UNDERSTANDING GENOME-WIDE GENE REGULATION IN PROKARYOTES USING OMICS DATA

by

Seyed Ehsan Seyedi Tabari

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
Bioinformatics and Computational Biology

Charlotte

2016

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Anthony Fodor

_____
Dr. Jun-tao Guo

_____
Dr. Xinghua Shi

_____
Dr. Bao-Hua Song

ABSTRACT

SEYED EHSAN SEYEDI TABARI. Understanding of genome-wide gene regulation in prokaryotes using omics data. (Under the direction of DR. ZHENGCHANG SU)

Gene annotation is a critical step for understanding functions of genomes and host organisms. However, accurately annotating thousands of sequenced genomes can be a challenging task. Further, although gene transcriptional regulation is crucial for many important biological functions of cells, our understanding of the complexity of this process in even prokaryotes is still limited. This dissertation addressed both of these problems. First, we developed a fast and scalable tool, PorthoMCL for annotating gene functions through identifying orthologous genes in a large number of prokaryotic genomes. Using this tool, we have predicted orthologous genes in the thousands of sequenced prokaryotic genomes for public use. Second, we systematically investigated the complexity of transcriptomes in *E. coli* K12 in response to a variety of environmental changes. By adopting a model for alternate splicing isoforms in eukaryotes, we revealed that ~22% of operons exhibited different forms of transcriptional units. i.e., alternative operon utilizations, and ~36% operons displayed varying transcriptional levels of their genes, i.e., dynamic operon utilizations, at different growth phases and culture conditions. Moreover, by simultaneously profiling directional transcriptomes and proteomes of *E. coli* K12 cells, we found that a varying portion of genes had antisense RNA (asRNAs) transcription in a growth phase- and culture condition-dependent manner. The detected asRNAs were generally short and overlapped the previously identified asRNAs. Intriguingly, the correlation between genes' protein levels and mRNA levels was

disrupted by increased relative expression levels of asRNA to mRNA, suggesting that asRNA may play an important role in gene expressional regulation.

DEDICATION

To whom I owe everything
and miss forever,
my late mother and father.

ACKNOWLEGMENTS

My greatest appreciation goes to my advisor, Dr. Zhengchang Su, for his support, guidance, encouragement and patience over the years. He is, without doubt, a distinguished mentor with a tireless dedication to science, unwearyingly providing tremendous support for all around him. This dissertation would have not been possible without his invaluable guidance and help.

I most sincerely thank my committee members Dr. Anthony Fodor, Dr. Jun-tao Guo, Dr. Xinghua Shi, and Dr. Bao-Hua Song for their valuable help and kind guidance throughout the years I have had the honor of knowing them.

I would like to thank my colleagues in our lab, Shan Li, Chen Xu, and especially Meng Niu. Meng has been an amazing friend during my years at UNC-Charlotte, as well as a great colleague. I am grateful to call him a best friend and hope to continue to do so in the years to come.

Additionally, I would like to express my deepest gratitude to my family for their unconditional love and support. The support and encouragement from my parents, Mr. Reza Tabari and Ms. Tahereh Zahedi, enabled me to pursue this route. The love and friendship of my siblings, Iman, Armin and Narges, kept me going.

Finally, 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.

INTRODUCTION

A well-stablished approach for gene annotation is through comparative genomics studies where functional similarity is established by finding orthologous genes of a gene in other species (Alexeyenko et al. 2006). Orthologs are genes in different species derived from the last common ancestor through speciation events and generally share the same biological functions in their host genomes. On the other hand, paralogs are genes that are resulted from gene duplication within a species, while their sequences can be highly conserved, paralogs may have different biological functions. Depending on whether duplication occurred before or after speciation, they are called outparalogs or inparalogs, respectively (Sonnhammer and Koonin 2002). Most existing orthology analysis tools, such as COG (Tatusov et al. 2003), InParanoid (Remm, Storm, and Sonnhammer 2001), OrthoMCL (Li, Stoeckert, and Roos 2003), and orthAgogue (Ekseth, Kuiper, and Mironov 2014) rely on sequence similarity whereas some tools use additional information such as synteny and other patterns of concomitant evolution. While multiple tools and databases have been developed for predicting orthologs (Alexeyenko et al. 2006), applying them to an ah-hoc set of genomes is not a straight forward task. Moreover, these tools are computationally expensive, in particular, with the exponential increase in the number of sequenced bacterial genomes as the cost of sequencing is decreasing.

On the other hand, prokaryotic genomes in comparison to the ones in eukaryotes are relatively small and less complex and they lay a perfect platform to study biological functions and their mechanisms. A prokaryotic genome usually consists of a circular chromosome of varying sizes encoding about a few thousand genes whose transcription

and resulting RNA and proteins determine the physiology of the organism. Therefore, a good understanding of their gene structures and transcriptional regulation is of upmost importance. In contrast to eukaryotes in which a gene consisting of multiple exons and introns is transcribed into a transcript that is subsequently spliced into one of more isofoms, in prokaryotes multiple genes arranged in tandem in the same strand of DNA are often transcribed into a single transcript by sharing the same promoter and terminator (Jacob et al. 1960). Such a string of co transcribed genes is called an operon, and in most cases, genes in an operon are involved in the same biological functions (Chuang et al. 2012). Hence, elucidation of operon structures in a genome can facilitate functional annotation of the genes.

Furthermore, the rapid advancement in sequencing technology has produced massive amount of RNA-Seq data that has sparked rethinking of prokaryotic operon structures. In recent years, it has been revealed that the structure of the operon is not as static as once thought and an operon can be transcribed into various transcription units (TUs) under different conditions (Cho et al. 2009). Moreover, extensive study of TU structures under multiple conditions has revealed that activation or repression of operon genes do not happen only at the beginning or the end of the operon, but also in the internal genes of the operons (Güell et al. 2009). These massive quantities of genomic and transcriptomic data available today generates an urgent need for new functional annotation tools that could take advantage of that and broaden our knowledge.

Moreover, until very recently bacterial transcriptomes have been considered to consist of mRNAs, rRNAs, tRNAs and some small RNAs. However, the advancements in high throughput sequencing technologies in the past few years is also challenging this

notion, since numerous studies revealed pervasive transcription from the reverse strands of protein coding genes, resulting in *cis*-antisense RNAs (asRNAs). But unfortunately, the lack of standard protocols and methods to find and categorize asRNAs has produced conflicting reports on the number of the genes in various prokaryotes that have antisense RNA (asRNA) transcription. Such highly varying reports even exist in the well-studied organisms such as *E. coli* K12 strain and casts doubts on the authenticity of most of the asRNAs in the bacterium (Slonczewski 2010). Futher, low asRNA conservation between *E. coli* K 12 and a closely related species *S. enterica* serovar Typhimurium raises more doubts that the majority of prokaryotic asRNA may have any biological functions (Raghavan, Sloan, and Ochman 2012).

In this dissertation, we address the complexities in functional annotation and transcriptomes in bacteria. First, we introduce PorthoMCL, a fast and scalable tool, to predict gene orthology among a large number of genomes. PorthoMCL can be run on a single machine or in parallel on HPC computer clusters and can facilitate comparative genomics analysis through exploiting the exponentially increasing number of sequenced genomes. Second, we analyzed alternative and dynamic operon utilizations in *E. coli* K12 at multiple time-points in three different stress conditions. Finally, we took a systems approach by simultaneously determining the transcriptomes and proteomes in *E. coli* K12 at different growth phases/time points under five culture conditions using a highly specific directional RNA-seq technique and a quantitative mass spectrometric technique, coupled with western blot validation of select genes. Our results suggest that asRNAs may directly or indirectly regulate translation and may play an important role in the

bacterium's responses to environmental changes during growth and adaption to different environments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

asRNA	antisense RNA
BLAST	Basic local alignment search tool
GB	Gigabytes
HPC	High performance computing
LB	Luria broth
MOPS	3-(N-morpholino) propanesulfonic acid
MCL	Markov clustering
MPI	Message passing interface
NPPH	Number of peptides per hundred amino acids
OD	Optical density
ORF	Open reading frame
RPK	Reads per kilo-base
SQL	Structured query language
TPC	Time-point/condition
TPM	Transcripts per million
TU	Transcription unit

CHAPTER 1: PORTHOMCL: PARALLEL ORTHOLOGY PREDICTION USING MCL FOR THE REALM OF MASSIVE GENOME AVAILABILITY

1.1 Background

Orthologs are genes in different species derived from the last common ancestor through speciation events. Orthologous genes generally share the same biological functions in their host genomes. Therefore, identification of orthologous genes among a group of genomes is crucial to almost any comparative genomic analysis (Alexeyenko et al. 2006). In contrast, paralogs, which are genes that are resulted from gene duplication within a species, may have different functions, though their sequences can be highly conserved. Depending on whether duplication occurred before or after speciation, they are called outparalogs or inparalogs, respectively (Sonnhammer and Koonin 2002). Thus, a major challenge in predicting orthologs of a gene is differentiating its orthologs from the orthologs of its paralogs.

Furthermore, due to the rapid advancement in sequencing technologies, sequencing a prokaryotic genome now occurs at an unprecedentedly fast speed and low cost. As a result, tens of thousands of prokaryotic genomes have been fully sequenced, and this number will soon reach hundreds of thousands. The availability of a large number of completed genomes makes comparative genomics an increasingly powerful approach for genome annotations, thereby addressing many important theoretical and application-based problems. However, the rate at which genomes are sequenced outpaces

that at which CPU speed increases. This poses a great challenge in comparative genomics that requires faster algorithms or adaptation of existing tools in parallel environments.

OrthoMCL (Li et al. 2003) is one of the most widely used algorithms for predicting orthologous genes across multiple genomes. Similar to many other orthology prediction algorithms (Gabaldón and Koonin 2013; Kuzniar et al. 2008), OrthoMCL is based on reciprocal best hits in all-against-all BLAST searches (ALTSCHUL et al. 1990) of complete proteomes of the genomes followed by applying the Markov Clustering algorithm (MCL) (Enright, Dongen, and Ouzounis 2002) to a weighted graph constructed based on these best hits (Dongen 2000; Enright et al. 2002). Specifically, OrthoMCL represents genes as nodes in the graph, and connects two nodes/genes by an edge if there are a pair of reciprocal best hits with a similarity greater than a cutoff. The weight of the edges is a normalized score ($\overline{w}$) based on the E-values of the reciprocal hits. This score for genes x_A and y_B in genomes A and B, respectively, is calculated using the following formulas:

$$w(x_A, y_B) = -\frac{\log_{10} \mathbf{Evalue}(x_A \to y_B) + \log_{10} \mathbf{Evalue}(y_B \to x_A)}{2} \tag{1.1}$$

$$\overline{w}(x_A, y_B) = \frac{w(x_A, y_B)}{\underset{\forall \alpha, \beta}{\text{average}}\big(w(\alpha_A, \beta_B)\big)} \tag{1.2}$$

Similarly, within-genome reciprocal hits that have a better normalized score than between-genomes hits are identified as paralogs (Li et al. 2003). Ortholog and paralog groups are then identified by finding the heavily connected subgraphs using the MCL (Enright et al. 2002). However, OrthoMCL relies on a relational database system to store the BLAST results and issues SQL commands to find reciprocal best hits, making it computationally inefficient when the number of genomes becomes large.

To overcome this problem and to speed up the method further, we developed PorthoMCL, a parallel orthology prediction tool using MCL. In addition to the parallelization, our sparse file structure that is more efficient makes PorthoMCL ultrafast and highly scalable. Furthermore, PorthoMCL is platform independent, thus can be run on a wide range of high performance computing clusters and cloud computing platforms.

