the levels of their prevalence can vary quite differently from different studies. For instance, Selinger et al. [15] reported that up to 4,000 *E. coli* K12 genes had anti-sense transcription based on a high-resolution "genome array", while using RNA-seq technique, Dornenburg et al. [154] only identified about 1,000 asRNA in the same genome under similar growth conditions. Therefore, how many antisense RNAs are there in the genome? What are the patterns of the antisense transcription under different culture conditions? What are their functions? To answer these questions, we applied directional RNA-seq technique to study the patterns of antisense transcription in *E. coli* K12 under a variety of culture conditions and growth phases.

3.3 Materials and methods

3.3.1 Bacteria culture

A frozen stock of *Escherichia coli* K12 strain MG1655 was thawed, inoculated in LB medium in a test tube by 1:100 dilution and cultured overnight at 37 ^{o}C and 250 rpm. The cells were then transferred to fresh LB medium in a flask by 1:100 dilutions, and cultured at 37 ^{o}C and 250 rpm. When the cells grew to the log phase with an optical density at 610

nm [OD_{610}] of 1.0, they were spun down at 3,200g for 25 min. For MOPS treatment (MOPS), the cell pellets were resuspended in the same volume of MOPS medium (100 ml of 10X MOPS mixture, 880ml of sterile H_2O, 10ml (0.132M) KH_2PO_4 and 10ml of 20% glucose, Teknova, Hollister, CA). For heat shock treatment (HS), the cell pellets were resuspended in the same volume of MOPS medium, and incubated at 48^oC and 250 rpm. For carbon-starvation treatment (M-C), the cell pellets were resuspended in the MOPS medium without KH_2PO4. Three milliliter cell suspension were collected in a tube containing 1.5ml RNA Later (Invitrogen) immediately after the cell pellets were resuspended in the indicated medium (0 min) and at the indicated time points thereafter (HS:15min, 30min, 1h, 2hrs, 4hrs; MOPS: 1hr, 2hrs, 4hrs, 6hrs; M-C: 1hr, 2hrs, 4hrs, 6hrs). Cells were spun down at 6,000g, 8 min and -4 ^oC, and the pellets were resuspended in 1.5 ml of *RNAlater*. The samples were stored at -80 ^oC until use.

3.3.2 Isolation and enrichment of mRNA

RNA was isolated using RiboPure^{TM} -Bacteria Kit (Ambion) following the manufacturer's instructions. Once isolated, ~10□ g total RNA was treated with 8 units DNase (Invitrogen) twice to remove genomic DNA, and the complete removal of DNA was confirmed by 35 cycles PCR amplification of a 196 bps fragment of the *crp* gene (5'-primer:AGCATATTTCGGCAATCCAG;3'-primer:TACAGCGTTTCCGCTTTTTC). rRNAs were removed from the total RNA using a MICROBExpress kit (Ambion) to enrich mRNAs.

3.3.3 Construction of directional RNA-seq libraries

In our earlier experiments, sequencing was done on an Illumina GAII platform at the sequencing core facility of the University of North Carolina at Chapel Hill, and the

directional RNA-seq libraries were constructed by following an Illumina's instruction using their Small RNA Sample Prep Kit with some modifications. Briefly, after the purified mRNA was fragmented using a RNA fragmentation kit (Ambion), the fragmented RNA was treated with Antarctic phosphatase (NEB) to remove the 5'-tri-phosphate groups of RNAs with an intact 5'-end. A mono-phosphate group was then added back to the 5'-end of RNAs by polynucleotide kinase (PNK, NEB) in the presence of 10 mM ATP. The v1.5 sRNA 3' Adaptor (5'/5rApp/ ATCTCGTATGCCGTCTTCTGCTTG /3ddC/) was ligated to the 3'-end of fragmented RNAs using truncated T4 ligase 2 (NEB), and the SRA 5' RNA adaptor (5'GUUCAGAGUUCUACAGUCCGACGAUC) was ligated to the 5'-end of fragmented RNAs using T4 ligase. To preserve short inserts from small RNAs we omitted the size selection step after PCR application of inserts. For our later experiments, sequencing was done on an Illumina HiSeq 2000 platform at David H. Murdock Research Institute of the North Carolina Research Campus (Kannapolis, NC), and we constructed the directional RNA-seq libraries using Illumina's TruSeq Small RNA Sample Prep Kit, so that multiplex sequencing can be achieved by using the barcoded PCR primers. The details of the method will be described elsewhere (Dong, Li and Su). Briefly, after similar treatments as described above, the 5' Adapter (RA5: 5' GUUCAGAGUUCUACAGUCCGACGAUC), and 3' Adapter (RA3: 5' TGGAATTCTCGGGTGCCAAGG) were ligated to 5'- and 3'-end of fragmented RNAs, respectively. Reverse transcription-PCR (RT-PCR) was performed using SuperScript II Reverse Transcriptase Kit using the SRA RT primer, followed by 16 cycles of PCR amplification. Again, the size selection was omitted on PCR products to preserve short

inserts from possible small RNAs. Single-end sequencing on the Illumina GA II platform was done with 76 cycles, while that on the HiSeq 2000 platform was done with 100 cycles. Some samples (M-C1h and M-C2h) were sequenced on both platforms.

3.3.4 Reads mapping and statistical analysis

The genome sequence of *E. coli* K12 substr. MG1655 was obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/). The gene annotation file and the experimentally verified operons in the bacterium were downloaded from RegulonDB [60] (http://regulondb.ccg.unam.mx/). A total of 4501 annotated genes (also including pseudo genes and non-coding small RNAs) are included in this analysis. As the reads were not size-selected during the library construction, we trimmed the 3' adapters attached to some short insertions. Adapter-free reads with lengths of <10 nucleotides (nts) were discarded; the remaining reads were mapped to the *E. coli* K12 genome using Bowtie [125]. For the reads of length 10-14, 15-29 and ≥30 nts, up to 1, 2, and 3 mismatches were allowed, respectively. Only uniquely mapped reads were used for further analysis. The alignment of mapped reads to the reference genome was visualized by Integrated Genome Browser (IGB) [133]. We used the assembly tool TruHmm we developed in Chapter 2 to assemble transcripts in each sample with a window size = 11nt. Different from operon assembly, we omitted the stitching step (for detail please refer to chapter 2) to reconstruct the ncRNAs or asRNAs. In addition, we used DAVID [161] to analyze functional annotation enrichment for the groups of genes.

3.4 Results and discussion

3.4.1 Our RNA-seq reads are of high quality

We prepared the directional RNA-seq libraries from 16 *E. coli* K12 samples collected

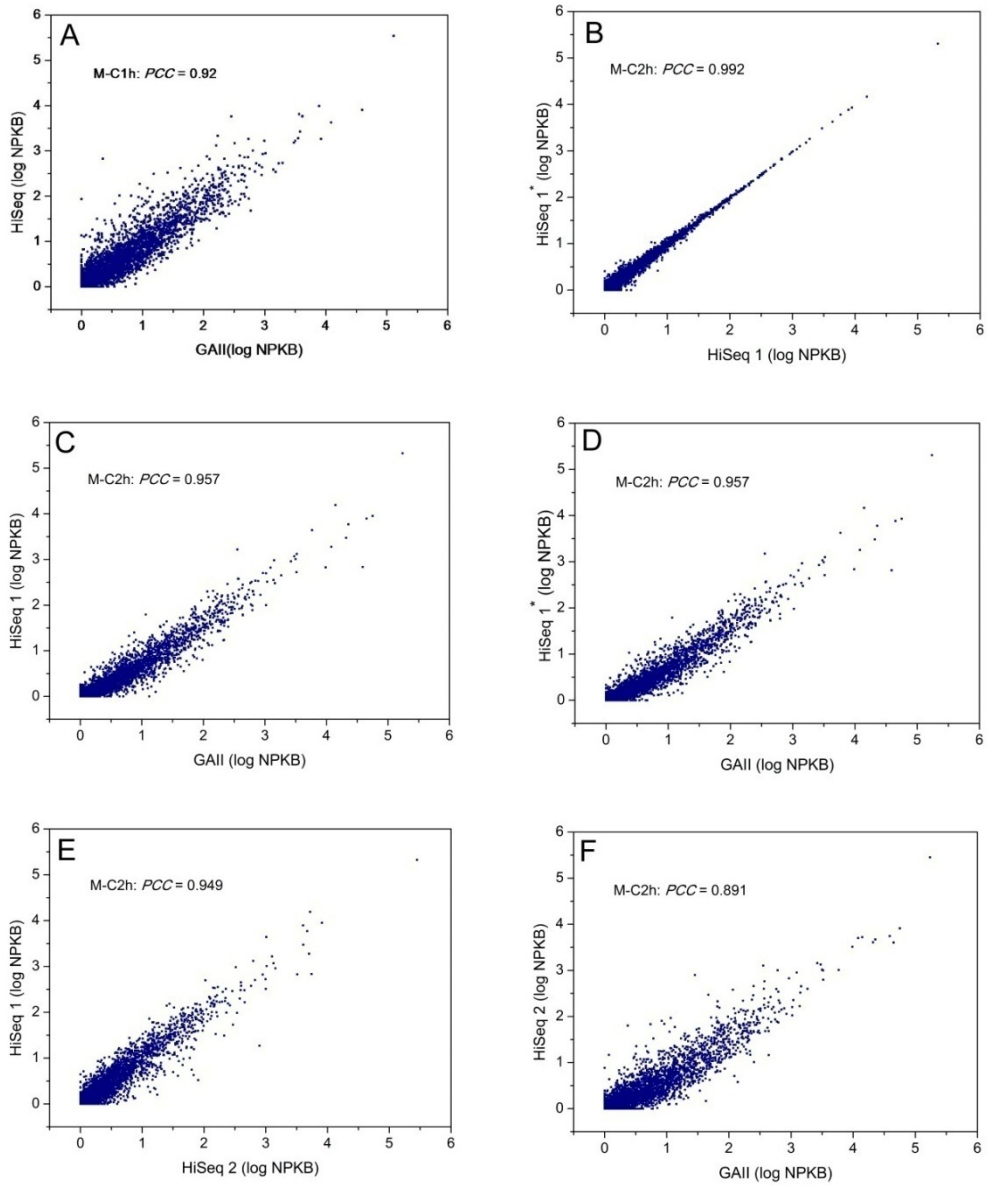


Figure 3.1. Correlation of expression levels of genes between any two replicates for samples M-C1h and M-C2h. Each dot represents a gene. The expression level is evaluated using log of the NPKB values. A) Correlation of expression levels for M-C1h between GAII reads and HiSeq reads. B) Correlation of expression levels for M-C2h between two technical replicates sequenced on HiSeq 2000 platform, HiSeq1 reads and HiSeq1* reads. C) Correlation of expression levels for M-C2h between two biological replicates sequenced on GAII reads and HiSeq1 reads, respectively. D) Correlation of expression levels for M-C2h between GAII reads and HiSeq1* reads. E) Correlation of expression levels for M-C2h between HiSeq1 reads and HiSeq2 reads. F) Correlation of expression levels for M-C2h between HiSeq2 reads and GAII reads.

at the log phase growth in LB, and different time points under MOPS (MOPS), heat shock (HS) or carbon starvation (M-C) treatments, denoted as LB0.5, LB1.0, LB3.0,

Table 3.1 Summary of mapping results.

Sample	Platform	Total reads	Uniquely mapped reads	Multipe mapped reads	Reads failed to map	Unique (%)	Multiple (%)	Failed (%)
LB0.5	HiSeq	35,456,265	3,141,933	22,113,900	10,200,432	8.86	62.37	28.77
LB1.0	HiSeq	44,278,709	4,753,962	26,356,555	13,168,192	10.74	59.52	29.74
LB3.0	HiSeq	39,089,273	3,400,925	32,077,448	3,610,900	8.70	82.06	9.24
HS15min	HiSeq	39,815,593	6,515,460	29,989,690	3,310,443	16.36	75.32	8.31
HS30min	HiSeq	34,593,476	5,257,054	25,654,089	3,682,333	15.20	74.16	10.64
HS1h	HiSeq	43,645,193	6,602,284	33,886,728	3,156,181	15.13	77.64	7.23
HS2h	HiSeq	38,782,211	4,978,191	30,815,704	2,988,316	12.84	79.46	7.71
HS4h	HiSeq	43,139,317	3,941,623	35,707,566	3,490,128	9.14	82.77	8.09
MOPS1h	HiSeq	28,239,285	4,991,358	21,423,634	1,824,293	17.68	75.86	6.46
MOPS2h	HiSeq	34,690,431	5,654,106	26,509,249	2,527,076	16.30	76.42	7.28
MOPS4h	HiSeq	32,390,937	6,349,465	23,724,159	2,317,313	19.60	73.24	7.15
MOPS6h	HiSeq	42,791,495	3,351,297	36,182,059	3,258,139	7.83	84.55	7.61
M-C1h	GAII + HiSeq	53,570,359	3,838,414	44,973,911	4,758,034	7.17	83.95	8.88
M-C2h	GAII + HiSeq	52,475,132	2,828,573	45,106,071	4,540,488	5.39	85.96	8.65
M-C4h	HiSeq	49,469,000	2,770,743	41,676,190	5,022,067	5.60	84.25	10.15
M-C6h	HiSeq	48,903,212	2,410,973	42,052,434	4,439,805	4.93	85.99	9.08