1.2 Implementation

1.2.1 Workflow

The workflow of PorthoMCL is similar to that of OrthoMCL (Figure 1.1). However, instead of depending on an external database server, PorthoMCL uses a sparse file structure for more efficient data storage and retrieval. In addition, we parallelized all the computationally intensive steps of OrthoMCL. First, PorthoMCL conducts all-against-all BLAST searches in parallel by performing individual-against-all BLAST searches for every genome independently. Second, it identifies the best between-genomes BLAST hits for each two genomes A and B in parallel by scanning the individual-against-all BLAST results. The BLAST hit for the gene x_A in genome B ($x_A \rightarrow y_B$) is considered to be the best hit if the E-value for x_A to gene y_B is the best E-value for all the searches of x_A for genes in genome B with E-

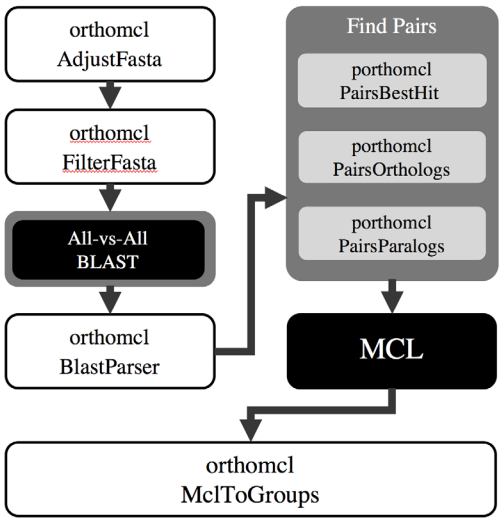


Figure 1.1: Flowchart of PorthoMCL. Original OrthoMCL steps are shown in white, and PorthoMCL steps are in grey shades. Black boxes are the external applications that PorthoMCL requires.

value/match-percentage better than the threshold. This step results in a single best hit file for each genome, and a self-hit file for paralogy-finding. Third, the algorithm finds reciprocal best hits between every two genomes and calculates the normalized score in parallel using Formula 1.2. This is the most computationally intensive step in the algorithm, so we used a sparse file for storage in addition to parallel processing, similar to the strategy used in orthAgogue (Ekseth et al. 2014). Specifically, for each parallel process, PorthoMCL loads at most two best-hit files at the same time to reduce the memory footprint, and every best-hit file is only loaded once to lower the I/O costs. Finally, PorthoMCL finds within-genomes reciprocal best hits and normalizes the score with the average score of all the paralog pairs that have an orthologs in other genomes.

These step are embarrassingly parallel computing problems and do not require shared memory, process coordination or data exchange platform (Graham, Woodall, and Squyres 2005) as used in orthAgogue. Hence, these steps are readily designed to be executed in parallel on a variety of high performance computing (HPC) environments. However, these steps are not totally independent as each step needs the output of the preceding step. The output of these steps are eventually collated to construct a sequence similarity graph that is then cut by the MCL program to predict orthologous and paralogous gene groups.

1.2.2 High Performance Computing Support

PorthoMCL is designed to predict orthologs in a very large number of sequenced genomes in a HPC environments, such as computing clusters or cloud computing platforms without the need of a database server or Message Passing Interface, which is an advantage over OrthoMCL and orthAgogue. We have included a TORQUE script in the

repository to facilitate its use in such environments. However, PorthoMCL also runs on a desktop or a server with minimal requirement using the provided wrapper script.

1.3 Results

To compare the computational efficiency of PorthoMCL and OrthoMCL, we applied the two programs to 10, 50, 100 and 500 randomly selected bacterial genomes. As OrthoMCL was not implemented for parallel computing, we ran both programs on a single computing node with four cores and 32GB of RAM to make the comparison fair. As shown in Table 1.1, PorthoMCL outperformed OrthoMCL in all sizes of datasets in runtime, and it is noteworthy noting that OrthoMCL failed to handle the data size of 500 genomes due to a memory error.

To illustrate the power of PorthoMCL, we applied it to 2,758 sequenced bacterial genomes obtained from GenBank using their annotated protein sequences. These genomes contain a total of 8,661,583 protein sequences with a median length of 270 amino acids. These sequences serve as both the query and the database for all-against-all BLAST searches. For this application, PorthoMCL split the query sequences into smaller files each containing about 10,000 sequences, and ran in the parallel mode on a cluster with 60 computing nodes (each node has 12 cores and 36GBs of RAM). PorthoMCL finished the job in 18 days, of which it spent 11 and 7 days on BLAST searches and the remaining steps that would have taken 549 and 1,634 days, respectively, if run on a single node. In contrast, OrthoMCL could not finish the job after 35 days running on a database server with 40 cores and 1TBs of RAM.

PorthoMCL identified 763,506,331 ortholog gene pairs and identified 230,815 ortholog groups in these genomes. The orthologous pairs (file size: 6.2GB) and

Table 1.1: Comparison of runtimes of OrthoMCL and PorthoMCL for different number orthologous groups (file size: 50MB) as well as paralogous pairs are available for of genomes.

Genomes	Proteins	BLAST Hits	OrthoMCL	PorthoMCL	Speedup
10	19,240	298,647	0:00:18	0:00:11	164%
	29,912	637,091	0:01:07	0:00:21	319%
	30,111	656,689	0:01:16	0:00:23	330%
	32,962	721,997	0:01:12	0:00:24	300%
50	126,020	5,771,483	0:15:55	0:05:55	269%
	127,724	6,363,917	0:27:53	0:06:08	455%
	133,974	6,418,035	0:08:29	0:06:15	136%
	138,258	7,008,798	0:24:06	0:06:18	383%
100	252,109	18,326,608	1:02:58	0:31:49	198%
500	1,327,716	283,850,847	-	17:38:55	∞

orthologous groups (file size: 50MB) as well as paralogous pairs are available for download at http://ehsun.me/go/porthomcl. We will periodically update our predictions when more genomes are available in the future. The options and arguments needed at each step are discussed in detail in the documentation of the PorthoMCL package that can be freely accessed from http://github.com/etabari/PorthoMCL.

1.4 Conclusion

PorthoMCL is fast tool with minimal requirements for identifying orthologs and paralogs in any number of genomes. While PorthoMCL uses the same mathematical basis as OrthoMCL to investigate orthology among genomes, it is much faster and a more scalable tool when handling a very large number of genomes than existing tools. PorthoMCL can facilitate comparative genomics analysis through exploiting the exponentially increasing number of sequenced genomes.

CHAPTER 2: ALTERNATIVE AND DYNAMIC OPERON UTILIZATIONS IN *E. COLI* IN RESPONSE TO ENVIRONMENTAL CHANGES

2.1 Background

Prokaryotic genomes typically consist of a circular chromosome of a few million base pairs encoding a few thousand genes. Multiple genes arranged in tandem with the same orientation are often transcribed into a single polycistronic transcript by sharing the same promoter and terminator (Jacob et al. 1960). Such co-transcribed genes are called an operon. In most cases these genes have similar or coordinated biological functions, and are involved in related biological pathways (Chuang et al. 2012; Overbeek et al. 1999; Salgado et al. 2000; Siefert et al. 1997; Wolf 2001). Understanding the architecture of operons in a prokaryotic genome is important to understand many aspects of the biology of the organism, and can help to predict functions of novel genes (Wang, MacKenzie, and White 2015).

However, recent applications of tiling arrays and new sequencing technologies in combination with mRNA enrichment techniques have revealed that the structures of operons are not as static as previously assumed, instead, an operon can be transcribed in various forms under different conditions, a phenomenon called alternative operon transcription or utilizations (Cho et al. 2009; Güell et al. 2009; Mao et al. 2014; Salgado et al. 2013; Sorek and Cossart 2010). Furthermore, genes in an operon can have varying-levels of transcriptions in a staircase-like manner, a phenomenon known as dynamic operon transcription or utilizations (Güell et al. 2009). The extent of such alternative and

dynamic operon utilizations can be comparable to alternative splicing in eukaryotes in which different mRNA isoforms with varying levels of expression can be produced from a pre-RNA molecule through alternative splicing (Güell et al. 2009). With this regard, the term operon might be more appropriately reserved for the longest set of adjacent genes from which some or all of the genes can transcribed as a transcriptional unit (TU) under certain conditions as previously suggested (Okuda et al. 2007). In other words, an operon can have different TUs, each results in a distinct transcript containing one or more genes.

Numerous studies have attempted to redefine operon or TU maps using RNA-seq techniques of eubacterial and archaeal species such as *L. innocua* (Toledo-Arana et al. 2009), *E. coli* (Cho et al. 2009; Conway et al. 2014), *B. anthracis* (Passalacqua et al. 2009), *L. monocytogenes* (Oliver et al. 2009), *S. enterica* serovar Typhi (Perkins et al. 2009; Wang et al. 2015), *B. cenocepacia* (Yoder-Himes et al. 2009), *S. solfataricus* P2 (Wurtzel et al. 2010), *Helicobacter pylori* (Sharma et al. 2010), *C. trachomatis* (Albrecht et al. 2010), *M. hyopneumoniae* (Güell et al. 2009; Siqueira, Schrank, and Schrank 2011), *S. elongatus* PCC 7942 (Vijayan, Jain, and O'Shea 2011), *M. gallisepticum* (Mazin et al. 2014), *C. thermocellum* (Chou et al. 2015). However, majority of these studies have only investigated changes in operon or TU structures under a single culture conditions, thus only a small portion of alternative operons have been revealed in these species, and dynamic transcription of operons has been largely ignored in most of these studies. Consequently, the patterns of changes and functional implications of alternative and dynamic operons utilizations in response to environmental changes remain largely unknown.

On the other hand, a challenge in studying alternative and dynamic operon utilizations using RNA-seq is to accurately assemble TUs and detect varying transcriptional levels of genes in operons due to the highly uneven coverage of short reads along the gene body (Dillies et al. 2012) and lack of sufficient training datasets (Chou et al. 2015). A few methods have been developed to predict TUs based on RNA-seq data. One of these methods is TruHMM (Li, Dong, and Su 2013) that uses a two-state hidden Markov model to model expressed and non-expressed part of a genomic sequence. The other methods attempt to predict whether or not a pair of consecutive genes on the same strand belong to a single TU based on the mapping of RNA-seq reads and other genomic features using machine learning algorithms. For instance, Rockhopper (McClure et al. 2013) assembles operon structures based on intergenic distance and correlation of expression levels of adjacent genes using a naïve Bayes classifier. SeqTU (Chou et al. 2015) detects the boundaries between adjacent TUs with the same orientation based on similar features using a support vector machine, and Fortino *et al.* (Fortino et al. 2014) combines well-studied genomic features (Brouwer, Kuipers, and van Hijum 2008) and transcriptomics data using multiple machine learning methods to investigate condition dependent TUs. Among these methods, RockHopper has been successfully applied in other studies (Chetal and Janga 2015; Fitzgerald, Bonocora, and Wade 2014; Pflaum et al. 2015; Saadeh et al. 2015; Wang et al. 2015), while others are compromised by their insufficient usability.

To address these questions, we analyzed alternative operon utilizations in *E. coli* K12 at multiple time-points in three different stress conditions, including heat shock (HS), phosphorous starvation (M-P) and carbon starvation (M-C) based on TUs

assembled using RockHopper. In addition, we analyzed dynamic transcription of TUs in the samples by adopting a model that has been successfully used in modeling varying transcriptional levels of alternate splicing isoforms in eukaryotes. We found that this model can accurately reflect the extent of dynamic operon transcription. Our results show that about 22% of operons have alternative TU transcription, and up to 36% of TUs display dynamic transcription in response to environmental changes for better adaptation.

2.2 Results

2.2.1 Defining the Structures of TUs in *E. Coli* K12 at Different Growth Phases and Culture Conditions.

We first identified expressed multi-gene TUs in samples collected at each time point under each culture condition (TPC) using Rockhopper (McClure et al. 2013). As shown in Table 1, we identified a similar number (812~835, average 826) of multi-gene TUs in each TPC, 78~80% of which are at least a subset of the 851 multi-gene operons documented in RegulonDB 9.0 (Gama-Castro et al. 2015), indicating that our TU

Table 2.1: The number of TUs detected at each TPC

Sample	No. TUs	RegulonDB Subunit (%)
LB 1.0	812	653 (80%)
M-P0h	834	661 (79%)
M-P2h	831	652 (78%)
M-P4h	834	658 (79%)
HS15min	835	658 (79%)
HS30min	828	659 (80%)
HS1h	816	651 (80%)
M-C1h	825	643 (78%)
M-C2h	821	641 (78%)
Total Distinct TUs	1,172	853 (73%)
Shared TUs	550	494 (90%)

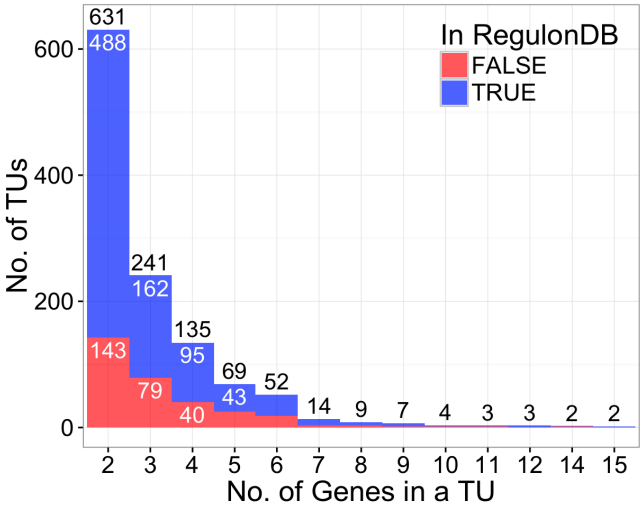


Figure 2.1: TUs detected by Rockhopper

assembling was quite accurate. In total, we identified 1,172 multi-gene TUs in all the nine TPCs, each containing 2~15 genes with an average of three genes (Figure 2.1). Of these 1172 multi-gene TUs, 853 (73%) are at least a subset of multi-gene operons documented in RegulonDB, the other 319 might be novel operons. Furthermore, 550 (46.9%) of the predicted 1,172 multi-gene operons were observed in all of the TPCs, and 494 of which (90%) are at least a subset of multi-gene operons documented in RegulonDB. Their ubiquitous expression in all the time points and culture conditions we examined indicate that these TUs might be involved in house-keeping functions (see below). On the other hand, the remaining 622 (53.1%) were only observed in a limited number (1~8) of TPCs (Figure 2.2A), indicating that they were expressed in a more or less TPC-dependent manner. In this regard, 99 of them (8%) were observed in only one TPC (Figure 2.2A). Interestingly, as shown in Figure 2.2B, TPCs sampled at adjacent time points and under the same culture condition were unambiguously clustered into a group based on the distances among the TU structures of the TPCs (see Methods), i.e., TPCs of LB, HS, M-C, and M-P were clustered in four distinct groups, far away from each other. These results further suggest that these 622 TUs were utilized in the bacterium in a growth phase- and condition-dependent manner.

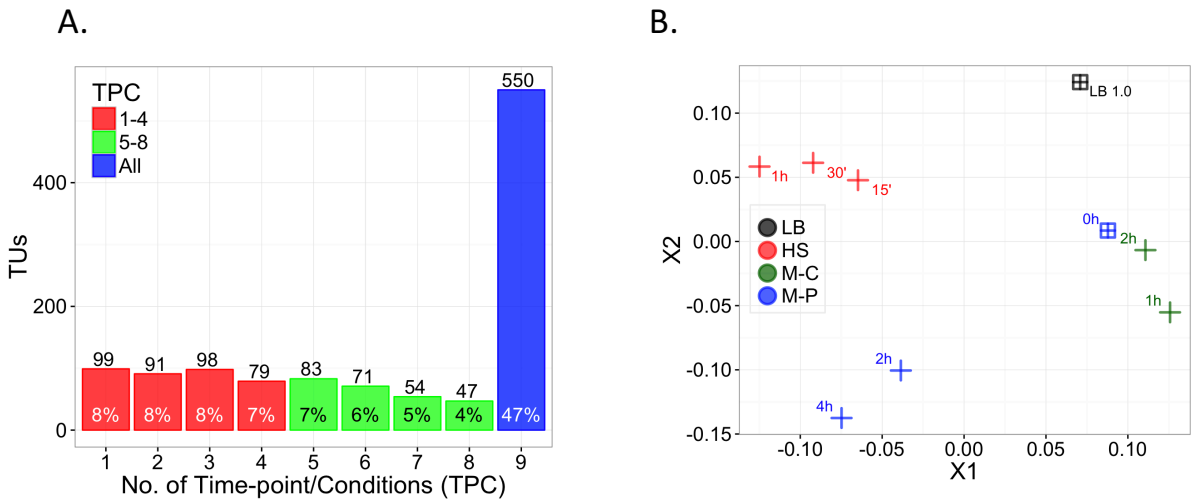


Figure 2.2: CTPs and TUs

2.2.2 A Considerable Portion of Operons Is Alternatively Utilized in a Growth Phase- and Culture-Dependent Manner.

We next examine the patterns of alternative operon utilizations in different growth phases and culture conditions. We found that 676 (57.7%) of the 1,172 TUs assembled in the TPCs were only transcribed in a single TU form in all the TPCs in which they presented, and thus consider them as single-TU operons. In contrast, the other 496 (42.3%) TUs contained at least a shared gene with another TU, suggesting that the TUs that shared genes were alternatively transcribed in different TPCs from a larger operon. To identify these larger operons, we stitched all TUs from all the TPCs that shared at least a gene to form an operon. In doing so, we combined the 496 TUs into 186 larger operons. As shown in Figure 2.3A, most (84.4%) of these stitched operons were transcribed in two or three different TU forms, although a few in more than four forms. Therefore, a considerable portion (22.0%) of the 862 identified operons (186 stitched

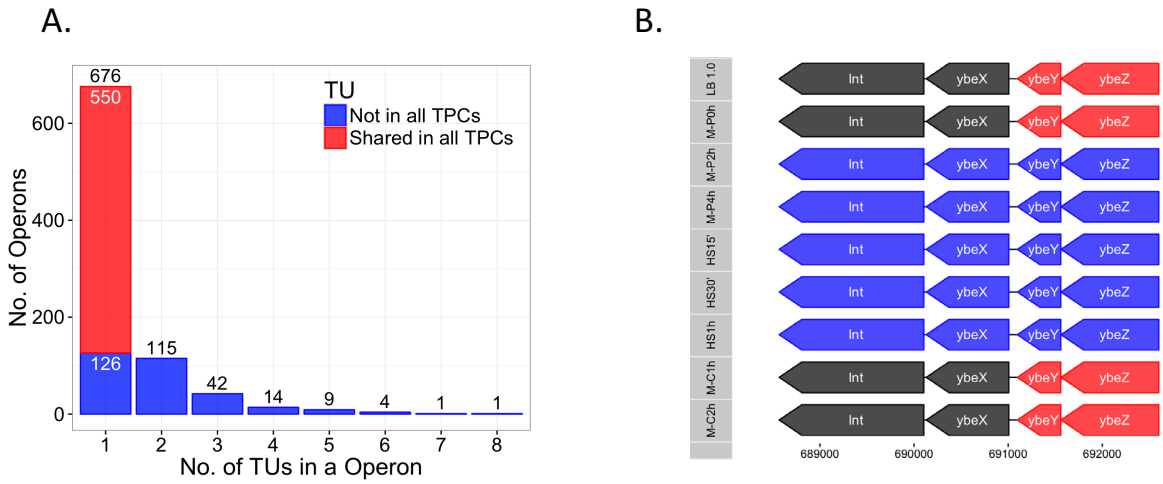


Figure 2.3: Constructed operons and their TUs.

operons plus 676 single-TU operons) were utilized in different forms in different growth phases and culture conditions. Figure 2.3B shows an example of such operons, ybeZYX-lnt, which is under control of the heat shock sigma factor δ^{32} (Nonaka et al. 2006). It has been shown that three different TUs can be transcribed from this operon, ybeZYX-lnt, ybeZY and ybeX-lnt (Salgado et al. 2013) that were all detected in our TPCs. Interestingly, while the full operon is known to be transcribed under heat shock, our results showed that it is was also transcribed in the later stages of phosphorous starvation. Furthermore, the above-mentioned 550 TUs observed in all the TPCs were only transcribed in a single TU form in all the TPCs, thus are a subset of the 676 single-TU operons. Therefore, all these ubiquitously used operons were not alternatively utilized. Functional enrichment analysis confirms our earlier conjecture that these ubiquitously transcribed single-TU operons are involved in house-keeping related biological functions, such as anaerobic/cellular respiration and cell mobility/localization (dataset DS2 in Appendix B).

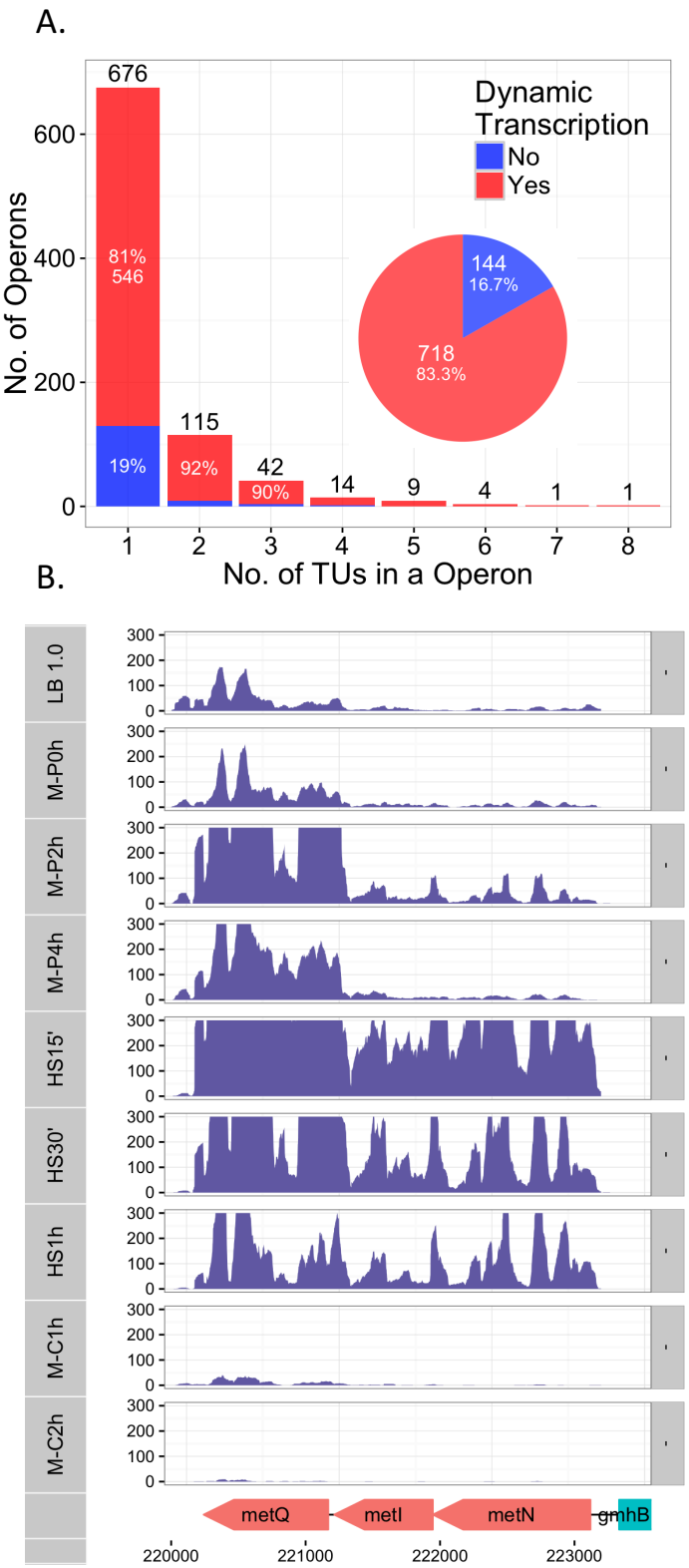


Figure 2.4: Dynamic Transcription of operons.

2.2.3 TU Are Dynamically Transcribed in a Growth Phase- and Culture-Dependent Manner.

Finally, we investigated dynamic expression patterns of TUs in different growth phase and culture conditions. As there is no available tools for such analysis, we adapted the DEXSeq algorithm that was originally developed to model varying transcriptional levels of alternative splicing isoforms in eukaryotes (Anders, Reyes, and Huber 2012) to this purpose (see Methods) by taking the advantage of the similarity between alternative splicing in eukaryotes and dynamic transcription in prokaryotes (Güell et al. 2009). As summarized in Figure 2.4A, we found that most (718, 83%) of the predicted multi-gene TUs

showed clear patterns of dynamic transcription. Dynamic transcription occurred in both single-TU operons and multi-TU operons, although operons with multi-TUs have a higher frequency of dynamic transcription than single-TU operons (Figure 2.4A). Figure 2.4B shows an example of single-TU operons with dynamic transcription, in which the *metNIQ* operon encoding the DL-methionine uptake system (Gál et al. 2002; Merlin et al. 2002) displayed varying transcriptional levels under different culture conditions. Specifically, the *metQ* gene coding for an ATPase, was expressed under all conditions, but its expression levels were much higher than the two other genes of the TU under all the culture conditions (the average expression ration of *metQ* over the other two is: 2.71) except for HS in which all the three genes were expressed in similar levels (the average expression ration of metQ over the other two is: 1.24). The other 144 operons displayed no dynamic transcriptional patterns even though they might be transcribed into multiple TUs (Figure 2.4A). To see how the dynamic transcription of operons vary in different TPCs, we clustered operons based on their expression profiles across all TPCs (see Methods). As shown in Figure 2.5A, the first 80 operons were single-TU operons were expressed in all the TPCs with no dynamic transcription. The next 64 operons (from 81 to 144) did not show any dynamic transcription either, although some of them have alternative transcription or not expressed in some TPCs. These two groups make up the 144 operons without dynamic operon transcription. The next 546 operons (from 145 to 690) were transcribed in a single TU with dynamical transcription dependent on the TPCs. The last 172 (from 691 to 862) operons showed both alternative and dynamic changes in different TPCs.