MOPS1h, MOPS2h, MOPS4h, MOPS6h, HS15min, HS30min, HS1h, HS2h, HS4h, M-C1h, M-C2h, M-C4h and M-C6h to reflect the treatment and sampling time points. The experimental procedure of our work is listed in Figure 2.7, chapter 2. The libraries were sequenced on either Illumina GA II or HiSeq 2000 platforms. Specifically, all the samples were sequenced using the HiSeq 2000 platform, except that samples M-C1h and M-C2h were sequenced using both HiSeq 2000 and GAII platforms. M-C1h has two biological replicates, and M-C2h has four biological/technical replicates, one was sequenced on GAII platform, and one was sequenced twice on HiSeq 2000 platform, the other one was sequenced on HiSeq 2000 platform. The reads obtained from different platforms for the same sample are highly correlated (Figure 3.1), thus the data for the same sample were

combined for the analysis. A total of 661,329,888 reads were generated from the 16 samples. The mapping statistics of the samples are summarized in Table 3.1 showing that 4.93~19.6% of reads could be uniquely mapped to the genome, resulting in 2,410,973 ~ 6,602,284 uniquely mapped reads in each sample, corresponding to 52~142 times

Table 3.2. Distribution of mapped nucleotides on coding (sense and antisense) and intergenic regions

Sample	Total nt counts	Sense nt counts	Antisense nt counts	Sense %	Antisense %	Intergenic %
LB0.5	1.69E+08	1.53E+08	1.93E+06	90.29%	1.14%	8.56%
LB1.0	3.25E+08	2.94E+08	2.98E+06	90.29%	0.92%	8.79%
LB3.0	1.99E+08	1.85E+08	1.28E+06	93.23%	0.64%	6.13%
HS15min	4.28E+08	3.87E+08	3.78E+06	90.21%	0.88%	8.91%
HS30min	3.00E+08	2.69E+08	3.15E+06	89.51%	1.05%	9.44%
HS1h	3.54E+08	3.08E+08	5.39E+06	87.02%	1.52%	11.46%
HS2h	2.94E+08	2.58E+08	4.58E+06	87.63%	1.56%	10.81%
HS4h	1.95E+08	1.62E+08	4.42E+06	83.24%	2.27%	14.49%
MOPS1h	2.63E+08	2.43E+08	1.81E+06	92.10%	0.69%	7.21%
MOPS2h	3.54E+08	3.25E+08	2.07E+06	91.94%	0.58%	7.47%
MOPS4h	4.07E+08	3.76E+08	2.91E+06	92.33%	0.71%	6.96%
MOPS6h	2.00E+08	1.91E+08	986918	95.48%	0.49%	4.03%
M-C1h	3.37E+08	3.19E+08	995483	94.87%	0.30%	4.83%
M-C2h	5.57E+08	5.36E+08	1.30E+06	96.29%	0.23%	3.48%
M-C4h	1.17E+08	1.16E+08	159837	98.60%	0.14%	1.27%
M-C6h	1.01E+08	9.94E+07	155885	98.67%	0.15%	1.17%

coverage of the genome. Of the 59.52~85.99% multiple mapped reads in each sample, over 99.8% were from duplicated tRNA/rRNA genes (data not shown). Furthermore, as shown in Table 3.2, in all the samples over 90% and less than 15% of the total mapped nucleotides were mapped to the sense strand and intergenic regions, respectively, with only 0.14~2.27% of the total mapped nucleotides mapped to the antisense strand. Therefore, as we discussed in chapter2, these results indicate that most of our reads were from the sense strand, and thus our libraries were highly strand specific.

Moreover, as the datasets we analyzed in chapter 2 (Figure 2.10A), the transcriptomes in this RNA-seq dataset also have the same gap-problem (Figure 3.2) caused by the loss of some RNA fragments during the library preparation due to the highly labile nature of prokaryotic RNAs as well as other technical artifacts [107, 111, 162]. Less than 50%

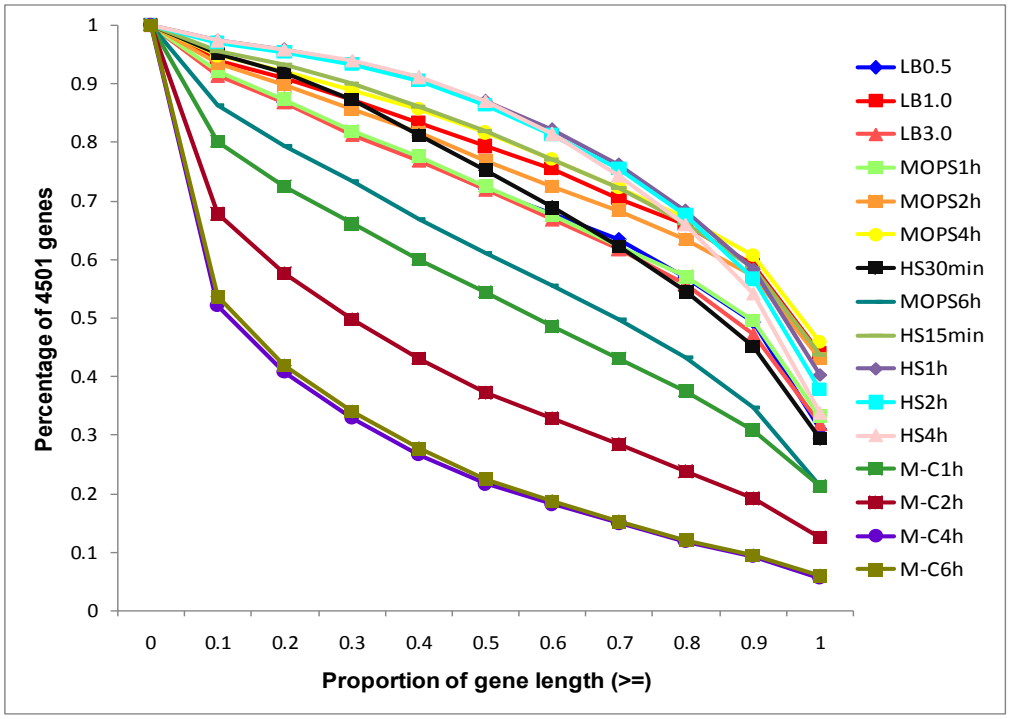


Figure 3.2 Distribution of the genes with more than the indicated percentage of their length covered by at least one read in the samples: Less than 50% of genes have their length completely covered by at least one read. Over 75% genes have over 50% of their length covered by at least one read except for samples MOPS6h and M-C treatment.

genes in the genome had their length completely covered by at least one read, while only less than 90% genes in the genome had at least 10% of their length covered by at least one read (Figure 3.2). Most of the samples except those of M-C treatment and MOPS6h have over 75% genes with ≥ 50% of the length covered by at least one read (Figure 3.2). The poor reads coverage on the gene-coding region is probably caused by the fact that the

cells stopped growing and began to die without carbon source as well as in the late stage of MOPS treatment when nutrients were exhausted. The reason will be discussed in detail later. On the other hand, as shown in Figure 3.3, our uniquely mapped reads consisted of well-balanced different sizes of RNA fragments.

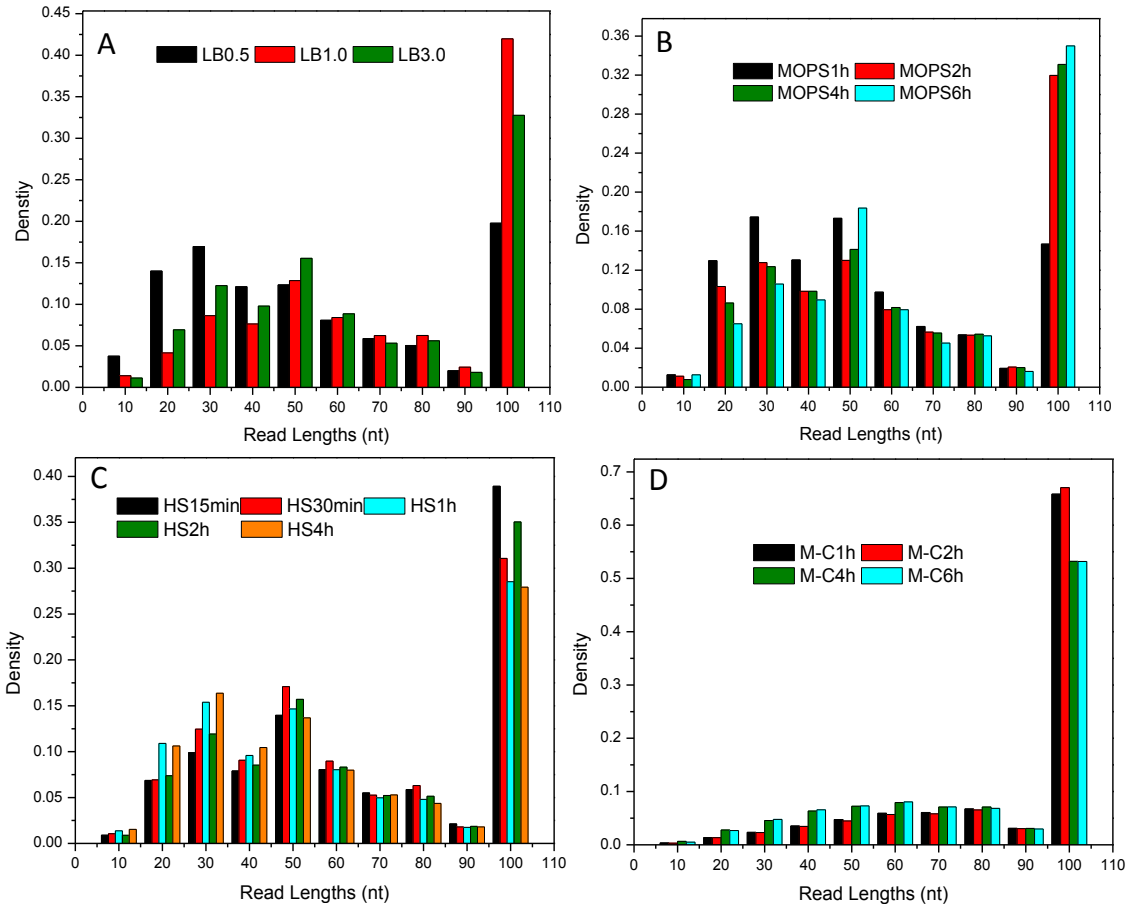


Figure 3.3. Distribution of the length of uniquely mapped reads in the samples: A) LB ; B) MOPS ; C) HS ; D) M-C, indicating that our library preparation protocol could potentially capture small RNA species such as asRNA and ncRNA, which were otherwise left out by a typical size selection step in the library preparation process. The data have been submitted to NCBI SRA database with accession number XXX.

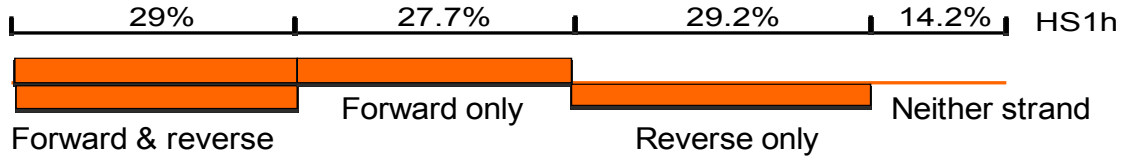
3.4.2 Antisense transcription is pervasive in *E. coli* K12

To see the source of transcription in the *E. coli* genome under different culture conditions and time points, we counted the percentage of positions of the genome from

which both the forward and reverse strands (forward & reverse), only the forward strand (forward only), only the reverse strand (reverse only) or neither strand was transcribed, respectively. Interestingly, as shown in Table 3.3, except for the samples in the late

Table 3.3. Summary of the coverage of the *E. coli* K12 genome by uniquely mapped reads on both the forward and reverse strands, on only one of the strands, or with no coverage, respectively. The cartoon shows the case of HS1h.

Sample	Forward & reverse	Coding region of both strand	Forward only	Coding region of forward only	Reverse only	Coding region of reverse only	Neither strand	Coding region of neither strand
LB0.5	11.1%	85.8%	30.7%	85.0%	32.3%	85.8%	25.9%	82.2%
LB1.0	16.3%	84.8%	31.6%	85.4%	33.2%	86.1%	18.8%	81.4%
LB3.0	8.7%	82.6%	30.4%	85.9%	32.9%	86.3%	28.0%	83.6%
HS15min	20.7%	85.3%	30.2%	83.4%	32.1%	83.8%	17.0%	81.2%
HS30min	17.0%	82.3%	28.9%	79.9%	30.8%	80.8%	23.3%	86.7%
HS1h	29.0%	83.7%	27.7%	80.1%	29.2%	81.5%	14.2%	85.9%
HS2h	26.7%	83.8%	28.8%	80.2%	29.9%	81.6%	14.7%	85.7%
HS4h	25.3%	83.3%	29.0%	80.7%	30.2%	81.5%	15.5%	86.5%
MOPS1h	11.1%	85.9%	30.3%	85.1%	32.2%	85.7%	26.4%	81.8%
MOPS2h	14.3%	85.3%	30.9%	85.1%	33.3%	85.9%	21.5%	81.9%
MOPS4h	18.3%	86.0%	31.0%	84.8%	32.9%	85.7%	17.9%	79.9%
MOPS6h	5.8%	84.5%	27.2%	84.6%	29.0%	85.2%	38.0%	85.0%
M-C1h	6.0%	82.1%	28.1%	85.4%	30.8%	86.0%	35.1%	84.9%
M-C2h	6.6%	85.5%	27.2%	84.2%	29.8%	84.6%	36.4%	84.6%
M-C4h	0.5%	75.1%	11.6%	83.6%	12.7%	83.5%	75.2%	87.9%
M-C6h	0.6%	75.7%	11.9%	84.0%	13.0%	84.6%	74.5%	87.8%



phases of M-C treatment (M-C4h and M-C6h) when cells were dying, the proportion of the genome that was transcribed from both strands was highly variable under different conditions and time points (from 5.8% in MOPS6h to 29.0% in HS1h), while those that were transcribed either from forward strand only or reverse strand only strand were less

variable and were quite balanced between the two strands (from 27.2% and 29.0% in MOPS6h to 31.6% and 33.2% in LB). Furthermore, of the positions where both strand were transcribed, around 80% are in gene-coding regions in all the 16 samples, indicating that coding regions are more likely to have transcription from both strands. Thus the resulting asRNAs might be related to the function of sense transcripts, in agreement with the early studies [84, 163]. This conclusion is also supported the condition and time dependent changes in the proportion of "forward & reverse" transcription (Table 3.3). For example, HS generally induced higher proportions of transcription of both strands at all the time points measured (17~29%), while C-starvation resulted in lower proportions of both strands transcriptions (0.5~6%), which decreased with the increase in incubation time. Moreover, samples taken at the latest time points of all the four different treatments have lower proportions of both strands transcription but highest proportion of neither strand transcription, suggesting that when the transcription activity decreased as the nutrients were exhausted and the cells became aging, both-strand transcription also decreased.