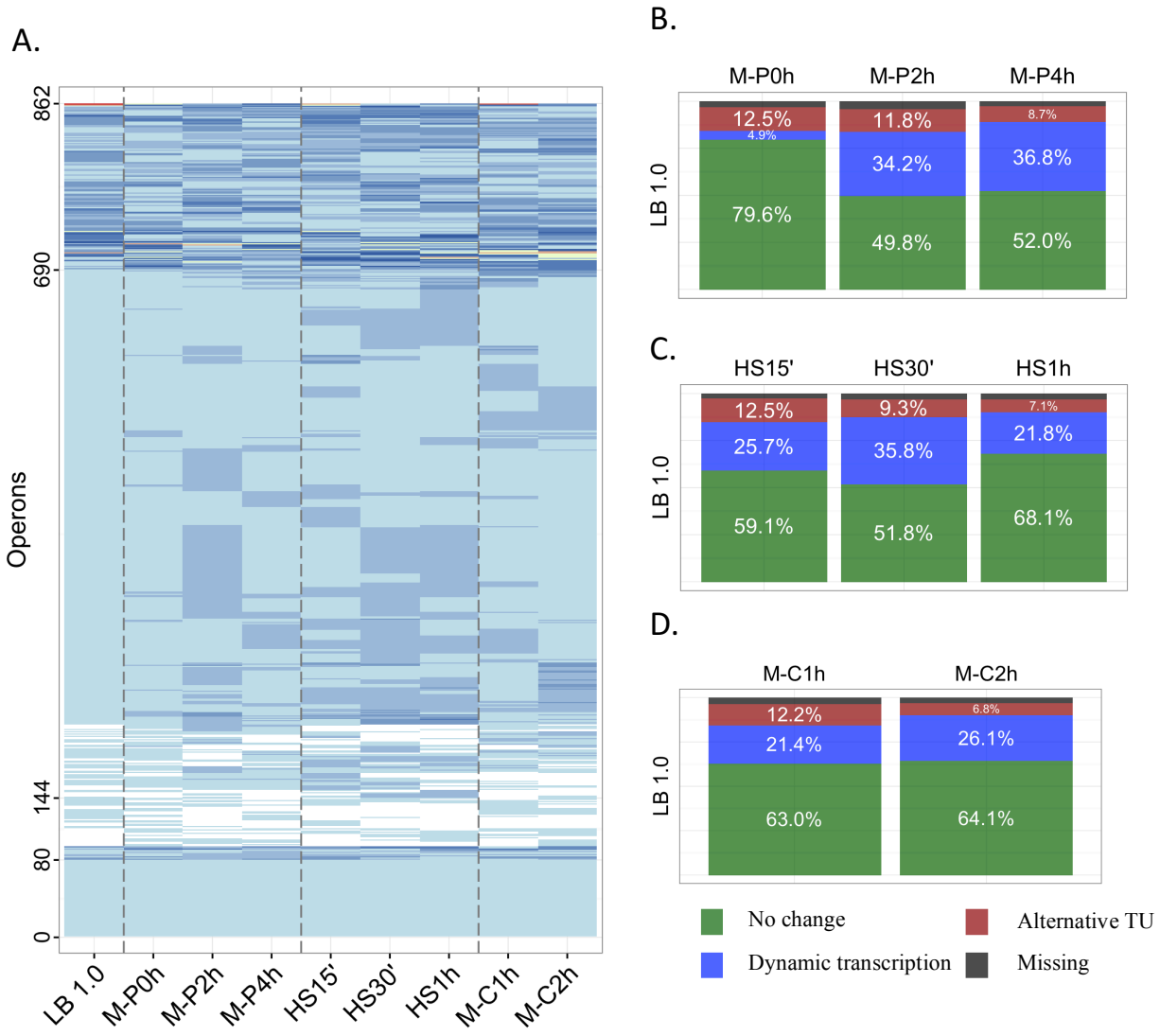


Figure 2.5: Changes in operon transcription at different time-points.
(A) Each line represents an operon, and different color across the line marks a change in transcription of the operon, a TU variation or alternate transcription (white=not expressed). Rate of alternative transcription in heat shock (B), phosphorous starvation (C) and carbon starvation (D) conditions. Operon at each consecutive time-point after LB 1.0 are divided in four groups, operons that were transcribed with no change to the previous time-point (green), operons that are transcribed into the same TU but with an dynamic transcription to the previous time-point (blue), operons that are transcribed into a different TU from the previous time-point (red) and operons from the previous time-point that are missing (black).

We then investigated how alternative and dynamic transcription changed at different growth phases under different culture conditions. To this end, for each culture condition we compared both alternative and dynamic transcriptional patterns of operons

at each time-point with its previous time-point starting from LB 1.0 from which the cells were exposed to different stress conditions. As shown in Figure 2.5B, ~80% of the operons at M-P0h are expressed in the same way as in LB 1.0, 4.9% were dynamically transcribed into the same TUs, and the rest were either alternatively transcribed into a different TU (12.5%) or not transcribed at all (3%). This is expected as the M-P0h samples were collected right after the cells were transferred to phosphorous starvation medium from LB 1.0. In contrast, only 49.8% and 52% of the operons were transcribed in the same way as in the previous time point for the M-P2h and M-P4h samples, respectively, while a larger portion of operons (~34% in M-P2h and ~37% in M-P4h) were dynamically transcribed from the previous time-points. Similarly, in other two stress conditions (Figure 2.5C and D), from 51.8% to 61% of the operons were transcribed in the same ways as in the previous time points, while from 21.4% to 35.8% of the operons were dynamically transcribed. Moreover, the proportion of the operons that were alternatively transcribed into different TUs compared to the previous time points was in the range of 6.8~12.5%, and generally decreased in all culture conditions with time. However, the proportion of operons that were not expressed compared to the previous time-points stayed relatively small and stable (2.5%~4.2%), regardless of the time-points and culture conditions. These results indicate that the bacterium can either alternatively transcribe operons into different TUs, or dynamically fine-tune the transcriptional levels of genes in the operons during the course of responses to environmental changes.

2.3 Discussion

A complex landscape of alternative and dynamic operon utilizations has been reported as a response to environmental changes in a variety of prokaryotic organisms including Gram-positive bacterium *M. pneumoniae* (Güell et al. 2009) and *B. Subtilis* (Nicolas et al. 2012); yet, an extensive investigation of this phenomenon is still missing in model Gram-negative bacterium, *E. coli*. Further, the time course and patterns of alternative and dynamic operon utilizations in bacteria in response various environmental changes were not fully understood. Moreover, although methods have been developed to analyze alternative operon utilizations, the highly non-uniform coverage of RNA-seq reads on transcribed regions (Dillies et al. 2012) render it a challenging task to accurately detect varying transcriptional levels of genes in a operon. To tackle these problems, we proposed to model dynamic operon transcription in the same way as modeling alternative splicing isoforms of genes in eukaryotes by treating genes in an operon as exons of a eukaryotic gene. In this study, we have successfully adapted the DEXSeq (Anders et al. 2012) algorithm to this purpose and revealed similarly complex landscape of alternative and dynamic operon utilizations in *E. coli* K12 cells in response to environmental changes in the course of various stress culture conditions. Our results suggest that transcriptional changes in bacteria in response to environmental changes are achieved through both alternative transcriptions of operons, by which genes in an operon can be selectively transcribed or skipped, and dynamic transcription, by which transcriptional levels of each gene in an operon can be quantitatively fine-tuned.

2.4 Methods

2.4.1 Datasets

The *E. coli* K12 genome and annotation files (version NC_000913.2) were downloaded from Genbank. Experimentally verified operons were downloaded from RegulonDB 9.0 (Gama-Castro et al. 2015). *E. coli* K12 cells were grown and treated as previously described (Li et al. 2013). Briefly, cells were first cultured in the rich medium Luria broth (LB) until a middle log phase (OD_{600}=1.0) was reached. The cells were then transferred to MOPS solution for heat shock (HS, 48°C), or MOPS without glucose (M-C) for carbon starvation, or MOPS without phosphorus (M-P) for phosphorus starvation. Cell samples were taken at different time points of the cultural conditions (TPCs), including at OD_{600}=1.0 growing LB (LB 1.0); right after transferring to the M-P medium (M-P0h), and two and four hours after the onset of phosphorous starvation (M-P2h and M-P4h, respectively); 15, 30 and 60 minutes of heat shock (HS15', HS30' and HS1h, respectively); one and two hours after transferring to the M-C medium (M-C1h and M-C2h, respectively). Directional RNA-seq libraries from the samples were prepared and sequenced as previously described (Li et al. 2013) with at least two biological replicates. The datasets have been deposited in GEO with accession numbers GSE48151 and GSE64021.

2.4.2 Predicting TUs and Reconstructing Operons

We predict TUs in a TPCs by running RockHopper 2.03 with default parameters on RNA-seq reads from all replicates for the TPC, using the protein and RNA annotation files of the bacterium and the LB 1.0 libraries as the baseline. We define a distance between the TU structures TU_i and TU_j of a pair of TPCs i and j, respectively, as

$d(i,j) = |TU_i \cup TU_j|/|TU_i \cap TU_j|$. To analyze and visualize relationships between operon structures in different TPCs, we performed a multidimensional scaling (Gower 1966) using the distance matrix. We reconstructed an operon by collating all the TUs predicted in all the TPCs that share at least a gene.

2.4.3 Modeling Dynamic Operon Transcription

We mapped the reads using bowtie 2.0 (Langmead and Salzberg 2012) with the option "--very-sensitive" and counted uniquely mapped reads to each gene using the GenomicAlignments package in Bioconductor (Lawrence et al. 2013). Then, we used DEXSeq (Anders et al. 2012) with following modifications to calculate the similarity of transcription profiles of a TU between two TPCs.

We assume that the reads mapped to gene *g* in TU *i* in TPC *t*, K_{git}, follow a negative binomial distributions (NB):

$$K_{git} \sim NB\left(mean = s_t \mu_{git}, dispersion = \alpha_{gt}\right)$$

where s_t is a scaling factor for TPC *t* that accounts for sequencing depth; μ_{git} is the linear predictor, which is decomposed into factors that account for the baseline expression of TU *i*, the expression of gene *g*, and the effect of TPC *t* on the expression of TU *i*; and α_{gt} is the dispersion parameter for gene *g* in TPC *t*, which is calculated using a Cox-Reid dispersion (Cox and Reid 1987) estimator (McCarthy, Chen, and Smyth 2012). The false discovery rate (FDR) for each gene in a TU was adjusted using the Benjamini-Hochberg method.

Assuming that there are T different TUs transcribed from an operon *o* ($TU^i, i = 1..T$), and TU^i is detected in n_i TPCs ($TU^i_{t=1..n_i}$), for each TU^i, we compute similarity scores of its transcription profiles between each pair of the n_i TPCs using DEXSeq.

Based on this similarity matrix, we identify TPC clusters in which TU^i is expressed similarly using Ward's hierarchical agglomerative clustering method (Murtagh and Legendre 2014). We consider each resulting cluster as a dynamic transcriptional pattern of TU^i. We define the dynamic transcriptional patterns (P_o) of an operon *o* is the union of the patterns of all its TUs. We sort the patterns in P_o by their sizes and label them by their ranks ($1..|P_o|$). We represent operon *o*'s expression pattern in all the TPCs by a vector ($C_o = [c^o_{LB1.0}, c^o_{MP0h}, \ldots, c^o_{MC2h}]$), where c^o_t is the label of the pattern observed in TPC *t*, or zero if any TU of the operon is not expressed in *t*. To find operons that have similar expression patterns shown in Figure 5A, we clustered operons based on their expression pattern vectors using the Euclidian distance and hierarchical clustering.

CHAPTER 3: ANTISENSE TRANSCRIPTION AND ITS ROLES IN RESPONSE TO ENVIRONMENTAL CHANGES IN *E. COLI* K12

3.1 Background

Bacterial transcriptomes have long been considered to consist of mRNAs, rRNAs, tRNAs, and some small *cis*- and *tran*-acting RNAs. However, in the past few years, applications of high-density directional tiling array, and in particular directional RNA-seq techniques, have revealed pervasive transcription from the reverse strands of protein coding genes, resulting in *cis*-antisense RNAs (hereafter denoted asRNAs). The asRNA molecules overlap with the 5'-end, 3'-end, middle, or the entire gene/operon (Georg and Hess 2011); their lengths vary from tens to thousands of nucleotides (nt) (Lasa, Toledo-Arana, and Gingeras 2012; Thomason and Storz 2010; Wade and Grainger 2014); and are reported in taxonomically distinct species, including: *M. pneumonia* (Güell et al. 2009), *B. anthracis* (Passalacqua et al. 2012), *Synechocystis sp.* PCC 6803 (Georg et al. 2009), *L. monocytogenes* (Toledo-Arana and Solano 2010; Wurtzel et al. 2012), *B. subtilis* (Nicolas et al. 2012; Rasmussen, Nielsen, and Jarmer 2009), *H. pylori* (Sharma et al. 2010), *P. syringae* (Filiatrault et al. 2010), *E. coli* (Dornenburg et al. 2010; Lybecker et al. 2014; Raghavan et al. 2012; Selinger et al. 2000; Thomason et al. 2015), *S. enterica* (Kröger et al. 2012), and *S. aureus* (Lasa et al. 2011). The highly pervasive asRNA transcription reported in some species strongly suggests that they may play important roles in the physiology of prokaryotes as the synthesis of asRNA is costly (Georg and Hess 2011). More recently, it was found that in Gram positive bacteria such as *S. aureus*

(Lasa et al. 2011; Lioliou et al. 2010) and *B. subtilis* (Lasa et al. 2011) asRNA expression may provide a unique genome-wide post-transcriptional regulatory mechanism to adjust mRNA levels through invoking RNase III-mediated digestion of mRNA/asRNA duplexes, while a different mechanism may be responsible for the mRNA/asRNA digestion in Gram negative bacterial such as *S. enteritidis* (Lasa et al. 2011, 2012). Moreover, it has been shown that asRNA transcription displayed non-random patterns under different culture conditions in *E. coli* (Dühring et al. 2006; Kawano et al. 2005; Opdyke, Kang, and Storz 2004; Thomason et al. 2015) and *B. anthracis* (Passalacqua et al. 2012); thus, the resulting asRNAs may be involved in the adaptation of the bacteria to the environments. These accumulative lines of evidence strongly suggest that pervasive asRNA transcription may play important roles in bacterial physiology (Lasa et al. 2012; Thomason and Storz 2010; Wade and Grainger 2014).

However, there are still many puzzles about asRNA transcription in prokaryotes. First, a highly varying proportion of open reading frames (ORFs) has been reported to have asRNA transcription in different species, ranging from 2.2% in *G. sulfurreducens* (Qiu et al. 2010) and 5.6% in *Synechocystis sp.* PCC 6803 (Georg et al. 2009), to 13% in *B. subtilis* (Nicolas et al. 2012) and 46% in *H. pylori* (Sharma et al. 2010), and 75% in *S. aureus* (Lasa et al. 2011) and 93% in *E. coli* (Selinger et al. 2000). This raises questions about the ubiquity of asRNA transcription pervasiveness in taxonomically distinct prokaryotes (Wade and Grainger 2014). Ironically, even inconsistent results have been reported in the most well studied *E. coli* K12 strain. For instance, up to 4,000 (~93%) genes in *E. coli* K12 were initially reported to have antisense transcription using a whole genome tiling array technique (Selinger et al. 2000), while two later studies only

identified 1,000 (Dornenburg et al. 2010) and 90 (Raghavan et al. 2012) asRNA species in the bacterium under similar growth conditions using RNA-seq techniques. To further confound the problem, a more recent study found a total of 316 asRNA species in RNase III mutant strains using a new technique that enriches double-stranded RNA using an antibody (Lybecker et al. 2014), whereas another study identified 5,495 asRNA species using a differential RNA-seq technique (Thomason et al. 2015). This inconsistency casts doubts on the authenticity of most of the asRNAs in the bacterium (Slonczewski 2010). Second, while some mammalian asRNAs (van Duin et al. 1989; Faghihi et al. 2010; Ling et al. 2013; Rossignol et al. 2004) and up to 86% of yeast asRNA (Goodman, Daugharthy, and Kim 2013; Swamy et al. 2014; Yassour et al. 2010) are evolutionarily conserved, only 14% of asRNAs are conserved between *E. coli* K 12 and a closely related species *S. enterica* serovar Typhimurium even though both displayed similarly extensive asRNA transcription (28% of ORFs), raising doubts that the majority of prokaryotic asRNA may have any biological functions (Raghavan et al. 2012).

Functional characterization of asRNAs has been hampered by the lack of a technique that disrupts the transcription of an asRNA without affecting the sense transcription (Lasa et al. 2012). To overcome this technical difficulty, thereby characterizing authentic asRNAs as well as their transcriptional patterns, functions and underlying mechanisms, we have taken a systems approach by simultaneously determining the transcriptomes and proteomes in *E. coli* K12 at different growth phases/time points under five culture conditions using a highly specific directional RNA-seq technique and a quantitative mass spectrometric technique, coupled with western blot validation of select genes. We found that asRNA transcription *E. coli* K12 is highly

pervasive, yet highly variable and dynamic, and that many genes change their relative asRNA levels to the mRNA levels at different growth phase/time points and under different culture conditions in well-defined manners. We show that such changes may have functional implications in the bacterium's responses to environmental changes through affecting protein translation directly or indirectly.

3.2 Results

3.2.1 Determination of Time Series Transcriptomes and Proteomes in *E. Coli* K12

Given the intrinsic difficulty of studying asRNAs using traditional genetic disruption methods (Lasa et al. 2012), we tested whether asRNAs could be more effectively studied by using a systems biology approach. To this end, we simultaneously profiled time series transcriptomes and proteomes of *E. coli* K12 using a highly strand-specific RNA-seq method (Li et al. 2013) and a quantitative tandem mass spectrometry method (Lee et al. 2013), respectively, under a variety of culture conditions, including early (OD_{600}=0.5) and middle (OD_{600}=1.0) log phases and the stationary phase (OD_{600}=3.0) growing in the rich medium Luria broth (LB), and at different time-points after transferring the cells growing in the LB (OD_{600}=1.0) to one of four stress conditions: minimum medium MOPS (MOPS), heat shock (HS) in MOPS, carbon starvation (M-C) and phosphorus starvation (M-P). We named sampling time points by concatenating the growth condition and the time point/growth phase at which the cultures were sampled (i.e. HS15' for the sampling time point at 15 minutes after the onset of heat shock, LB 0.5 for the sampling time point when the cell growth reached an OD_{600}=0.5 in LB, and M-P6h for the sampling time point at 6 hours after the onset of phosphorus starvation).

As shown in Figure A1A&B (in Appendix A), both the growth rate and protein levels of the cells in the four stress cultures had varying levels of reduction at all the time

points measured compared with those of the cells in LB, which is consistent with the current understanding of the severity of the stresses that each of these conditions would exert on the cells. Specifically, the cells in the minimum medium MOPS continued to grow with a moderately elevated protein production, while the cells in HS, M-C, and M-P largely stopped growing (Figure A1A), maintaining slightly decreased yet steady-state protein levels, with the exception of the cells in M-C, in which the protein levels initially increased slightly and then dropped markedly later. Interestingly, the differences in the protein levels normalized by the cell density (OD_{600} value) were less distinct among the five cultures, in particular for LB, MOPS, and M-C (Figure A1C), suggesting that the prolonged culture in LB slowed down and stresses (HS, M-C and M-P) almost completely suppressed cell proliferation; and the cells tended to maintain similar protein concentrations, presumably in order to adapt to the harsh environments though active synthesis of required proteins.

The RNA-seq reads obtained from different sequencing platforms for the same sample or from different biological replicates were highly correlated (Figure A2), thus, we pooled the reads of libraries for cells collected at the same growth phase/time point and culture condition for further analyses, and refer the pooled reads as a *sample* for the sampling time point and culture condition. As summarized in Table A1, a total of 1,085,996,619 reads were generated from the 20 samples, with an average of 21.8% of the reads uniquely mapped to the genome. From 2,941 (in M-C4) to 4,497 (in HS30min) of the 4,567 annotated ORFs were expressed in the samples (Table A2). On the other hand, the peptides detected for proteins of the technical replicates for the same samples (Figure A3A) and biological replicates for the same sampling time points (Figure A3B)

were highly repeatable, so we also pooled the results from the technical and biological replicates for further analysis. We detected from 1,270 (in M-C6h) to 1,951 (in LB1.0) proteins in each sample (Figure A3C). In total, we identified 2,436 proteins in the 20 samples, each of which was present in 4.47±1.43 samples. Thus, the number of proteins detected is about half the number of mRNA detected, indicating that the RNA-seq method was more sensitive than the mass spectrometric method for detecting the respective type of molecules.

3.2.2 Properties of asRNA

As summarized in Table A3, depending on the sampling time points/growth phases and culture conditions, a varying small percentage (0.15%–2.01%) of nucleotides of the reads in the samples were uniquely mapped to the antisense strand of coding regions, yet the number of genes that had at least one mapped antisense read was relatively high and also highly varying, ranging from 804 (17.6%) to 4,270 (93.5%). In total, we assembled 6,613 antisense transcripts that passed a minimum expression level. These results are consistent with the earlier findings (Selinger et al. 2000). However, as shown in Figure 3.1A, 4,984 (75.4%) of the assembled antisense transcripts appeared in only one or two samples and were likely to be noise transcripts. To at least partially eliminate possible noise antisense transcripts, we consider the rest 1,629 of the assembled antisense transcripts that occurred in at least three samples as asRNAs (dataset DS3A in the Appendix B) (see Methods).

Although a few of these 1,629 predicted asRNAs were several thousand nucleotides long, vast majority (92%) of them were shorter than 1kb with a median length of 439nt (Figure 3.1B). Most (~70%) of the predicted asRNAs overlapped with a single gene, while a few overlapped with multiple genes (Figure 3.1C). As a result, the 1,629 predicted asRNAs overlap with a total of 1,487 (32%) genes of *E. coli* K12. About 80%

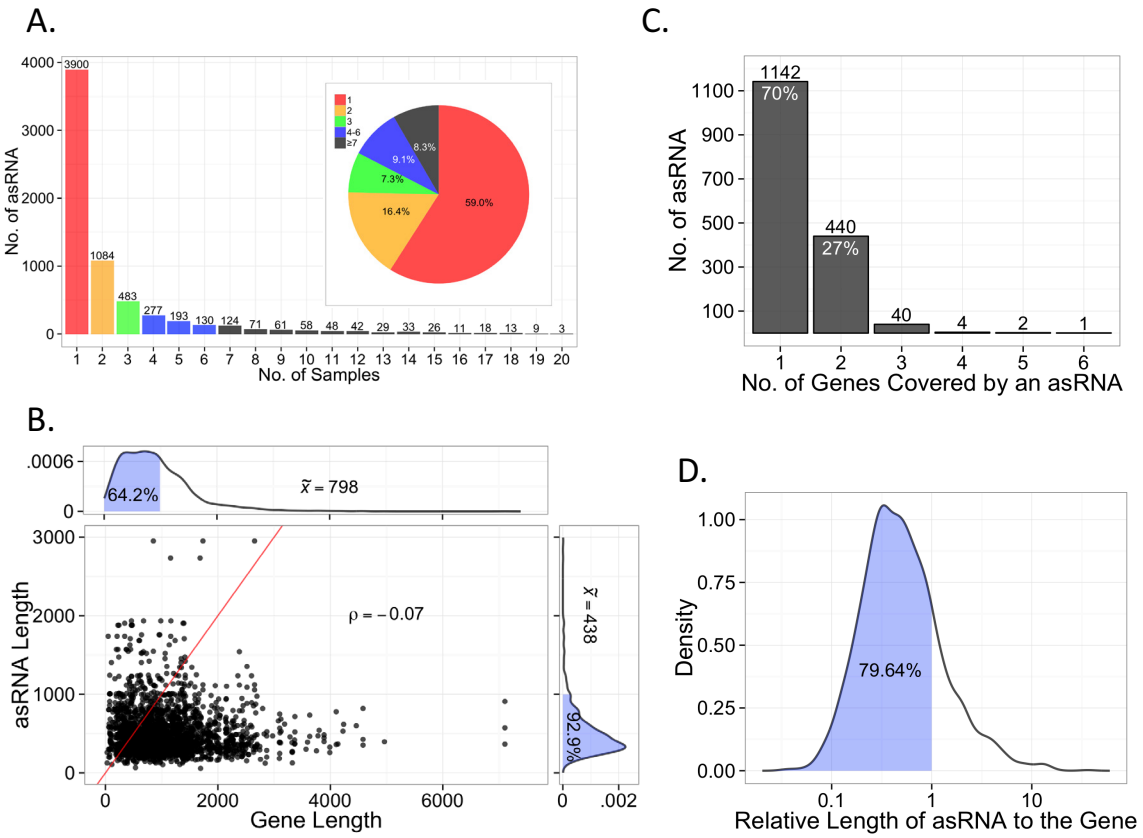


Figure 3.1: Properties of the predicted asRNAs.
(A) Number of samples in which the same antisense TSS were observed. (B) The lengths of asRNAs and of the cognate genes are not correlated. ρ is the Spearman correlation coefficient with a p-value of 0.001. The red line represents the identity line. Marginal distributions of the lengths of asRNAs and cognate genes are shown along the respective axes. The median lengths of asRNA and cognate genes are 438nt and 798nt, respectively and 64.9% of the genes in *E. coli* K12 are shorter than 1,000nt while more than 92% of the asRNAs are shorter than 1,000nt. (C) Number of asRNAs spanning different numbers of genes. (D) Relative length of asRNAs to the cognate genes.