3.4.3. Modes of sense and antisense transcriptions

In principle, given an ORF on a locus of the chromosome, there are four possible transcription modes associated with it: 1) the ORF is transcribed but there is no associated antisense transcription (sense-only); 2) the ORF is not transcribed but there is associated antisense transcription (antisense-only), 3) the ORF is transcribed and there is also associated antisense transcription (sense & antisense), and 4) there is no transcription on neither sense nor antisense (silent). We first examined the ORFs with the sense & antisense pattern in the 16 samples, to see whether or not there is a correlation between

the transcription levels of such ORFs and those of their asRNAs. As shown in Figure 3.4, the levels of sense and antisense transcripts of the ORFs are highly correlated in most samples except for M-C4h and M-C6h, which behaved quite differently from the others

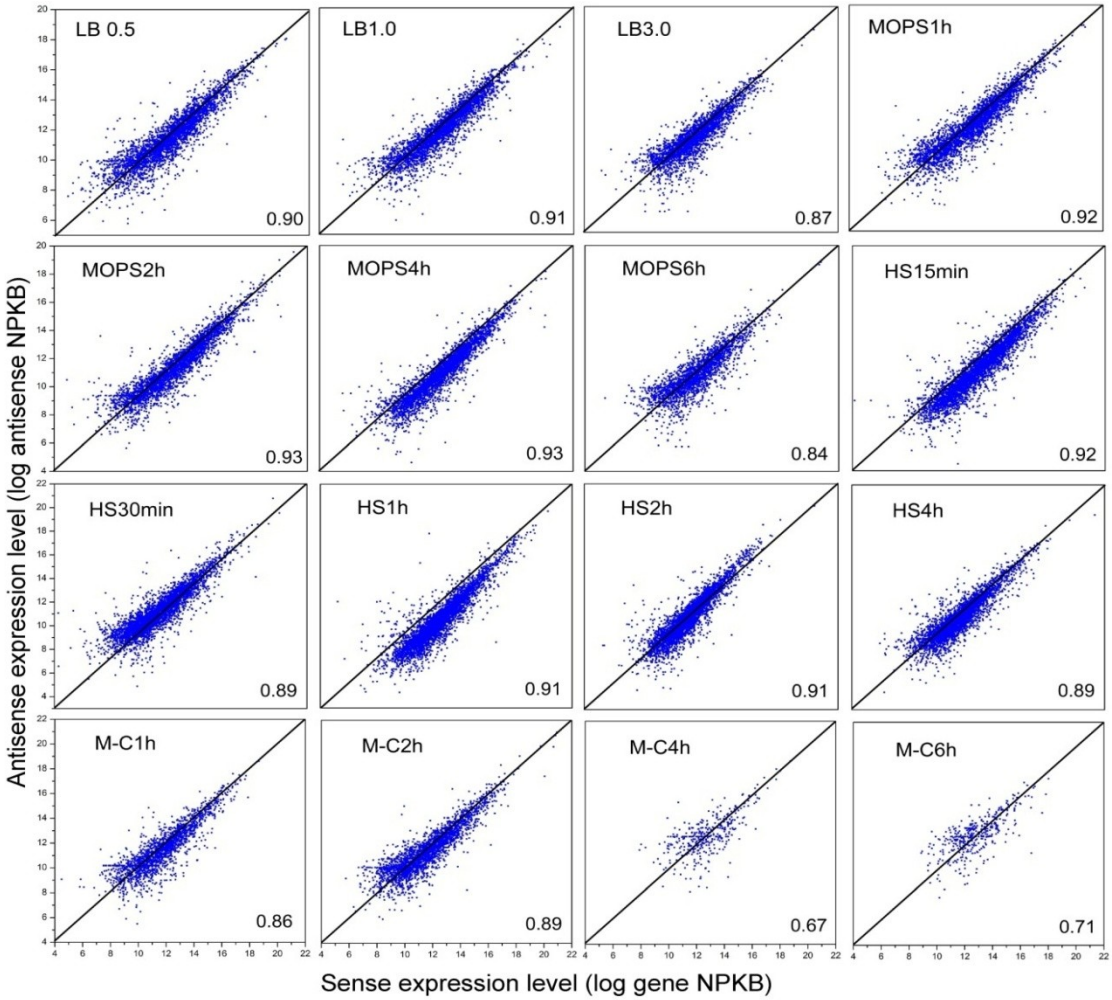


Figure 3.4. Correlations between the expression levels of sense and antisense transcripts in all 16 samples. The plots show the dependence of the averaged antisense expression level vs. the sense expression level ($\log_2$ of NPKB value). The averaged antisense expression level is used here considering that one ORF tends to have multiple antisense transcripts on the reverse complementary strand. The line y=x aims to sort out the stronger signal from the sense and antisense transcripts. The number in the bottom right corner is the Pearson correlation coefficient (PCC). Only the genes transcribed with antisense transcription are included in the plot.

probably because the cells were dying due to the lack of the most demanding element carbon. There were also notable different biases to the sense transcription level in different samples (Figure 3.4). As asRNAs are likely to execute their functions by forming complementary duplexes with their sense transcripts, to further investigate the

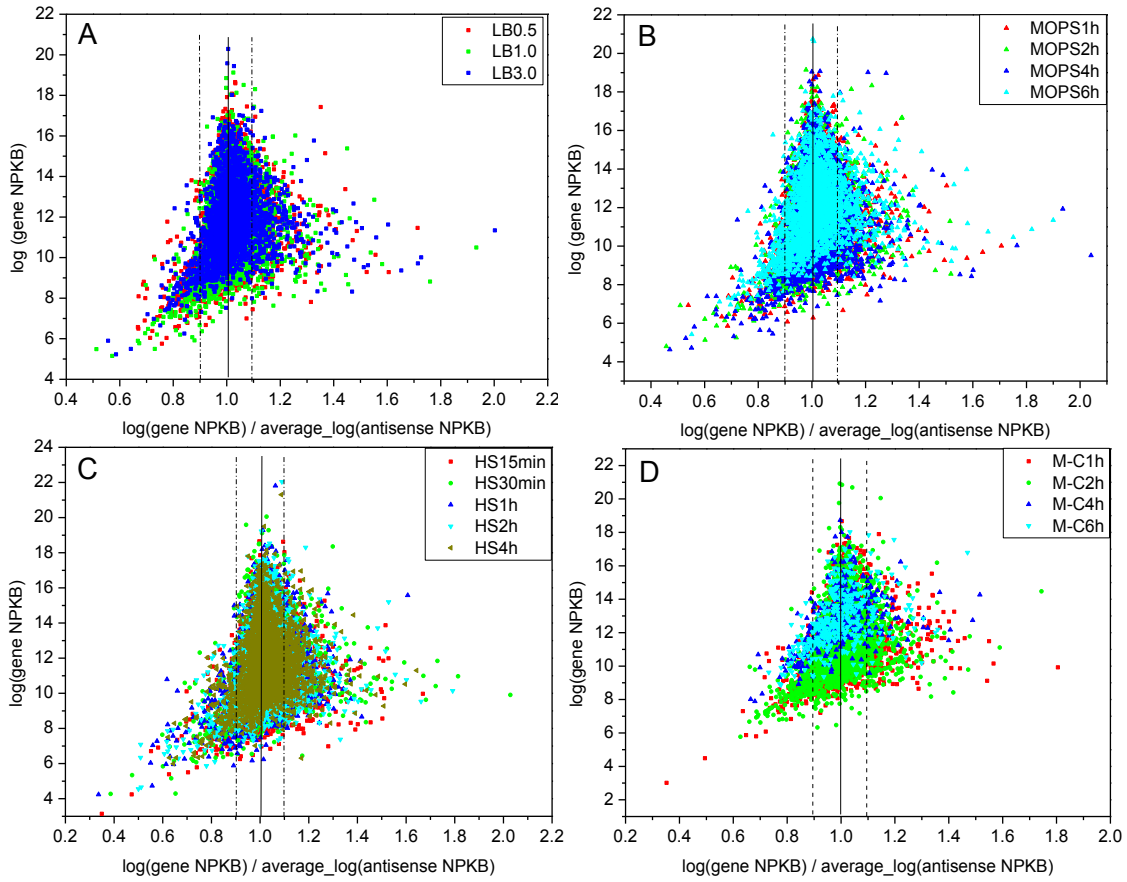


Figure 3.5 Expression levels of genes against the ratio (γ) of gene/antisense transcripts. A) LB treatment. B) MOPS treatment. C) HS treatment. D) M-C treatment. The genes in the plot are the expressed ones with antisense transcription. The ratio of the expression levels of genes over their antisense transcripts are centered at around 1.02 in all the samples, except samples M-C4h and M-C6h centered at γ =0.99. The interval [0.9, 1.1] composed by the two dashed line contains the majority of genes in each sample.

relationship between sense and antisense transcription, we plotted the transcription level of an ORF as a function of the ratio (γ) of the transcription level to the average

transcription levels of its asRNAs (an ORF could have multiple asRNAs). As shown in Figure 3.5, in all the four cultural conditions, the transcription level and the ratio γ showed a very similar but complex triangular relationship. When the ratio γ was in a narrow interval around 1, the transcription level of an ORF could vary dramatically without a significant change of the ratio (the upper angle), most ORFs fell in this regime; when the ratio γ is lower than 1, the transcription level of an ORF was positively correlated with the ratio (left angle); when the ratio γ is higher than 1, the transcription level of an ORF might not change significantly with a dramatic change of the ratio γ (the right angle). Relatively few ORFs fell in the latter two regimes. They might represent three different modes of antisense transcription and thus sub-transcription modes of the sense & and antisense mode. Based on the distribution of the γ values of the ORFs as shown in Figure 3.6 and the above observations in Figure 3.5, we further divided the

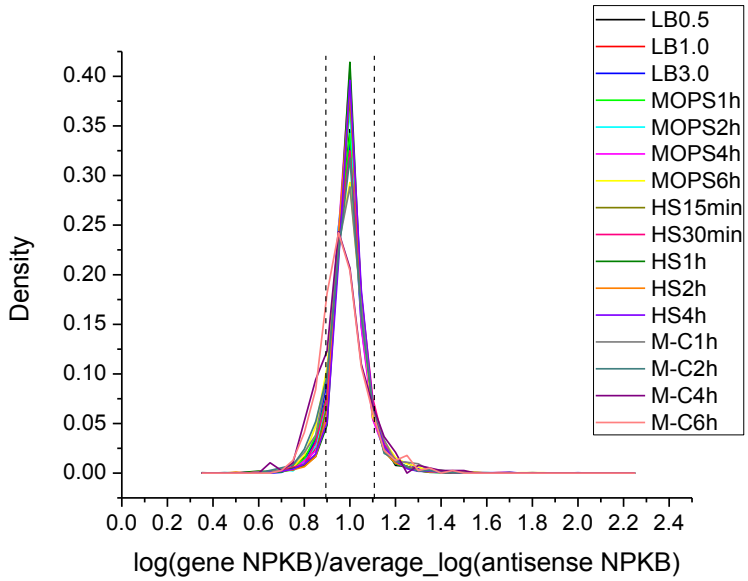


Figure 3.6. Distribution of ratio (γ) of ORF transcription level to the average transcription levels of its asRNAs.

sense & antisense transcription mode into three modes: sense dominant ($\gamma > 1.1$), equal transcription ($0.9 \leq \gamma \leq 1.1$), and antisense dominant ($\gamma < 0.9$). The interval [0.9, 1.1] was chosen because the standard deviation of ratio γ was around 0.1 in all the 16 samples (Figure 3.6). Thus we classify transcription of an ORF into a total of six modes, i.e., sense only, sense dominant, equal transcription, antisense dominant, antisense only and silent modes.

Table 3.4 summarizes the transcription modes of the ORFs in the genome in the 16 samples. As shown in the table, in all the samples except M-C1h, M-C4h, M-C6h, and MOPS6h, more than 60% of transcribed ORFs had at least one asRNA, and on average

Table 3.4 Summary of transcription modes of genes in the 16 samplers according to their sense and antisense transcription modes: sense only, sense dominant, equal transcription, antisense dominant, antisense only, and silent. The numbers in the parentheses are the number of antisense transcripts.

Sample	Sense only	Sense dominant	Equal transcription	Antisense dominant	Antisense only	Neither strand	Total antisense	Expressed genes	Gene with asRNA (%)	Average # asRNA per gene
LB0.5	1103	138 (1449)	3061 (5942)	24 (457)	10 (10)	165	7858	4326	74.50	1.82
LB1.0	1163	155 (1279)	3006 (6003)	155 (353)	10 (10)	149	7645	4479	74.03	1.71
LB3.0	1674	121 (842)	2466 (3933)	20 (291)	17 (17)	203	5083	4281	60.90	1.19
HS15min	827	185 (1942)	3344 (6927)	37 (549)	12 (13)	96	9413	4393	81.17	2.14
HS30min	751	200 (2302)	3422 (6829)	38 (762)	15 (15)	75	9908	4411	82.97	2.25
HS1h	547	165 (2686)	3693 (9248)	61 (773)	9 (9)	26	12716	4466	87.75	2.85
HS2h	664	150 (2337)	3587 (8273)	42 (629)	9 (9)	49	11248	4443	85.06	2.53
HS4h	632	177 (2410)	3595 (8209)	47 (605)	18 (18)	32	11242	4451	85.80	2.53
MOPS1h	1349	134 (1259)	2818 (5326)	21 (400)	8 (8)	171	6993	4322	68.79	1.62
MOPS2h	1261	133 (1275)	2932 (5841)	23 (395)	12 (12)	140	7523	4349	71.00	1.73
MOPS4h	1071	136 (1375)	3150 (6626)	30 (390)	11 (11)	103	8402	4387	75.59	1.92
MOPS6h	1928	96 (668)	2135 (2867)	18 (289)	10 (10)	314	3834	4177	53.84	0.92
M-C1h	2293	100 (484)	1700 (2092)	23 (218)	16 (16)	369	2810	4116	44.29	0.68
M-C2h	1608	104 (930)	2487 (4081)	25 (455)	12 (12)	265	5478	4224	61.93	1.30
M-C4h	2735	5 (65)	373 (289)	4 (70)	7 (7)	1377	431	3117	12.26	0.14
M-C6h	2722	12 (54)	377 (310)	3 (61)	5 (5)	1382	430	3114	12.59	0.14

each transcribed gene had more than one asRNAs. Interestingly, although similar numbers of genes were transcribed in all the conditions except carbon starvation (samples

M-C4h and M-C6h), HS generally resulted in the most pervasive antisense transcription, as more than 81% of the transcribed genes had antisense transcription. HS also had the highest number of genes with the equal transcription mode, but fewest number of genes with sense only transcription (Table 3.4). In contrary, carbon starvation gave rise to the least pervasive antisense transcription, as only 12~62% of the transcribed genes had antisense transcription. Carbon starvation also had the fewest genes with the equal transcription mode, but largest number of genes with sense only transcription (Table 3.4). Meanwhile, the nutrient sufficient cultures MOPS and LB induced intermediate levels of antisense transcription, as 54~75% of the transcribed genes had antisense transcription. They also had intermediate numbers of ORFs with sense only and equal transcription modes (Table 3.4). These results strongly suggest that both the extent and modes of antisense transcription in the organism are culture condition dependent, stress can dramatically change the extent and modes of antisense transcription. Furthermore, the number of transcribed ORFs having antisense transcription dropped with the time of incubation in all the culture conditions except HS, and the same was true for the average number of asRNA per transcribed ORF, while the number of ORFs with sense only transcription increase with time of incubation except HS (Table 3.4). These results strongly suggest that both the extent and modes of antisense transcription in the organism are also time or growth phase dependent. In addition, dependent on culture conditions and growth phases, the most dominant form of the six transcription modes could be either equal transcription or sense only transcription, and the least used form was always antisense only, while sense dominant and antisense dominant were in the middle. Taken together, our data indicates that antisense transcription might play an important role not

only in stress responses such as HS and carbon starvation but also in nutrient sufficient cultures such as MOPS and LB.

3.4.4. ORFs switch their transcription modes in a time and condition dependent manner

We next asked whether or not an ORF switched its transcription mode under different growth phases of a culture condition and under different culture conditions. To this end,

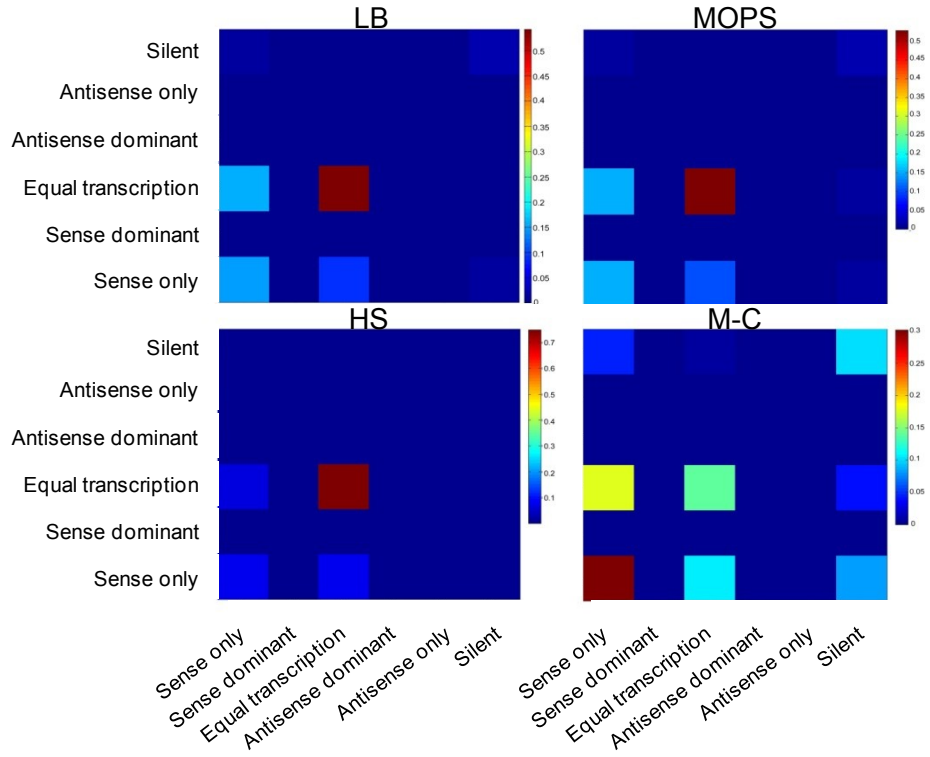


Figure 3.7. Transitions among different transcription modes between two adjacent time points in LB, MOPS, HS and M-C cultures. Row represents earlier stage and column represents immediate later stage.

we first counted each of all possible 6x6=36 transitions occurring between each two adjacent time points under a cultural condition. As shown in Figure 3.7, for all the four culture conditions, the most frequently occurring transitions between different modes were from equal transcription mode to sense only mode, and from sense only mode to

equal transcription nodes, although there were clearly subtle quantitative differences among the four conditions. For instance, there were more transitions from equal transcription mode to sense only mode under LB, MOPS and M-C cultures, while it was not true under HS. There were also considerable transitions from silent mode to sense only mode under LB, MOPS and particular M-C cultures, but not under HS culture. Furthermore, a large number of ORFs with equal transcription mode or sense only modes remained in the same modes between adjacent time points under all the four culture conditions examined (Figure 3.7). However, under carbon starvation (M-C), there were considerable numbers of transitions from equal transcription and sense only modes to silent transcription mode. In summary, the nutrient sufficient LB and MOPS cultures induced a very similar transitions pattern of transcription modes of ORFs, which were quite different from those induced by stress responses under HS and M-C cultures, suggesting that antisense transcription played important roles in the transition of growth phases under a variety of culture conditions.

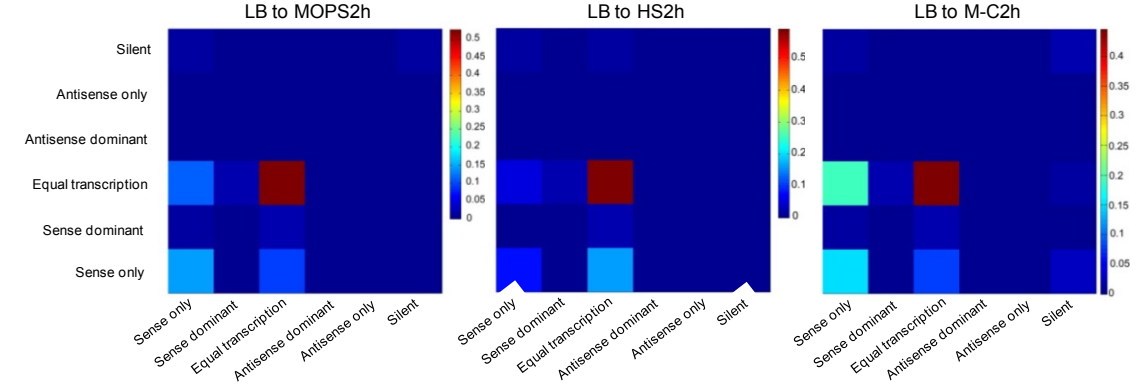


Figure 3.8. Transitions among different transcription modes when cells were transferred from LB (OD=1.0) to MOPS, HS and M-C cultures for two hours. Row represents earlier stage and column represents immediate later stage.

To further understand the patterns that ORFs changed their transcription modes under different culture conditions, we counted each of the 6x6=36 possible transitions by comparing an ORF's transcription mode in sample LB1.0 to the samples taken 2 hours after growing in HS, M-C and MOPS (i.e., samples HS2h, M-C2h and MOPS2h) when cells had fully adapted to the new growth conditions. As shown in Figure 3.8, when cells transferred from LB1.0 to any of the new growth culture conditions, the dominant transitions between different modes were from equal transcription mode to sense only mode or sense dominant mode, from sense only mode or sense dominant mode to equal transcription mode, while larger portions of ORFs are staying in the same modes of either equal transcription or sense only. However, there were striking differences among the transition patterns under different new cultures. For instance, for both the LB to MOPS and to M-C transfers, there were more transitions from equal transcription mode to sense only mode than for the reverse order transitions, but for the LB to HS transfer, the opposite was true. There were also unique transition patterns to specific culture transfers. For instance, the transitions from silent mode to equal transcription mode for the LB to HS transfer, and the transitions from sense only mode to silent mode for the LB to M-C transfer (Figure 3.8). These results again strongly suggest that antisense transcription plays a critical role for the adaptation of the bacterium to the environment.

Figure 3.9 shows an example of how the gene, *sulA* encoding the cell division inhibitor, switched its transcription mode from equal transcription in LB to sense dominant at HS15min and HS30min, back to equal transcription at HS1h and HS2h, and

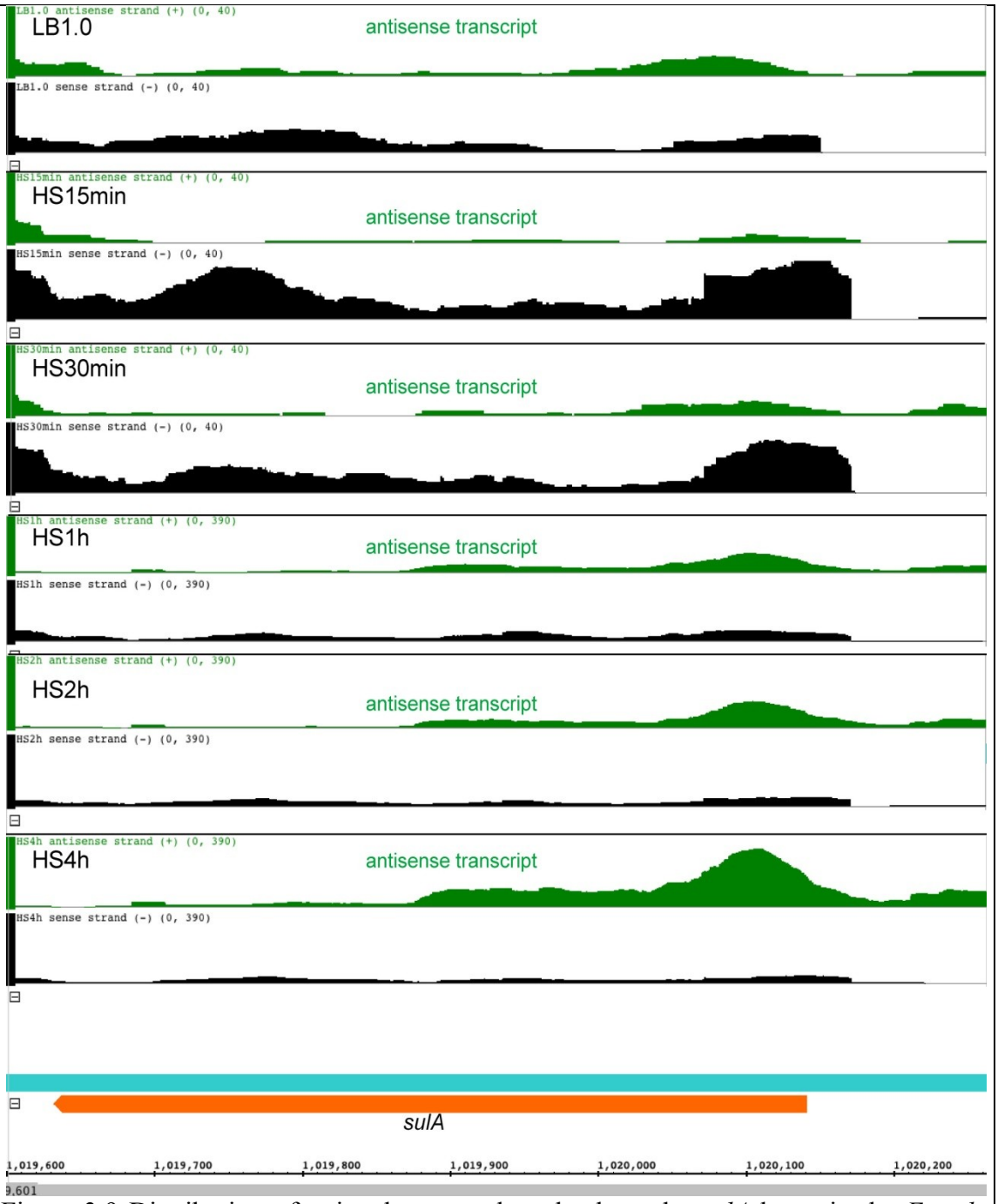


Figure 3.9 Distribution of uniquely mapped reads along the *sulA* locus in the *E. coli* genome before (LB1.0) and at different time points of heat shock. The vertical axis is the number of reads covered at the position. The dark cyan and orange bars at the bottom of the graph represent the forward and reverse strands, respectively. The arrowed orange segment represents the *sulA* gene. To make the expression levels for *sulA* and its antisense transcripts in different samples visible and comparable, the same vertical axis scale (40) is used for LB1.0, HS15min and HS30min, and the same vertical axis scale (390) is used for the other samples.