of the predicted asRNAs were shorter than the cognate genes (Figure 3.1D), and the lengths of asRNA were independent of the lengths of the cognate genes (Figure 3.1B). Moreover, while most (74%) of these genes had only one asRNA initiated in their bodies or upstream, we detected multiple asRNA transcription start sites (TSS) for some genes (Figure 3.2A). Most (83.5%) asRNA TSSs were located inside the gene body, and more preferentially at the two ends of the genes, although some asRNAs were initiated from the 3' downstream regions of the genes (Figures 1.5B and 1.5C). Inherently, the proportion of the asRNAs that were initiated inside the gene body increased from 60%

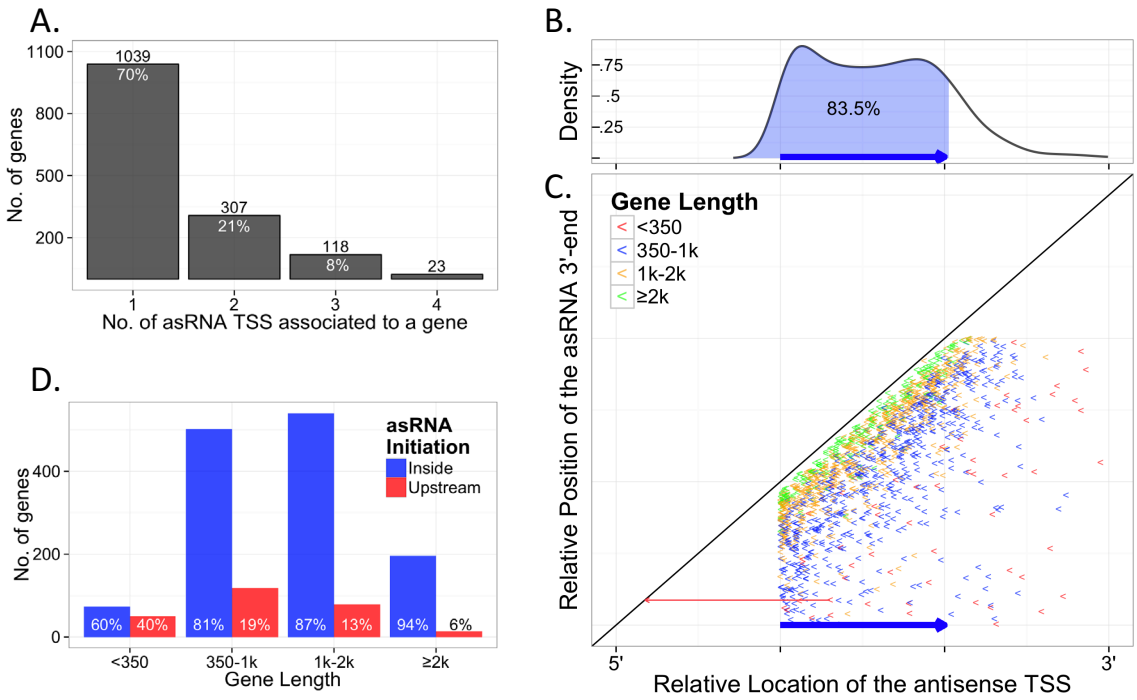


Figure 3.2: Properties of predicted asRNA TSSs.
(A) Number of genes having different numbers of associated asRNAs initiated in the coding or 3'-UTR region of the genes. (B) Distribution of asRNA TSS locations relative to the gene body. (C) Relative asRNA TSS locations to the gene body. Each < represents an antisense TSS. The relative length of an asRNA extends from the TSS to the diagonal line, and the red arrow shows an example of asRNA. In (B) and (C) the x-axis is the relative coordinates of nucleotides from the 5'-end to the 3'-end, and the blue arrows represent the coding region of the genes. (D) Number of the genes having antisense TSSs located in the inside or upstream of the genes with different lengths.

for genes shorter than 350bp to 94% for genes longer than 2kbp (Figure 3.2D), i.e., the longer the gene, the more likely its asRNA was initiated inside the gene body.

3.2.3 Most asRNAs Are Expressed in a Culture Condition-Dependent Manner

We next compared the asRNAs observed under different culture conditions. As shown in Figure 3.3A, although 202 (12%) of the predicted asRNA were transcribed under all the culture conditions, the majority of asRNAs were only seen in certain specific culture conditions, thus their expression was more or less culture condition-dependent. Specifically, 204 (13%) of the asRNAs were only transcribed in one culture condition. Interestingly, most of these highly specifically expressed asRNAs were occurred in HS and M-P conditions (Figure 3.3B). The other 1,223 (85.8%) were utilized in varying combinations of culture conditions (Figure 3.3B). For example, the two predicted asRNAs for the *mcrB* gene were observed only in HS, while the predicted asRNA for *symE*, an SOS-induced gene, encoded in the opposite strand was seen in LB, MOPS and HS conditions but not in M-P and M-C (Figure A4). It has been shown that

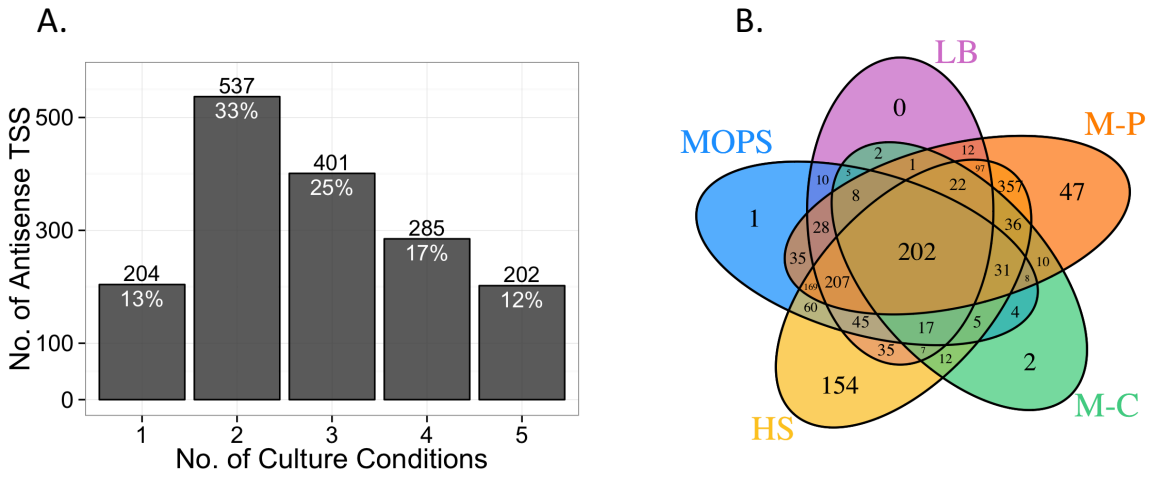


Figure 3.3: Condition dependency of antisense TSS.
(A) Number of asRNAs observed in different number of culture conditions. (B) Sharing of antisense TSSs observed in the five culture conditions.

this asRNA was expressed under LB and MOPS growth conditions (Kawano et al. 2005), and that it repressed the transcription of *symE* (Kawano, Aravind, and Storz 2007). Some predicted asRNA TSSs are more active in certain conditions than in other conditions. For example, one asRNA (t_2) for the *sulA* gene was more transcribed in HS than in other conditions, while the other asRNA (t_1) for the gene was more active in the other conditions (Figure A5). Taken together, these results indicate that the expression of most asRNAs were highly condition dependent, thus they may play a role in the adaptation of the bacterium to the respective culture conditions.

3.2.4 Our Predicted Antisense RNAs Largely Overlap with Those from Earlier Studies

As we indicated earlier, highly varying numbers of asRNAs have been reported in *E. coli* K12 by different research groups, ranging from 90 to 5,495 (Dornenburg et al. 2010; Lybecker et al. 2014; Raghavan et al. 2012; Salgado et al. 2013; Thomason et al. 2015). Although there are still intense debates about the authenticity of these asRNAs (Georg and Hess 2011), true asRNAs are more likely to be detected by multiple research groups using similar or different methods and culture conditions. Thus, to further validate our predicted asRNAs, we compared them with those from these earlier studies using two

Table 3.1: Comparison of our predicted asRNA with those reported in earlier studies.

		No. of asRNA	Matching TSS		Genes with asRNA	Matching genes	
Thomason	**2015**	5,495	709	43%	3,916	1,357	91%
Lybecker	**2014**	316	26	8%	261	110	42%
Salgado (RegDB 8.6)	**2014**	121	17	14%	102	42	41%
Raghavan	**2012**	90	45	50%	86	61	71%
Dornenburg	**2010**	1,005	118	12%	704	363	52%

metrics. First, we compared the TSSs of our predicted asRNAs with those of asRNAs reported earlier by different research groups. If the distance between our predicted antisense TSS and an earlier reported one is within a specific cutoff, we considered them to be the same TSS. Second, we compared the genes that we predicted to have asRNAs to those for which an earlier study also reports to have asRNAs. As shown in FigureA6, increasing the distance cutoff in the first metric had a minimal effect on the recovery rate when the cutoff was greater than 3nt, thus we chose 3nt as the cutoff for the validation. With this cutoff, our predicted asRNAs recovers 8%–50% of those reported by Dornenburg *et al.* (Dornenburg et al. 2010), Raghavan *et al.* (Raghavan et al. 2012), Salgado *et al.* (Salgado et al. 2013), and Lybecker *et al.* (Lybecker et al. 2014), who identified a smaller number of asRNAs than we did (Table 3.1). Furthermore, 43% of our predicted asRNAs match those reported by Thomason and colleagues (Thomason et al. 2015), who identified far more asRNAs than we did (Table 3.1). By the second metric, 41%, 42%, 52%, 71% and 91% of the genes that we predicted to have asRNAs also have asRNAs reported by Salgado *et al.* (Salgado et al. 2013), Lybecker *et al.* (Lybecker et al. 2014), Dornenburg *et al.* (Dornenburg et al. 2010), Raghavan *et al.* (Raghavan et al. 2012), and Thomason *et al.* (Thomason et al. 2015), respectively. Thus, although Thomason and colleagues have noted that there were very limited (4%–33%) overlaps between the previously reported asRNAs (Thomason et al. 2015), our predicted asRNAs had a rather high matching rate with these earlier reports. As an example, Figure A7 shows that our three predicted antisense TSSs for the *intS* and *yfdGHI* operons match with those reported by Thomason *et al.* (Thomason et al. 2015) and Salgado *et al.* (Gama-Castro et al. 2011). However, note that Thomason *et al.* (Thomason et al. 2015) identified

a total of eight TSSs, and the other five were not seen in our samples or the other studies. On the other hand, as shown in Figure A7, although we predicted asRNAs for the *sulA* gene may overlay with those reported by Thomason *et al.* (Thomason et al. 2015) and Dornenburg *et al.* (Dornenburg et al. 2010), they had quite different TSSs. These results indicate that our predicted asRNAs are more likely to be genuine, and the number of these putative asRNAs could be a good estimate of asRNAs in the *E. coli* K12 genome.

3.2.5 Transcriptional Modes Defined by Relative Levels of Antisense and Sense Transcription Are Dependent on Culture Conditions

To investigate the possible effect of an asRNA on the expression of its cognate gene, we analyzed the relationship between mRNA and asRNA levels for genes. To avoid the effects of possible overlapping transcription between divergent or convergent genes on measuring antisense transcription levels, we only considered for this analysis the subset of adjacent genes with the same orientation (2,330 genes) in the *E. coli* K12 genome (dataset DS3B in Appendix B). We found that sense and antisense transcription levels were independent in all the samples

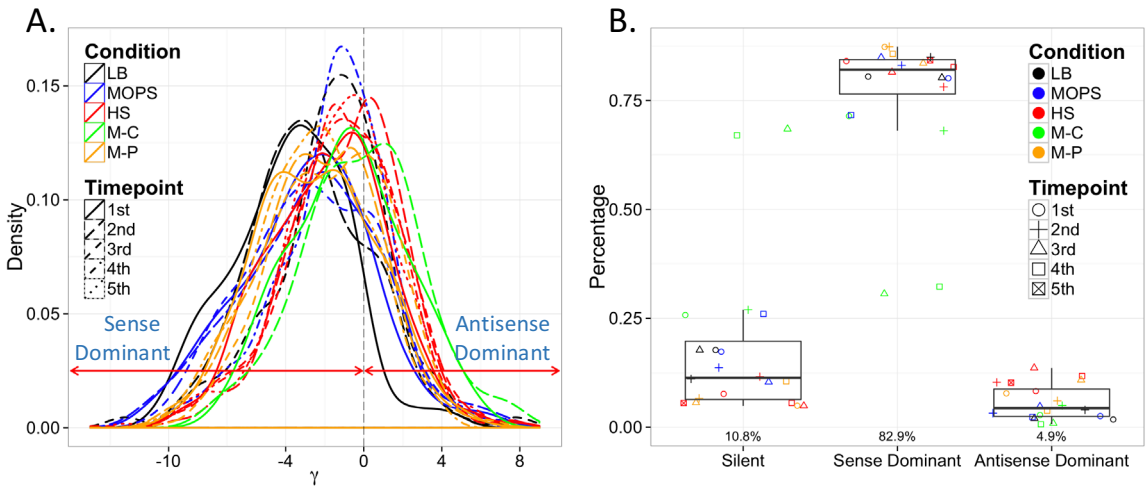


Figure 3.4: Transcriptional modes of genes.
(A) Distribution of log odds ratio of antisense and sense transcription levels (γ) in the samples. (B) Percentage of the genes in each transcriptional mode in the samples; the medians are shown along the *x* axis.