then to antisense dominant at HS4h. Such dynamic changes in transcription mode might be explained by the function of the *sulA* gene: at log growth phase in LB (OD=1.0), the

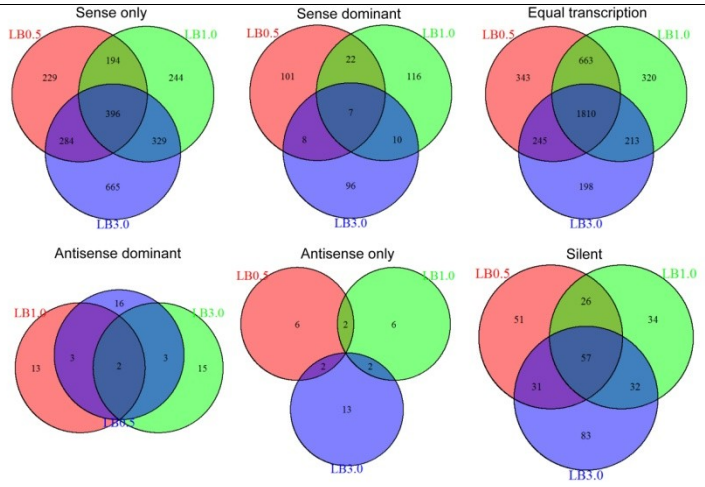


Figure 3.10. Venn diagram showing common ORFs with the same transcription mode among the three samples from LB culture taken at OD=0.5, 1.0 and 3.0.

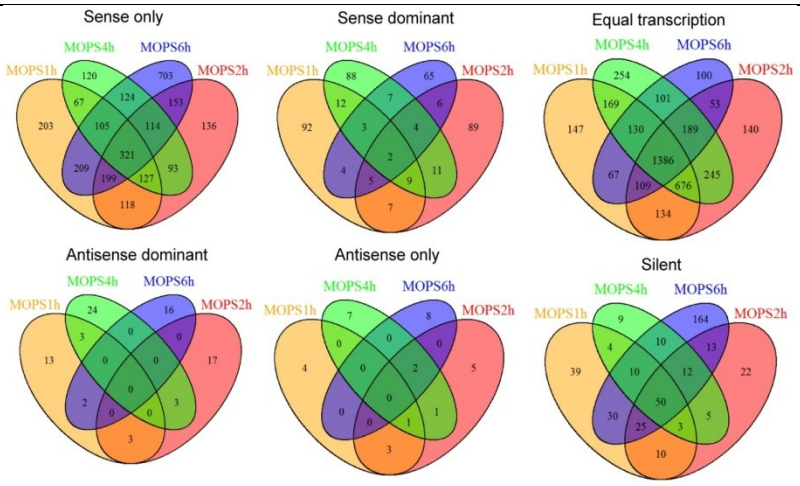


Figure 3.11. Venn diagram showing common ORFs with the same transcription mode among the four samples from MOPS culture taken at 1, 2, 4 and 6 hrs after transfer from LB1.0 to MOPS.

sulA locus was transcribed with equal transcription mode, so the activity of *sulA* was fined tuned by equivalent amount of its asRNAs. When the cells first responded to heat shock, the activity of *sulA* was enhanced by increased sense transcription and decrease

antisense transcription (Figure 3.9), so the locus was transcribed with sense dominant mode to prevent the cells from dividing before heat shock induced DNA damage had

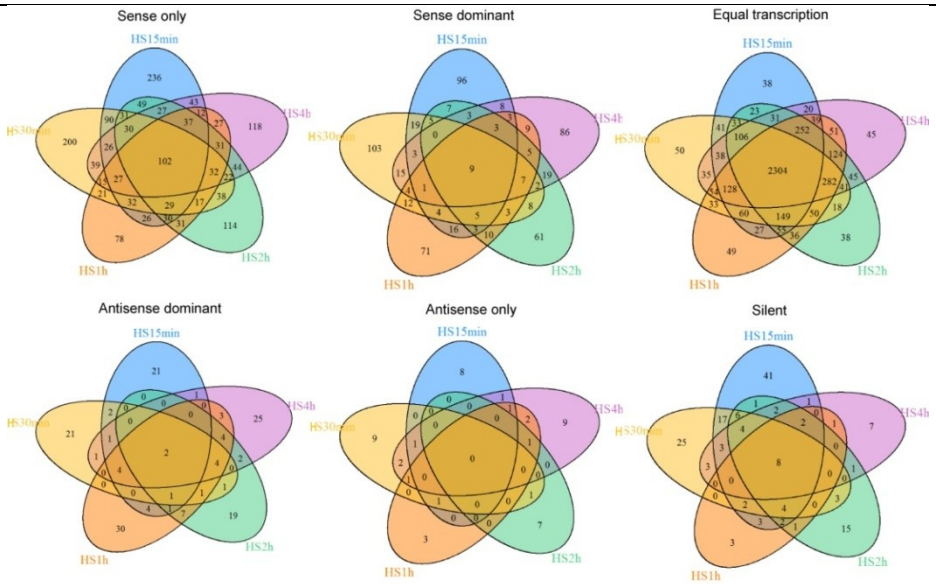


Figure 3.12. Venn diagram showing common ORFs with the same transcription mode among the five samples from HS culture taken at 15min, 30min, 1 hr and 2 hrs after transfer from LB1.0 (37ºC) to MOPS (48ºC).

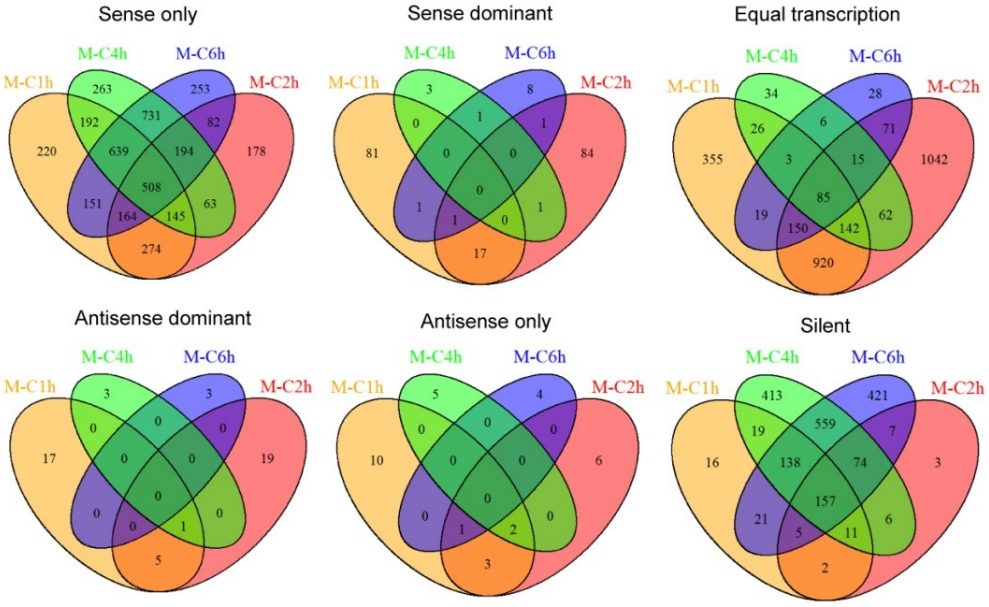


Figure 3.13. Venn diagram showing common ORFs with the same transcription mode among the four samples from MOPS culture taken at 1, 2, 4 and 6 hrs after transfer from LB1.0 to MOPS without carbon (M-C).

been repaired. However, when cells gradually adapted to heat shock, SulA was less needed, thus the sense transcription decreased while the antisense increased, so the activity of *sulA* was fined tuned in the equal transcription mode. After the cells fully adapted to HS, the activity of *sulA* was attenuated by overwhelmingly larger amount of antisense transcription in the antisense dominant mode, thereby the cell division resumed.

To understand which ORFs changed and which ORFs remained their transcription modes between the time points examined, we analyzed using Venn diagrams ORFs with the same transcription mode at different time points under the four culture conditions. As shown in Figures 3.10, 3.11, 3.12, and 3.13 for LB, MOPS, HS and M-C, respectively, and we have already indicated earlier, there were many ORFs remained the same transcription mode of equal transcription and sense only. GO term analysis revealed that these ORFs are mainly involved in housekeeping functions. On the other hand, there were very few or none ORF sharing the same transcription node of sense dominant, antisense dominant and antisense only, suggesting that house-keeping genes were not transcribed in these modes. Furthermore, there were always unique ORFs with a certain mode of transcription under the four culture conditions, indicating that there are always some ORFs changing their transcription to that mode from different ones. Some ORFs were shared only by adjacent time points for a certain transcription mode, suggesting these genes change their transcription mode slowly, but eventually switched to a different one. Some ORFs were shared only by none-adjacent time points, indicating that these ORFs return to their original transcription mode after changing to a different one.

To understand which ORFs changed and which ORFs remained their transcription modes under different culture conditions, we analyzed using Venn diagrams ORFs with

the same transcription mode when growing in LB and after being transferred to different culture conditions for 2 hours when the cells were fully adapted to the new culture conditions. As shown in Figure 3.14, there were larger number of ORFs with sense only and equal transcription common to any two, three and all of the four samples (LB1.0, MOPS2h, HS2h, and M-C2h), suggesting that these ORFs were transcribed in the same mode at least two different conditions. GO term analysis revealed that the common genes

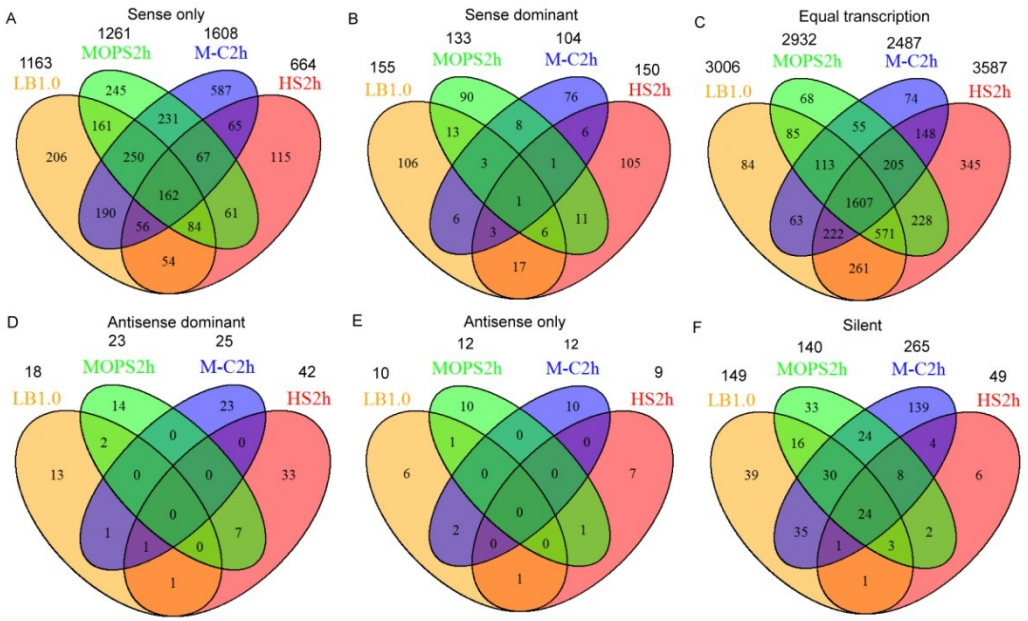


Figure 3.14. Venn diagram showing common ORFs with the same transription mode when cells were growing in LB and after being transferred to MOPS, HS and M-C for two hours.

were mainly involved house-keeping functions. For instance, the 1607 common ORFs equal transcription in all the four samples were enriched for housekeeping functions such as amino sugar and nucleotide sugar metabolism (KEGG ID: eco00520), Citrate cycle (TCA cycle) (KEGG ID: ecg00020), phage recognition/detection of virus (GO:0009597), RNA degradation (KEGG ID: ecj03018), cell wall macromolecule biosynthetic process

(GO:0044038), alcohol catabolic process (GO:0046164), Pentose phosphate pathway (KEGG ID: ecr00030), pentose and glucuronate interconversions (KEGG ID: ecf00040), bacterial secretion system (KEGG ID: ect03070), etc (Table S1 in supplementary file 1). Furthermore, there were ORFs unique to each sample for each transcription mode, suggesting that they might be involved in functions of cells under specific conditions. For instance, the 345 ORFs unique to sample HS2h with equal transcription were enriched for functions of DNA damage-induced stress responses such as mismatch repair (KEGG ID: eck03430), homologous recombination (KEGG ID: ecc03440), DNA replication (KEGG ID: ecc03030), and base excision repair (KEGG ID: ecf03410) (Table S2). It is well-known that under some extreme conditions such as heat shock, these SOS response genes

Table 3.5 Number of ORFs with a certain transcription mode in different samples.

Transcription mode	Union				Intersection				Union of 16 samples	Intersection of 16 samples
	LB	MOPS	HS	M-C	LB	MOPS	HS	M-C		
Sense only	2341	2792	1654	4057	396	321	102	508	4329	13
Sense dominant	360	404	600	198	7	2	9	0	1097	0
Equal transcription	3792	3900	4295	2958	1810	1386	2304	85	4380	59
Antisense dominant	52	81	155	48	2	0	2	0	272	0
Antisense only	31	31	47	31	0	0	0	0	110	0
Silent	314	406	155	1852	57	50	8	157	1867	8

are induced for the survival of cell through the stress. Thus, our data presented here might indicate that this mode of transcription might play a role in fine tuning the expression of these SOS genes. Furthermore, the 74 ORFs unique to sample M-C2h (Figure 3.8A) with equal transcription mode were enriched for functions of cell division, cell cycle and cell inner membrane, etc (Table S3). However, as ORFs with antisense only mode was all unique to a sample, this mode seemed to be restricted on a set of highly unique genes.

Finally, as summarized in Table 3.5, although there were a large number of ORFs with sense only, sense dominant, or equal transcription modes and a considerable number of ORFs with antisense dominant or antisense only modes in at least one the 16 samples, few or none ORFs with these transcription mode were shared by all the 16 samples,

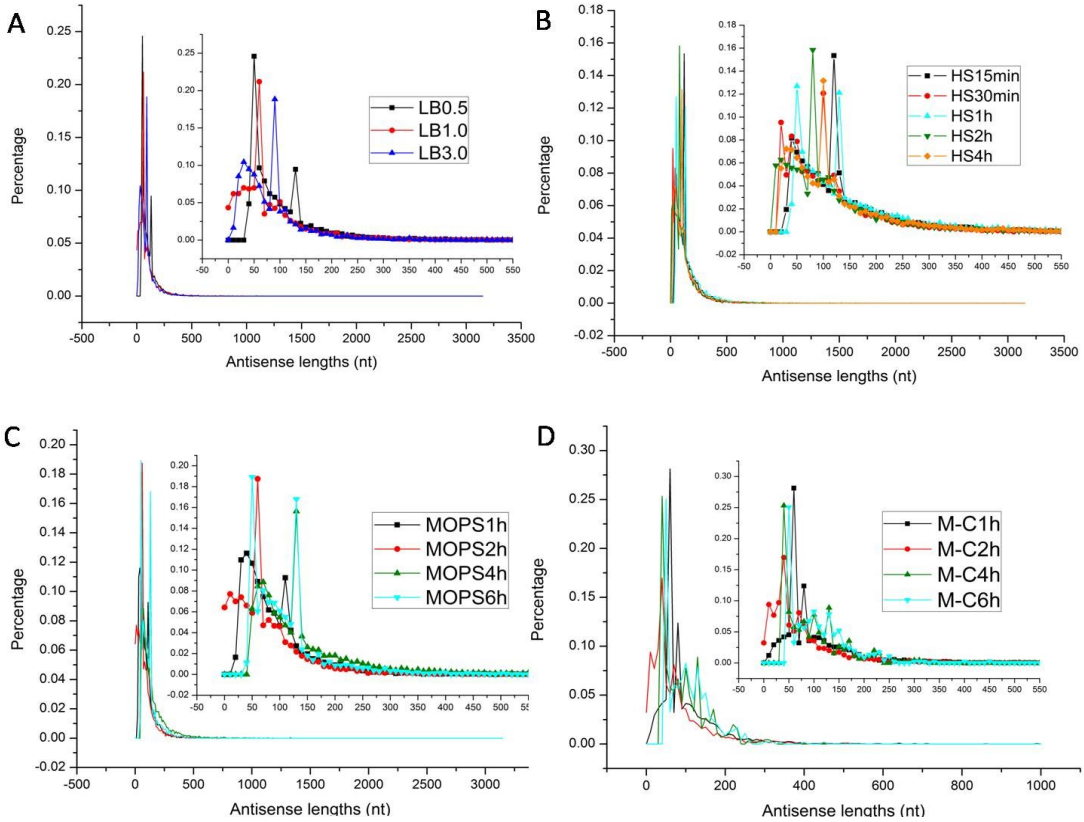


Figure 3.15 Length of assembled antisense transcripts in all the samples. A) LB treatment. B) HS treatment. C) MOPS treatment. D) M-C treatment.

except for equal transcription and sense only modes, where 59 and 13 house-keeping genes remained in the same mode in all the samples, respectively. Thus the results clearly

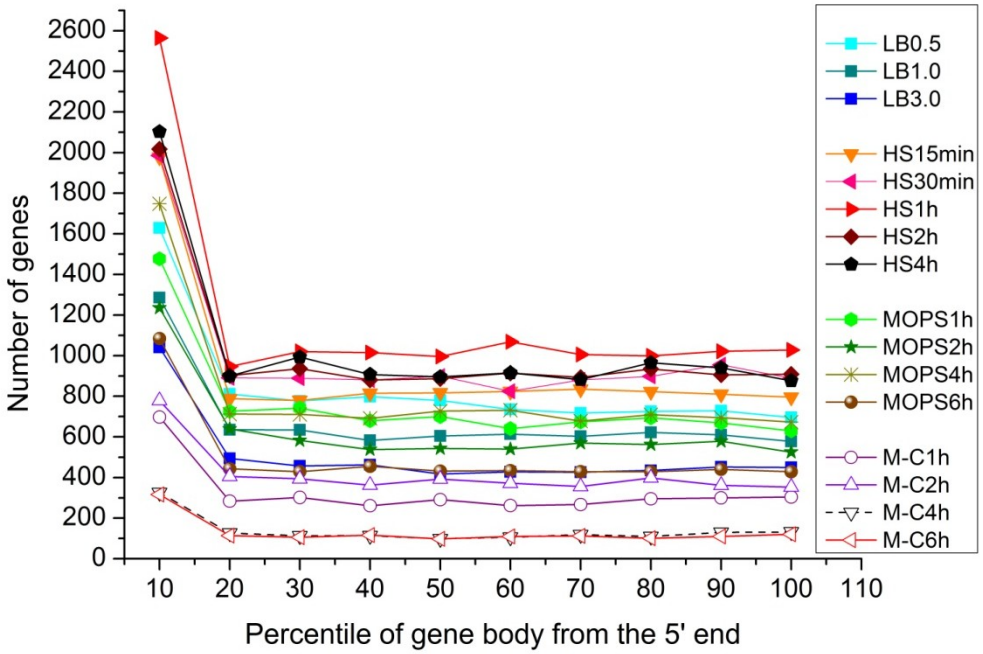


Figure 3.16. Relative locations of antisense transcripts on the gene body.

indicate that vast majority of ORFs changed their transcription modes at different growth phases and under different environmental conditions. In particular, ORFs with sense dominant and antisense dominant seemed to always switch to other transcription modes.

3.4.5 Antisense transcripts are initiated at and restricted to 5' ends

As shown in Figure 3.15, the lengths of assembled antisense transcripts varied from tens to several thousand nucleotides, but the vast majority antisense transcripts were smaller than 150nt. However, asRNAs in samples from different treatments may have quite different length distributions. For instance, samples from HS treatment seemed to have longer antisense transcripts than other treatments, while samples from M-C treatment had much shorter antisense transcripts. To see whether there was a pattern of the location of the short asRNA relative to the ORF on the sense strand, we divided each ORF to 10 equal portions (the 5'UTR also belong to the 1st bin), the relative position of

the 5'end of an asRNA on the gene body determines which percentile this asRNA falls into. Figure 3.16 shows the distribution of relative locations of antisense transcripts on the gene body. Surprisingly, the majority of antisense transcripts are located to the most 5' end of the genes or on their 5' UTRs, especially for the samples in HS treatment, where a lot more genes are highly expressed than in other conditions. Since the antisense transcripts near the 5' end of gene or in 5' UTR can probably interfere with transcription on the sense strand or block translation on the sense strand either in a direct or indirect way [83], these large amount of antisense transcripts appearing near the 5' end are highly likely to repress or fine tune the expression on the sense strand.

3.5 Conclusion

By applying RNA-seq technique to *E. coli* K12, we reconstructed its transcriptomes under different growth phases and culture conditions, and found that antisense transcription was a common and widespread phenomenon and was much more pervasive than originally anticipated. We found that up to one third of the genome had transcripts from both the forward and reverse strands, and between 13 and 87% of transcribed ORFs had at least one asRNA, dependent on growth phases and culture conditions. ORFs could have six different modes of transcription in a growth phase and culture condition dependent manner: sense only, sense dominant, equal transcription, antisense dominant, antisense only and silent modes. The modes sense dominant, equal transcription, antisense dominant might present different levels of regulation by antisense transcripts. Almost all transcribed genes in our dataset changed their transcription modes between different growth phases and culture conditions, except for dozens of housekeeping genes that tends to remain in sense only or equal transcription modes. Moreover, we found that

antisense transcriptions can be initiated anywhere along an ORF, but strongly biased and restricted to the 5'end the ORF, suggesting that a large portion of asRNA might achieve the regulation roles through transcription interference or translation blocking. Therefore, antisense transcription is very prevalent in *E. coli* K12, and may play important roles in various aspects of the bacterium's physiology through modulation transcription or translation processes.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation is an extensive investigation on evolution and dynamics of transcriptional regulation in bacteria using a combination of computational and experiment approaches. By studying the evolution of LexA regulons in cyanobacteria, we furthered our understanding of how the *cis*-regulatory elements such as the LexA binding sites evolve in the closely related species, thereby rewiring the transcriptional regulation networks, and how the divergence of *cis*-regulatory elements plays an important role in organisms' adaptation to environments during the course of evolution. By developing a transcriptome assembler tailored to prokaryotes using RNA-seq short reads, we reconstructed the alternative operon structures in the model microbial organism *Escherichia coli* K12 under different growth phases and culture conditions, and elucidated the modes of pervasive antisense transcription in a condition and stage dependent manner. Elucidation of these rules in bacteria is essential to better understanding transcriptional regulatory mechanisms and to the physiology of prokaryotic cells.

In chapter 1 we utilized a comparative genomic approach to study the LexA regulon in cyanobacteria. Specifically, we applied a regulon prediction algorithm [164] that we developed earlier to elucidate the evolution of the transcription factor LexA and its regulons in cyanobacteria. We found that in most cyanobacterial genomes that we analyzed, LexA appears to function as the transcriptional regulator of the key SOS

response genes. There are possible couplings between the SOS response and other biological processes. In some cyanobacteria like *Synechocystis* PCC 6803, LexA has adapted distinct functions, and might no longer be a regulator of the SOS response system. In some other cyanobacteria, *lexA* appears to have been lost during the course of evolution. The loss of *lexA* in these genomes might lead to the degradation of its binding sites. Moreover, we conclude that cyanobacteria inherited the *lexA* gene from their last common ancestor; however, substantial genome-wide turnover seems to have led to the high degree of variation of the LexA regulons in some species during evolution. Moreover, the divergence within *cis*-regulatory elements or the binding sites turn over facilitates the transcriptional rewiring and phenotypic adaptation. Of course, numerous important questions related to this topic remain to be elucidated. For instance, it is very interesting to study the co-evolution of the DNA binding domain of a TF and its binding sites in a wide spectrum of evolution distances.

Chapter 2 focused on development of the HMM-based transcriptome assembler tailored to prokaryotes using RNA-seq short reads. Although numerous transcriptome assembly algorithms and tools have been developed in the past several years using RNA-seq short reads generated by next-generation sequencing (NGS) technologies, these tools are mainly designed for assembling eukaryotic isoforms, and cannot be used for prokaryotic transcriptome assembly. Furthermore, as has been shown earlier [10-14, 101, 102] and we indicated in this chapter that, the coverage of reads on transcribed regions in these studies are highly non-uniform, and there are even numerous zero coverage positions in transcribed regions [103-105], leading to gaps in otherwise an overlapping mapping of reads to a transcribed region [106-108]. Therefore, we developed a HMM

based gap-tolerant algorithm and tool, TruHmm, for simultaneous assembly of full-length prokaryotic transcripts using a sliding-window strategy. When evaluated on a directional RNA-seq dataset collected in *Escherichia coli* K12 str. MG1655 (*E. coli* K12) under different culture conditions and time points, TruHmm is able to reconstruct known operons with very high sensitivity and specificity. Nevertheless, the limitation of our current prototype model is that it can only infer the expression or non-expression state for each position on the genome without estimation of the transcription level for one transcript to determine the portions of dynamic operons with 'stair-case' manner, as discovered by Guell*, et. al* [10] in a genome-reduced bacterium. Hence, our model still needs to be upgraded to detect the dynamic (varying levels) expression along the assembled operon. With this tool, we will reconstruct alternative operons as well as anti-sense and non-coding expression patterns under various growth conditions and time points in *E. coli* K12 by using a multiplex directional RNA-seq method for capturing RNA fragment of various lengths and types.

In chapter 3, we analyzed the patterns of antisense transcription in E. coli K12 under different growth phase and culture conditions using directional RNA-seq and the assembler TruHmm we developed in Chapter 2. We found that up to one third of the genome had transcripts from both the forward and reverse strands, and between 13 and 87% of transcribed ORFs had at least one asRNA, dependent on growth phases and culture conditions. ORFs could have six different modes of transcription in a growth phase and culture condition dependent manner: sense only, sense dominant, equal transcription, antisense dominant, antisense only and silent modes. The modes sense dominant, equal transcription, antisense dominant might present different levels of

regulation by antisense transcripts. Almost all transcribed genes in our dataset changed their transcription modes between different growth phases and culture conditions, except for dozens of housekeeping genes that tends to remain in sense only or equal transcription modes. Moreover, we found that antisense transcriptions can be initiated anywhere along an ORF, but are strongly biased and restricted to the 5'end the ORF, suggesting that asRNA might achieve the regulation roles through transcription interference or translation blocking. Therefore, antisense transcription is very prevalent in *E. coli* K12, and may play important roles in various aspects of the bacterium's physiology through modulation transcription or translation processes. Still, many open questions remain in the field. For example, how is antisense transcription initiated and regulated? What triggers the switch of transcription modes under certain conditions? Furthermore, for the multi-gene operon, which gene is targeted for antisense transcripts?