Table 3.2: Summary of the transcriptional modes of genes in the 20 samples.

Sample	Sense Dominant		Antisense Dominant		Silent		Gene with asRNA	
LB 0.5	1876	80.5%	41	1.8%	413	17.7%	157	6.7%
LB 1.0	1980	85.0%	92	3.9%	258	11.1%	326	14.0%
LB 3.0	1870	80.3%	47	2.0%	413	17.7%	118	5.1%
MOPS1h	1867	80.1%	59	2.5%	404	17.3%	180	7.7%
MOPS2h	1936	83.1%	75	3.2%	319	13.7%	216	9.3%
MOPS4h	1978	84.9%	111	4.8%	241	10.3%	330	14.2%
MOPS6h	1670	71.7%	54	2.3%	606	26.0%	117	5.0%
HS15min	1959	84.1%	193	8.3%	178	7.6%	568	24.4%
HS30min	1820	78.1%	240	10.3%	270	11.6%	506	21.7%
HS1h	1899	81.5%	317	13.6%	114	4.9%	805	34.5%
HS2h	1927	82.7%	274	11.8%	129	5.5%	698	30.0%
HS4h	1963	84.2%	238	10.2%	129	5.5%	667	28.6%
M-C1h	1665	71.5%	65	2.8%	600	25.8%	114	4.9%
M-C2h	1586	68.1%	116	5.0%	628	27.0%	179	7.7%
M-C4h	714	30.6%	21	0.9%	1,595	68.5%	27	1.2%
M-C6h	752	32.3%	16	0.7%	1,562	67.0%	19	0.8%
M-P1h	2034	87.3%	181	7.8%	115	4.9%	711	30.5%
M-P2h	2035	87.3%	141	6.1%	154	6.6%	524	22.5%
M-P4h	1947	83.6%	252	10.8%	131	5.6%	712	30.6%
M-P6h	1997	85.7%	88	3.8%	245	10.5%	309	13.3%

(Figure A8). Since an asRNA is likely to execute its functions by forming complementary duplexes with its sense transcript (Brantl 2007; Georg and Hess 2011; Lasa et al. 2012; Thomason and Storz 2010; Wade and Grainger 2014), we computed the logarithmic ratio of asRNA and mRNA levels for each gene in the subset, γ, which measures the relative levels of antisense and sense transcription of the gene. Interestingly, as shown in Figure 3.4A, γ had a left-skewed and bell-shaped distribution in all the samples. This prompted us to divide the transcriptional activities of a gene into three possible distinct *transcriptional modes* according to the extent to which the mRNA and asRNA levels differ: sense-dominant mode ($\gamma \leq 0$ or asRNA$=0$), in which the gene has higher sense

transcription than antisense transcription; antisense-dominant mode (γ>0 or mRNA=0), in which the gene has higher antisense transcription than sense transcription (Figure 3.4A); and silent, if the gene has neither the sense nor antisense transcription. As shown in Figure 3.4B and Table 3.2, there were more genes in the sense-dominant mode than in the antisense-dominant or silent modes in the samples with except for the samples taken at the prolonged carbon starvation phases (M-C4h and M-C6h, in which there were more genes in the silent mode. However, the proportions of genes in these modes changed dramatically at different growth phases/time points and culture conditions (Table 3.2). Interestingly, the number of genes in the antisense-dominant mode generally increased after the onsets of all the cultures, and then decreased at the end of the cultures after peaking at the second or the third sampling time-points (Table 3.2). These consistent patterns of dynamic antisense transcription activities suggest again that the resulting asRNAs may play a role in the bacterium's adaptation of to environmental changes.

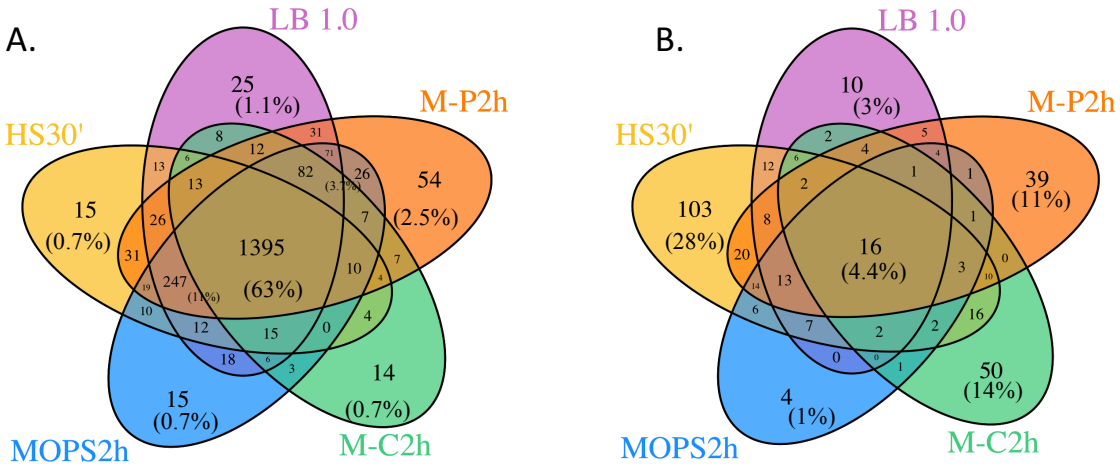


Figure 3.5: Number and percentage of genes in different transcriptional modes.
(A) Sense dominant mode and (B) antisense dominant mode shared by the second time-point of the five culture conditions.

In order to understand the effects of culture conditions on the transcriptional modes that genes adopted, we compared genes' transcriptional modes at the second time-point of the five culture conditions (i.e. LB 1.0, HS30', MOPS2h, M-C2h and M-P2h) when the cells were presumably in the steady states. As shown in Figure 3.5, 63% of genes in sense-dominant mode were shared by all the five samples, while only 4.4% of genes in the antisense-dominant mode were shared, indicating that genes in this mode were more likely to be dependent to culture conditions than those in the sense-dominant mode for adapting the respective transcriptional modes. GO term analyses on the shared and unique sense-dominant and antisense-dominant genes in these conditions (dataset DS3C in Appendix B) showed that housekeeping GO terms such as cell membrane, transport, ribosomal terms were enriched in the sense-dominant genes shared among the five culture conditions, whereas specific functional GO terms were enriched for genes that were uniquely in antisense-dominant or sense-dominant modes in specific conditions. For instance, sugar transport (GO:0051119, GO:0008643) and ATP binding (GO:0005524) were enriched in unique antisense-dominant genes in carbon starvation (M-C2h); metal binding (GO:0051536, GO:0051539, GO:0005506) and oxidoreductase (GO:0016491) were enriched in unique antisense-dominant genes in heat shock treatment (HS30'); phosphorous metabolic process (GO:0006793), ion transport (GO:0006811, GO:0006820, GO:0015711) and organic phosphonate transport (GO:0015716) were enriched in unique sense-dominant genes in phosphorous starvation (M-P2h). These results again strongly suggest that asRNA may play an important role in the bacterium's adaptation to different environments.

3.2.6 Genes Change Their Transcriptional Modes at Different Growth Phases/Time Points in a Culture Condition

To determine whether and how a gene changed its transcriptional mode at different growth phases/time points in a culture condition, we counted the number of times that genes change their transcriptional modes between two adjacent sampling times in the five culture conditions (Figure A9). As shown in 3.6A, most genes (73%–81%) stayed in the same transcriptional mode at the sampling time points in all the cultures except M-C, especially those in the sense-dominant mode. However, many genes

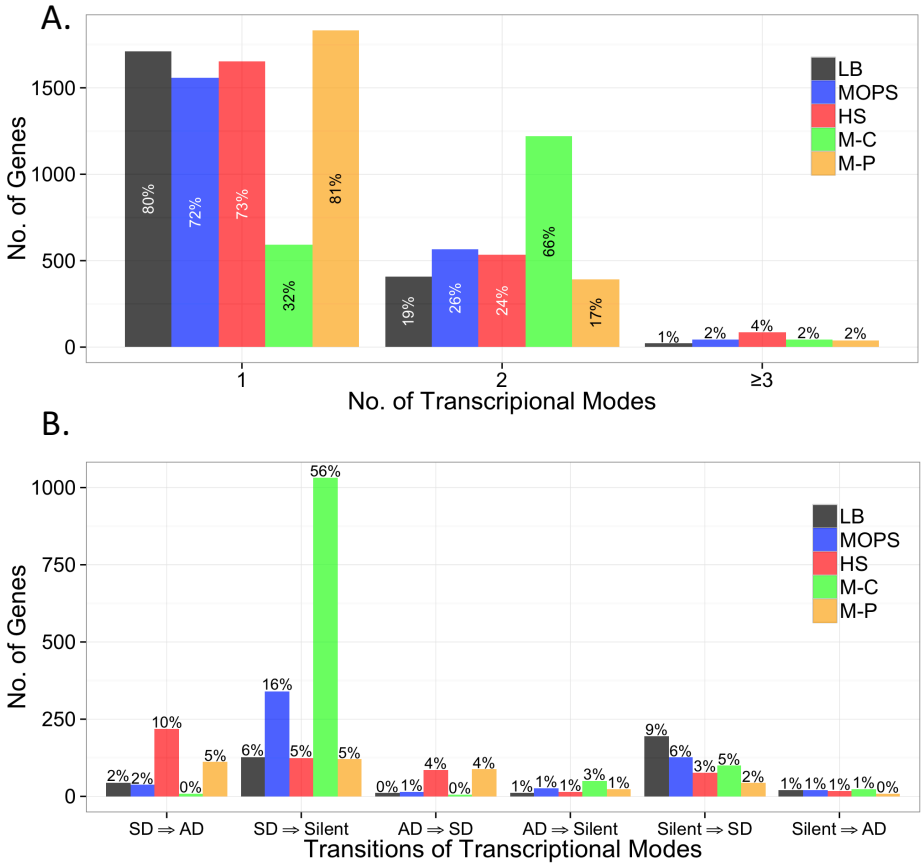


Figure 3.6: Transcriptional mode change abundance. **A.** Number and percentage of genes undergoing different numbers of transcriptional modes in each growth condition. **B.** Number of genes undergoing the single indicated transcriptional mode transitions in each culture condition. The value on each bar represents the percentage of genes with the transition under each condition. SD and AD stand for sense-dominant and antisense-dominant, respectively.