To summarize, first, our investigation on the LexA regulon has largely furthered our understanding of the evolution of transcriptional networks in prokaryotes. Second, our tool for prokaryotic transcriptome assembling has proven to be very useful for our research to reveal the complexity of prokaryotic transcriptome. As RNA-seq becomes a routine for probing transcriptomes in prokaryotes, we hope our software can be a useful tool for understanding the complexity of transcriptomes and the underlying mechanisms in prokaryotic cells. Third, our analyses on alternative operon utilized and antisense transcription in *E. coli* K12 have greatly enhanced our understanding of the prevalence, patterns and molecular mechanisms of these two newly discovered important transcriptional regulation in prokaryotes.

REFERENCES

1. Liu JM, Camilli A: A broadening world of bacterial small RNAs. *Curr Opin Microbiol* 2010, 13(1):18-23.

2. Toledo-Arana A, Solano C: Deciphering the physiological blueprint of a bacterial cell: revelations of unanticipated complexity in transcriptome and proteome. *Bioessays* 2010, 32(6):461-467.

3. Sorek R, Cossart P: Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* 2010, 11(1):9-16.

4. Guell M, Yus E, Lluch-Senar M, Serrano L: Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat Rev Microbiol* 2011, 9(9):658-669.

5. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH: Structure and complexity of a bacterial transcriptome. *J Bacteriol* 2009, 191(10):3203-3211.

6. van Vliet AH, Wren BW: New levels of sophistication in the transcriptional landscape of bacteria. *Genome Biol* 2009, 10(8):233.

7. Pinto AC, Melo-Barbosa HP, Miyoshi A, Silva A, Azevedo V: Application of RNA-seq to reveal the transcript profile in bacteria. *Genet Mol Res* 2011, 10(3):1707-1718.

8. Filiatrault MJ: Progress in prokaryotic transcriptomics. *Curr Opin Microbiol* 2011, 14(5):579-586.

9. van Vliet AH: Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol Lett* 2010, 302(1):1-7.

10. Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S *et al*: Transcriptome complexity in a genome-reduced bacterium. *Science* 2009, 326(5957):1268-1271.

11. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R *et al*: The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* 2010, 464(7286):250-255.

12. Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S *et al*: Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science* 2012, 335(6072):1103-1106.

13. Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY *et al*: Prevalence of transcription promoters within archaeal

operons and coding sequences. *Mol Syst Biol* 2009, 5:285.

14. Hovik H, Yu WH, Olsen I, Chen T: Comprehensive transcriptome analysis of the periodontopathogenic bacterium Porphyromonas gingivalis W83. *J Bacteriol* 2012, 194(1):100-114.

15. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat Biotechnol* 2000, 18(12):1262-1268.

16. Rasmussen S, Nielsen HB, Jarmer H: The transcriptionally active regions in the genome of Bacillus subtilis. *Mol Microbiol* 2009, 73(6):1043-1057.

17. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ *et al*: A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet* 2009, 5(7):e1000569.

18. Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R: Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A* 2009, 106(10):3976-3981.

19. McGrath PT, Lee H, Zhang L, Iniesta AA, Hottes AK, Tan MH, Hillson NJ, Hu P, Shapiro L, McAdams HH: High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol* 2007, 25(5):584-592.

20. Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penades JR, Valle J *et al*: Genome-wide antisense transcription drives mRNA processing in bacteria. *Proc Natl Acad Sci U S A* 2011, 108(50):20172-20177.

21. Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, Waldor MK: RNA-Seq-based monitoring of infection-linked changes in Vibrio cholerae gene expression. *Cell Host Microbe* 2011, 10(2):165-174.

22. Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T: Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res* 2010, 38(3):868-877.

23. Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, Vogel J, Rudel T: The transcriptional landscape of Chlamydia pneumoniae. *Genome Biol* 2011, 12(10):R98.

24. Wang Y, Li X, Mao Y, Blaschek HP: Single-nucleotide resolution analysis of the transcriptome structure of Clostridium beijerinckii NCIMB 8052 using RNA-Seq. *BMC Genomics* 2011, 12:479.

25. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K *et al*: The Listeria transcriptional landscape from saprophytism to virulence. *Nature* 2009, 459(7249):950-956.

26. Flaherty BL, Van Nieuwerburgh F, Head SR, Golden JW: Directional RNA deep sequencing sheds new light on the transcriptional response of Anabaena sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genomics* 2011, 12:332.

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

28. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R: A single-base resolution map of an archaeal transcriptome. *Genome Res* 2010, 20(1):133-141.

29. Pop M: Genome assembly reborn: recent computational challenges. *Brief Bioinform* 2009, 10(4):354-366.

30. Flicek P, Birney E: Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009, 6(11 Suppl):S6-S12.

31. Martin JA, Wang Z: Next-generation transcriptome assembly. *Nat Rev Genet* 2011, 12(10):671-682.

32. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, 28(5):511-515.

33. Little JW, Mount DW: The SOS regulatory system of Escherichia coli. *Cell* 1982, 29(1):11-22.

34. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R: Identification of additional genes belonging to the LexA regulon in Escherichia coli. *Mol Microbiol* 2000, 35(6):1560-1572.

35. Groban ES, Johnson MB, Banky P, Burnett PG, Calderon GL, Dwyer EC, Fuller SN, Gebre B, King LM, Sheren IN *et al*: Binding of the Bacillus subtilis LexA protein to the SOS operator. *Nucleic Acids Res* 2005, 33(19):6287-6295.

36. Wojciechowski MF, Peterson KR, Love PE: Regulation of the SOS response in Bacillus subtilis: evidence for a LexA repressor homolog. *J Bacteriol* 1991, 173(20):6489-6498.

37. Mazon G, Erill I, Campoy S, Cortes P, Forano E, Barbe J: Reconstruction of the evolutionary history of the LexA-binding sequence. *Microbiology* 2004, 150(Pt 11):3783-3795.

38. Erill I, Campoy S, Barbe J: Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS Microbiol Rev* 2007, 31(6):637-656.

39. Horii T, Ogawa T, Nakatani T, Hase T, Matsubara H, Ogawa H: Regulation of SOS functions: purification of E. coli LexA protein and determination of its specific site cleaved by the RecA protein. *Cell* 1981, 27(3 Pt 2):515-522.

40. Luo Y, Pfuetzner RA, Mosimann S, Paetzel M, Frey EA, Cherney M, Kim B, Little JW, Strynadka NC: Crystal structure of LexA: a conformational switch for regulation of self-cleavage. *Cell* 2001, 106(5):585-594.

41. Slilaty SN, Little JW: Lysine-156 and serine-119 are required for LexA repressor cleavage: a possible mechanism. *Proc Natl Acad Sci U S A* 1987, 84(12):3987-3991.

42. Michel B: After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biol* 2005, 3(7):e255.

43. Walker GC: Mutagenesis and inducible responses to deoxyribonucleic acid damage in Escherichia coli. *Microbiol Rev* 1984, 48(1):60-93.

44. Fogh RH, Ottleben G, Ruterjans H, Schnarr M, Boelens R, Kaptein R: Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy. *Embo J* 1994, 13(17):3936-3944.

45. Knegtel RM, Fogh RH, Ottleben G, Ruterjans H, Dumoulin P, Schnarr M, Boelens R, Kaptein R: A model for the LexA repressor DNA complex. *Proteins* 1995, 21(3):226-236.

46. Wertman KF, Mount DW: Nucleotide sequence binding specificity of the LexA repressor of Escherichia coli K-12. *J Bacteriol* 1985, 163(1):376-384.

47. Sjoholm J, Oliveira P, Lindblad P: Transcription and regulation of the bidirectional hydrogenase in the cyanobacterium Nostoc sp. strain PCC 7120. *Appl Environ Microbiol* 2007, 73(17):5435-5446.

48. Mazon G, Lucena JM, Campoy S, Fernandez de Henestrosa AR, Candau P, Barbe J: LexA-binding sequences in Gram-positive and cyanobacteria are closely related. *Mol Genet Genomics* 2004, 271(1):40-49.

49. Gutekunst K, Phunpruch S, Schwarz C, Schuchardt S, Schulz-Friedrich R, Appel J: LexA regulates the bidirectional hydrogenase in the cyanobacterium Synechocystis sp. PCC 6803 as a transcription activator. *Mol Microbiol* 2005, 58(3):810-823.

50. Oliveira P, Lindblad P: LexA, a transcription regulator binding in the promoter region of the bidirectional hydrogenase in the cyanobacterium Synechocystis sp.

PCC 6803. *FEMS Microbiol Lett* 2005, 251(1):59-66.

51. Patterson-Fortin LM, Colvin KR, Owttrim GW: A LexA-related protein regulates redox-sensitive expression of the cyanobacterial RNA helicase, crhR. *Nucleic Acids Res* 2006, 34(12):3446-3454.

52. Patterson-Fortin LM, Owttrim GW: A Synechocystis LexA-orthologue binds direct repeats in target genes. *FEBS Lett* 2008, 582(16):2424-2430.

53. Domain F, Houot L, Chauvat F, Cassier-Chauvat C: Function and regulation of the cyanobacterial genes lexA, recA and ruvB: LexA is critical to the survival of cells facing inorganic carbon starvation. *Mol Microbiol* 2004, 53(1):65-80.

54. Sugita C, Ogata K, Shikata M, Jikuya H, Takano J, Furumichi M, Kanehisa M, Omata T, Sugiura M, Sugita M: Complete nucleotide sequence of the freshwater unicellular cyanobacterium Synechococcus elongatus PCC 6301 chromosome: gene content and organization. *Photosynth Res* 2007, 93(1-3):55-67.

55. Gupta RS, Mathews DW: Signature proteins for the major clades of Cyanobacteria. *BMC Evol Biol*, 10:24.

56. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009, 37(Web Server issue):W202-208.

57. Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127-138.

58. Sierro N, Makita Y, de Hoon M, Nakai K: DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 2008, 36(Database issue):D93-96.

59. Erill I, Jara M, Salvador N, Escribano M, Campoy S, Barbe J: Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. *Nucleic Acids Res* 2004, 32(22):6617-6626.

60. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A *et al*: RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* 2011, 39(Database issue):D98-105.

61. Mahony S, Benos PV: STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 2007, 35(Web Server issue):W253-258.

62. Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: a sequence logo

generator. *Genome Res* 2004, 14(6):1188-1190.

63. Tamura K, Dudley J, Nei M, Kumar S: MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007, 24(8):1596-1599.

64. Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16(1):16-23.

65. Su Z, Olman V, Mao F, Xu Y: Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res* 2005, 33(16):5156-5171.

66. Xu M, Su Z: Computational prediction of cAMP receptor protein (CRP) binding sites in cyanobacterial genomes. *BMC Genomics* 2009, 10:23.

67. Su Z, Olman V, Xu Y: Computational prediction of Pho regulons in cyanobacteria. *BMC Genomics* 2007, 8:156.

68. Kelley WL: Lex marks the spot: the virulent side of SOS and a closer look at the LexA regulon. *Mol Microbiol* 2006, 62(5):1228-1238.

69. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H *et al*: RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 2008, 36(Database issue):D120-124.

70. Cirz RT, Chin JK, Andes DR, de Crecy-Lagard V, Craig WA, Romesberg FE: Inhibition of mutation and combating the evolution of antibiotic resistance. *PLoS Biol* 2005, 3(6):e176.

71. Butala M, Zgur-Bertok D, Busby SJ: The bacterial LexA transcriptional repressor. *Cell Mol Life Sci* 2009, 66(1):82-93.

72. Bisognano C, Kelley WL, Estoppey T, Francois P, Schrenzel J, Li D, Lew DP, Hooper DC, Cheung AL, Vaudaux P: A recA-LexA-dependent pathway mediates ciprofloxacin-induced fibronectin binding in Staphylococcus aureus. *J Biol Chem* 2004, 279(10):9064-9071.

73. Cirz RT, Jones MB, Gingles NA, Minogue TD, Jarrahi B, Peterson SN, Romesberg FE: Complete and SOS-mediated response of Staphylococcus aureus to the antibiotic ciprofloxacin. *J Bacteriol* 2007, 189(2):531-539.

74. Sycheva LV, Permina EA, Gel'fand MS: Taxon-specific regulation of SOS-response in gamma-proteobacteria. *Mol Biol (Mosk)* 2007, 41(5):908-917.

75. Dam P, Olman V, Harris K, Su Z, Xu Y: Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 2007, 35(1):288-298.

76. Hulsen T, Huynen MA, de Vlieg J, Groenen PM: Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006, 7(4):R31.

77. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994, 2:28-36.

78. Metropolis N, Ulam S: The Monte Carlo method. *J Am Stat Assoc* 1949, 44(247):335-341.

79. Hu J, Li B, Kihara D: Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* 2005, 33(15):4899-4913.

80. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al*: Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, 23(1):137-144.

81. Stormo GD: Consensus patterns in DNA. *Methods Enzymol* 1990, 183:211-221.

82. Felsenstein J: Phylogeny Inference Package (Version 3.2). *Cladistics* 1989, 5:164-166.

83. Thomason MK, Storz G: Bacterial antisense RNAs: how many are there, and what are they doing? *Annu Rev Genet* 2010, 44:167-188.

84. Georg J, Hess WR: cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev* 2011, 75(2):286-300.

85. Repoila F, Darfeuille F: Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol Cell* 2009, 101(2):117-131.

86. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T *et al*: EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res* 2011, 39(Database issue):D583-590.

87. Sierro N, Makita Y, de Hoon M, Nakai K: DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 2007, 36:D93-96.

88. Chen X, Su Z, Xu Y, Jiang T: Computational Prediction of Operons in *Synechococcus sp.* WH8102. *Genome Inform Ser Workshop Genome Inform* 2004, 15(2):211-222.

89. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI: Operon prediction without a training set. *Bioinformatics* 2005, 21(7):880-888.

90. Price MN, Huang KH, Alm EJ, Arkin AP: A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 2005, 33(3):880-892.

91. Tran TT, Dam P, Su Z, Poole FL, 2nd, Adams MW, Zhou GT, Xu Y: Operon prediction in Pyrococcus furiosus. *Nucleic Acids Res* 2007, 35(1):11-20.

92. Bergman NH, Passalacqua KD, Hanna PC, Qin ZS: Operon prediction for sequenced bacterial genomes without experimental information. *Appl Environ Microbiol* 2007, 73(3):846-854.

93. Mao F, Dam P, Chou J, Olman V, Xu Y: DOOR: a database for prokaryotic operons. *Nucleic Acids Res* 2009, 37(Database issue):D459-463.

94. Taboada B, Verde C, Merino E: High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res* 2010, 38(12):e130.

95. Livny J: Efficient annotation of bacterial genomes for small, noncoding RNAs using the integrative computational tool sRNAPredict2. *Methods Mol Biol* 2007, 395:475-488.

96. Tjaden B: Prediction of small, noncoding RNAs in bacteria using heterogeneous data. *J Math Biol* 2008, 56(1-2):183-200.

97. Pichon C, Felden B: Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics* 2008, 24(24):2807-2813.

98. Luban S, Kihara D: Comparative genomics of small RNAs in bacterial genomes. *OMICS* 2007, 11(1):58-73.

99. Brouwer RW, Kuipers OP, Hijum SA: The relative value of operon predictions. *Brief Bioinform* 2008.

100. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, 10(1):57-63.

101. Vivancos AP, Guell M, Dohm JC, Serrano L, Himmelbauer H: Strand-specific deep sequencing of the transcriptome. *Genome Res* 2010, 20(7):989-999.

102. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 2010, 7(9):709-715.

103. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008,

5(7):621-628.

104. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 2011, 12(3):R22.

105. Cheung MS, Down TA, Latorre I, Ahringer J: Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res* 2011, 39(15):e103.

106. Sendler E, Johnson GD, Krawetz SA: Local and global factors affecting RNA sequencing analysis. *Anal Biochem* 2011, 419(2):317-322.

107. Wu Z, Wang X, Zhang X: Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* 2011, 27(4):502-508.

108. Li J, Jiang H, Wong WH: Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* 2010, 11(5):R50.

109. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO: The transcription unit architecture of the Escherichia coli genome. *Nat Biotechnol* 2009, 27(11):1043-1049.

110. Ciesiolka J, Michalowski D, Wrzesinski J, Krajewski J, Krzyzosiak WJ: Patterns of cleavages induced by lead ions in defined RNA secondary structure motifs. *J Mol Biol* 1998, 275(2):211-220.

111. Hansen KD, Brenner SE, Dudoit S: Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 2010, 38(12):e131.

112. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J *et al*: RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* 2011, 17(9):1697-1712.

113. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB: Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res* 2012, 40(7):e54.

114. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R: Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* 2011, 39(21):e141.

115. Risso D, Schwartz K, Sherlock G, Dudoit S: GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 2011, 12:480.

116. Benjamini Y, Speed TP: Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012, 40(10):e72.

117. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011, 12(2):R18.

118. Minoche AE, Dohm JC, Himmelbauer H: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 2011, 12(11):R112.

119. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H *et al*: Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 2011, 39(13):e90.

120. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ: FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 2010, 7(2):130-132.

121. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M: Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 2009, 27(7):652-658.

122. Raz T, Causey M, Jones DR, Kieu A, Letovsky S, Lipson D, Thayer E, Thompson JF, Milos PM: RNA sequencing and quantitation using the Helicos Genetic Analysis System. *Methods Mol Biol* 2011, 733:37-49.

123. Kent WJ: BLAT--the BLAST-like alignment tool. *Genome Res* 2002, 12(4):656-664.

124. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9):1105-1111.

125. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10(3):R25.

126. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010, 28(5):503-510.

127. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011, 29(7):644-652.

128. Schulz MH, Zerbino DR, Vingron M, Birney E: Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012, 28(8):1086-1092.

129. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al*: De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010, 7(11):909-912.

130. Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z: Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* 2010, 11:663.

131. Surget-Groba Y, Montoya-Burgos JI: Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res* 2010, 20(10):1432-1440.

132. Martin J, Zhu W, Passalacqua KD, Bergman N, Borodovsky M: Bacillus anthracis genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics* 2010, 11 Suppl 3:S10.

133. Nicol JW, Helt GA, Blanchard SG, Jr., Raja A, Loraine AE: The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 2009, 25(20):2730-2731.

134. Bullard JH, Purdom E, Hansen KD, Dudoit S: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010, 11:94.

135. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008, 18(9):1509-1517.

136. Jones DC, Ruzzo WL, Peng X, Katze MG: A new approach to bias correction in RNA-Seq. *Bioinformatics* 2012, 28(7):921-928.

137. Srivastava S, Chen L: A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* 2010, 38(17):e170.

138. Durbin R, Eddy S, Krogh A, Mitchison G: Biological sequence analysis. Cambrage, UK.: Cambridge University Press; 1998.

139. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008, 320(5881):1344-1349.

140. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B *et al*: Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* 2009, 4(10):e7526.

141. Yus E, Guell M, Vivancos AP, Chen WH, Lluch-Senar M, Delgado J, Gavin AC,

Bork P, Serrano L: Transcription start site associated RNAs in bacteria. *Mol Syst Biol* 2012, 8:585.

142. Makino K, Kim SK, Shinagawa H, Amemura M, Nakata A: Molecular analysis of the cryptic and functional phn operons for phosphonate use in Escherichia coli K-12. *J Bacteriol* 1991, 173(8):2665-2672.

143. Hove-Jensen B, Rosenkrantz TJ, Zechel DL, Willemoes M: Accumulation of intermediates of the carbon-phosphorus lyase pathway for phosphonate degradation in phn mutants of Escherichia coli. *J Bacteriol* 2010, 192(1):370-374.

144. Iqbal S, Parker G, Davidson H, Moslehi-Rahmani E, Robson RL: Reversible phase variation in the phnE gene, which is required for phosphonate metabolism in Escherichia coli K-12. *J Bacteriol* 2004, 186(18):6118-6123.

145. Jochimsen B, Lolle S, McSorley FR, Nabi M, Stougaard J, Zechel DL, Hove-Jensen B: Five phosphonate operon gene products as components of a multi-subunit complex of the carbon-phosphorus lyase pathway. *Proc Natl Acad Sci U S A* 2011, 108(28):11393-11398.

146. Chen CM, Ye QZ, Zhu ZM, Wanner BL, Walsh CT: Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the phn (psiD) genes involved in alkylphosphonate uptake and C-P lyase activity in Escherichia coli B. *J Biol Chem* 1990, 265(8):4461-4471.

147. Metcalf WW, Wanner BL: Evidence for a fourteen-gene, phnC to phnP locus for phosphonate metabolism in Escherichia coli. *Gene* 1993b, 129(1):27-32.

148. Kononova SV, Nesmeyanova MA: Phosphonates and their degradation by microorganisms. 2002, 67:184-195.

149. Shi W, Zhou Y, Wild J, Adler J, Gross CA: DnaK, DnaJ, and GrpE are required for flagellum synthesis in Escherichia coli. *J Bacteriol* 1992, 174(19):6256-6263.

150. Rashid MH, Rao NN, Kornberg A: Inorganic polyphosphate is required for motility of bacterial pathogens. *J Bacteriol* 2000, 182(1):225-227.

151. Filiatrault MJ, Stodghill PV, Bronstein PA, Moll S, Lindeberg M, Grills G, Schweitzer P, Wang W, Schroth GP, Luo S *et al*: Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. *J Bacteriol* 2010, 192(9):2359-2372.

152. Jager D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA: Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* 2009, 106(51):21878-21882.

153. Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR: Evidence for a major

role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol* 2009, 5:305.

154. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT: Widespread antisense transcription in Escherichia coli. *MBio* 2010, 1(1).

155. Neidhardt FC, Curtiss III R, INgraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE: EcoSal : Escherichia coli and Salmonella : cellular and molecular biology. Washington D.C.: ASM Press; 2002.

156. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: The EcoCyc Database. *Nucleic Acids Res* 2002, 30:56-58.

157. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J: Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet* 2005, 21(1):16-20.

158. Busby S, Ebright RH: Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* 1994, 79(5):743-746.

159. Browning DF, Busbym SJW: The regulation of bacterial transcription initiation. *Nat Rev Microbiol* 2004, 2:57-65.

160. Hershberg R, Altuvia S, Margalit H: A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res* 2003, 31(7):1813-1820.

161. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003, 4(5):P3.

162. Edwards MT, Rison SC, Stoker NG, Wernisch L: A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic Acids Res* 2005, 33(10):3253-3262.

163. Gripenland J, Netterling S, Loh E, Tiensuu T, Toledo-Arana A, Johansson J: RNAs: regulators of bacterial virulence. *Nat Rev Microbiol* 2010, 8(12):857-866.

164. Su Z, Olman V, Mao F, Xu Y: Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acid Res* 2005, 33(16):5156-5171.

APPENDIX A: LINKS OF SUPPLEMENTARY DATA OF EACH CHAPTER

The additional files for chapter 1 can be downloaded from
http://www.biomedcentral.com/1471-2164/11/527

The supplementary files for chapter 2 and chapter 3 can be downloaded from
http://bioinfolab.uncc.edu/ShanLi_dissertation_supplementary/

VITA

Shan Li
University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC 28223
Cell Phone: (704) 807-7481
E-mail:lishan989@gmail.com

EDUCATION	University of North Carolina Charlotte, Charlotte, NC, USA Candidate for PhD in Computing and Informatics, Department of Bioinformatics and Genomics, spring 2013 (GPA:3.816/4.0). Shandong University, Jinan, Shandong, China Bachelor of Science in Mathematics, fall 2005 (GPA: 3.7/4.0)
EXPERTISE	• Mathematical modeling and combinatorial optimization • Approximation algorithms design and complexity analysis • Parallel computing and cluster computing • Large-scale data mining • Programming in Perl (proficient), C/C++ (proficient), Matlab (proficient), R (prior experience), etc. • Familiar with Linux and Microsoft Windows
PUBLICATIONS	1. *Shan Li*, Xia Dong and Zhengchang Su. Reconstruction of Operon Structures in Prokaryotes from Short Directional RNA-seq Reads Using a Hidden Markov Model. (under review) 2. *Shan Li*, Minli Xu and Zhengchang Su. Computational analysis of LexA regulons in Cyanobacteria. BMC Genomics 2010, 11:527.(PMID: 20920248) 3. Shaoqiang Zhang, *Shan Li*, Meng Niu, Phuc Pham and Zhengchang Su. MotifClick: prediction of cis-regulatory binding sites via merging cliques. BMC Bioinformatics, 2011 Jun 16;12:238. (PMID:21679436) 4. Zhang S, *Li S*, Pham PT, Su Z. Simultaneous prediction of transcription factor binding sites in a group of prokaryotic genomes. BMC Bioinformatics. 2010;11:397. (PMCID: 2920276) 5. Shaoqiang Zhang, MinLi Xu, *Shan Li*, and Zhengchang Su. Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. Nucleic Acids Research, 2009. doi:10.1093/nar/gkp248. (PMID: 19383880)

SOFTWARE	• TruHmm is a RNA-seq reads assembly program for prokaryotes, coded by standard C++ and perl, and compiled by GNU C++ compiler under Linux, Mac and Cygwin (http://bioinfolab.uncc.edu/TruHmm_package/)
RESEARCH EXPERIENCE	2008 – Present:: Pre-dissertation Research (Research Assistant) PI: Dr. Zhengchang Su, Department of Bioinformatics and genomics, UNC Charlotte • Reconstruction dynamic transcriptome of Escherichia coli K12 using RNA-seq data based on a Hidden Markov Model • Computational analysis of LexA regulons in Cyanobacteria • Co-evolution of a transcription factor and its cis-regulatory binding sites: a lesson learned from the phoB and its binding sites in bacteria May – June 2010: Reviewing manuscript for Bioinformatics journal January 2007 – May 2008: Rotation project PI: Dr. Xintao Wu, Department of Computer Science, UNC Charlotte • Prediction of protein-protein interaction network
HONORS AND AWARDS	2012: Travel Fellowship of ISMB conference, Long Beach, CA 2007 – 2012: UNC Charlotte Graduate Assistant Scholarship 2003 – 2005: Scholarship in Shandong University, China 1999: Second prize in the National Physics Contest, Shandong Province, China 1998: Second prize in the National Olympic Mathematics Contest for High School students in China
TEACHING EXPERIENCE	2009 – 2010: Teaching Assistant, BINF 6200 (ITSC 8200) UNC Charlotte
CONFERENCE AND MEETINGS	• Poster presentation in ISMB, July 2012, Long Beach, CA • Podium presentation in UNCC Graduate Research Fair, March 2010

MEMBERSHIP	• Member of AAAS/Science • Member of ISCB