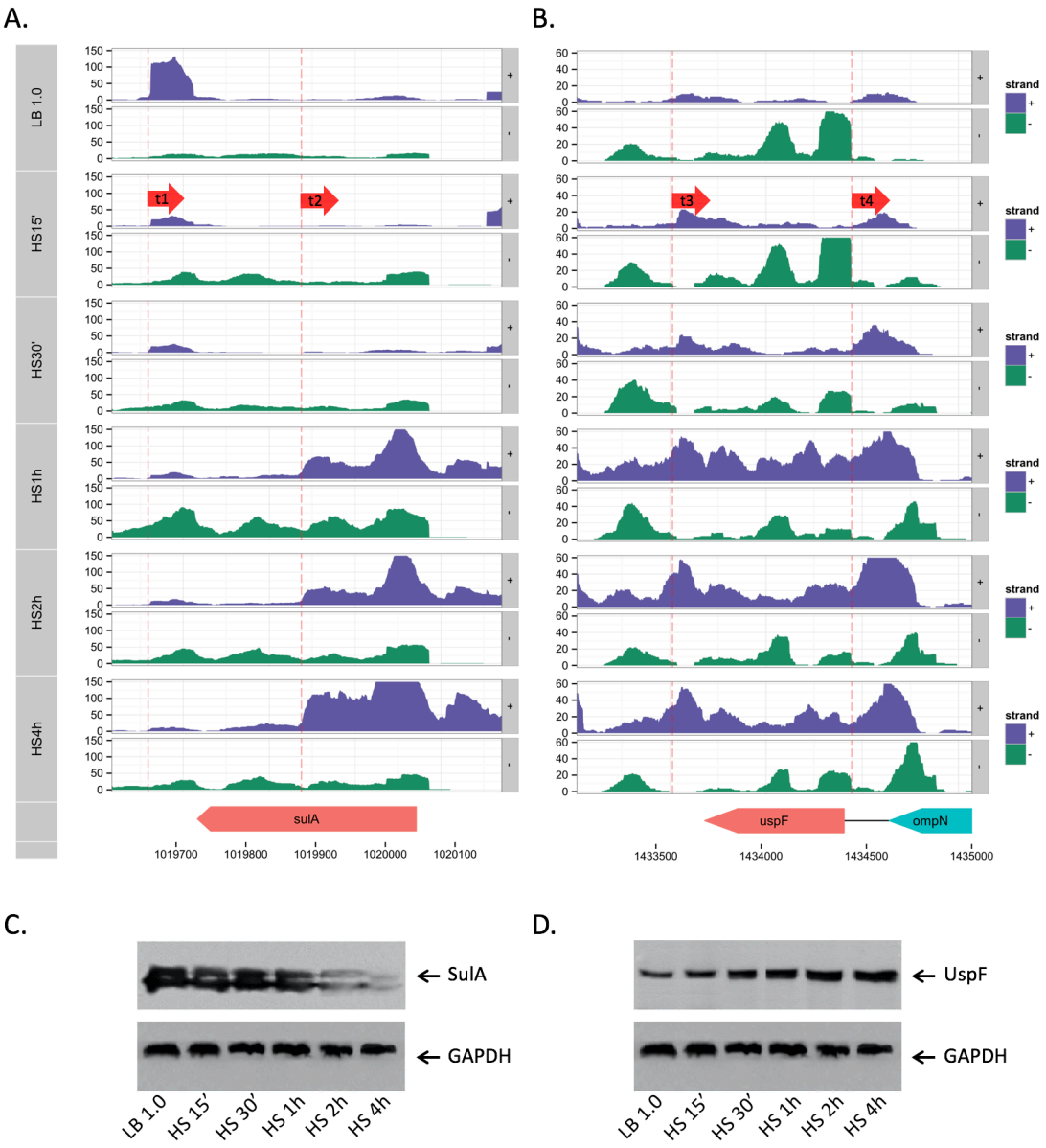


Figure 3.7: Examples of transcriptional mode changes under heat shock stress. (A) The *sulA* gene changed from the antisense-dominant mode to the sense-dominant mode when the cells were transferred from LB to heat shock, and then from the sense-dominant to the antisense-dominant mode during heat shock. Two asRNAs were identified and the dashed vertical lines, marked t_1 and t_2, indicate their predicted TSS. The activity of the t_2 TSS increased in the later stages of heat shock compared with that of t_1. (B) The gene *uspF* going from sense-dominant to antisense-dominant mode. Two identified asRNA TSSs are marked t_3 and t_4. In (A) and (B), the sense and antisense read coverages are colored in green and purple, respectively. (C) Western blot verification of the expression of the SulA protein under heat shock stress condition. (D) Western blot verification of the expression of the UspF protein under heat shock stress condition. In (C) and (D) GAPDH served as the standard control.

changed their transcriptional modes at different stages of the cultures, and some even changed their transcriptional modes multiple times during the sampling process (Table A4). Intriguingly, as summarized in Figure 3.6B, genes showed distinct patterns of transcriptional mode changes under different culture conditions. Specifically, the transition from the sense-dominant mode to the silent mode was most predominant under the carbon starvation and MOPS culture conditions, while the transition from the sense-dominant mode to the antisense-dominant mode was most predominant under heat shock, and the transition from the silent mode to the sense-dominant mode was most predominant under the LB culture. To investigate the functional implication for such transcriptional mode transitions, we performed GO term enrichment analysis on genes that changed their transcriptional modes once under heat shock and phosphorous transport (GO:0006865) terms were enriched for the genes that transition from the antisense-dominant mode to the sense-dominant mode; and under phosphorous starvation, phosphate-binding loop (p-loop) motif containing genes were enriched for the genes with the same form of transitions. As an example, Figure 3.7 show how the *sulA* gene encoding the cell division inhibitor (George, Castellazzi, and Buttin 1975; Al Mamun et al. 2012) and the *uspF* gene encoding the universal stress response protein, changed their transcriptional modes at different stages of the heat shock culture. These results suggest that changes in the transcriptional modes

of genes may play a role during the growth and adaptation of the bacterium under these culture conditions.

3.2.7 Protein Levels of Genes Are Stoichiometrically Affected by the Associated asRNAs

To see the possible effects of asRNA transcription on the protein expression of the associated genes, we quantified the levels of detected proteins in terms of the number of peptides per hundred amino acids (NPPH). As shown in Figure A10A&B, genes generally had similar distribution of protein levels in all the 20 samples, nonetheless, genes in different transcriptional modes had quite different protein levels. Specifically,

although protein levels of genes in the silent and antisense-dominant modes had similar distributions (adjusted p-value of 0.46), they were significantly differently different from that of the protein levels of genes in the sense-dominant mode (Figure A10C&D). We analyzed the relationship between protein levels and mRNA levels as well as asRNA levels of the genes in each sample. As shown in Table A5,the protein and mRNA levels in all the samples were strongly correlated with a spearman correlation coefficient (ρ) ranging from ρ=0.23 in M-C6h to ρ=0.63 in LB 0.5; and ρ=0.43 when the data were pooled from all the samples (Figure 3.8A). In contrast, the protein levels and asRNA levels were not correlated (ρ=0.033, Figure A11), suggesting that the effects of an asRNA on the expression of its cognate gene if any, are irrelevant to the absolute level of the asRNA in the cells. We next examined the relationships between protein levels and mRNA levels of genes in the sense-dominant and antisense-dominant transcriptional modes. Since some samples do not have enough data points for a reliable analysis (Table A6), we pooled the data from all the samples for this analysis. As expected, the protein levels and mRNA levels for sense-dominant genes were highly correlated (ρ=0.424, p-value=2.2e-16), but this was not the case for antisense-dominant genes (ρ=0.086, p-value=0.16) (Figure 3.8B). Since genes in the antisense-dominant

mode tended to have lower mRNA levels than did those in the sense-dominant mode, to eliminate possible effects of mRNA levels on the correlation between protein levels and mRNA levels for sense-dominant genes, we reexamined the relationship between protein levels and mRNA levels for the subset of sense-dominant genes whose mRNA levels were in the same range as those of antisense-dominant genes. As shown in Figure 3.8B, the correlation between the proteins levels and mRNA levels for this subset of sense-dominant genes was 0.182 with a p-value of 2.2e-16, which was significantly higher than that for antisense-dominant genes. These results indicate that a higher asRNA level relative to the cognate mRNA level somehow disrupted the positive correlation between the protein levels and mRNA levels of genes.

To further verify this conclusion, we monitored the protein expression levels of two select genes *sulA* and *uspF* in the samples using Western blots. As shown in Figure 3.7C, the SulA protein, a cell division inhibitor that plays a role in SOS responses (George et al. 1975; Al Mamun et al. 2012), showed a decrease in the expression in the two later time points under heat shock. While this phenomenon has been noted earlier (Vasil'eva and Makhova 2003), little is known about the mechanism of the expression reduction. As shown in Figure 3.7A, there were minimal changes in the mRNA levels of *sulA* ($23.6 \pm 8.8_{TPM}$) at multiple time-points under heat shock, however, its antisense expression levels dramatically increased 4 hours after the onset of heat shock (from

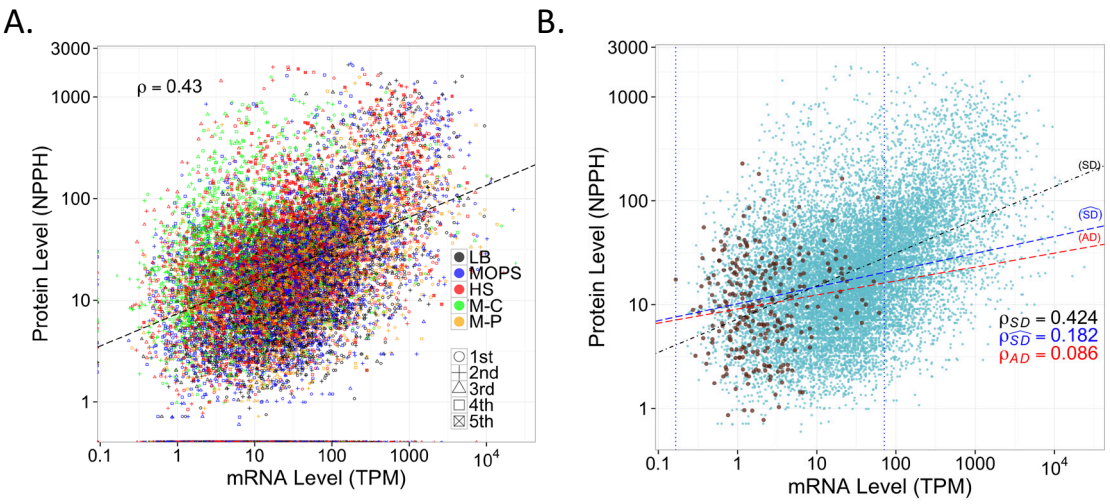


Figure 3.8: Relationship between the sense (mRNA) transcription levels and protein levels of genes.
(A) Correlation for genes when all samples are pulled together. (B) Correlation for genes in different transcriptional modes. Genes in the sense-dominant (SD) transcription modes are shown in blue, while genes in the antisense-dominant (AD) mode are shown in red. The $\widehat{SD}$ genes are a subset of sense-dominant genes whose mRNA levels were in the same range as those of antisense-dominant genes (vertical dotted borders). The Spearman correlation between the protein levels and mRNA levels for these three subsets of genes are shown in black (SD), blue (AD) and red ($\widehat{SD}$). The asymptotic *t* approximation p-values for these correlations are 2.2e-16, 2.2e-16 and 0.1565 respectively.

11.5_{RPK} to 333.7_{RPK}), resulting a transition from the sense-dominant mode at the first two time-points to the antisense-dominant mode at the later time-points. Hence the antisense transcription may play an inhibitory role in the expression of the protein. Furthermore, it has been demonstrated that UspF, a universal stress response protein that promotes adhesion of cells at the expense of motility (Nachin, Nannmark, and Nyström 2005; Saveanu et al. 2002), is upregulated under a glucose limited condition (Raman et al. 2005). As shown in Figure 3.7D, UspF was also upregulated under heat shock. Interestingly, the mRNA level of *uspF* decreased from 23.9_{TPM} in HS15' to 11.60_{TPM} in HS4h, while its antisense level increased from 83.9_{RPK} to 290.1_{RPK}, resulting in a transition from the sense-dominant mode in LB 1.0 and HS15' to the antisense-dominant

mode in the later stages of heat shock (Figure 3.7B). Thus this antisense transcript appeared to enhance the protein expression. On the contrary, under the MOPS culture, both genes showed minimal changes in the mRNA and asRNA levels, and their protein levels did not change significantly (Figure A12).

3.3 Discussions and Conclusions

It has been reported that a highly varying portion (1%–93%) genes in various prokaryotes have antisense RNA (asRNA) transcription. It is not clear whether these differences are due to biological or technical variations or both. In this study, we analyzed the transcriptomes and proteomes in *E. coli* K12 at different growth phases/time points under five culture conditions using a strand-specific RNA-seq method and a quantitative mass spectrometry method. The resulting data allowed us to systematically analyze the pervasiveness and patterns of asRNA transcription during the course of cell growth and adaptation to different environments. In agreement with the early report (Selinger et al. 2000), we found that up to 93.5% of the annotated genes in the genome have reads mapped to their antisense strands (Table A3), however, some of them may be transcriptional noise. To identify authentic asRNAs, we invoked a rather rigorous criterion to call asRNAs in this study, and found that from 0.8% to 34.5% of the annotated genes had asRNA transcription depending on sampling time points and culture conditions (Table 3.2). We predicted a total of 1,629 asRNAs, which is more than the number of asRNAs reported by most earlier studies (Dornenburg et al. 2010; Gama-Castro et al. 2011; Lybecker et al. 2014; Raghavan et al. 2012), but much smaller than the number reported by Thomason and colleagues (Thomason et al. 2015). Thus, some earlier reports may have overestimated the prevalence of antisense transcription in terms

of the number of genes having antisense transcription. Moreover, our results indicate that antisense transcription in *E. coli* K12 is highly variable and dynamic dependent on different time points of growth and adaptation and culture conditions. Thus, the earlier different reports regarding the pervasiveness of asRNA transcription may be due partially to identifying noise antisense transcripts as asRNAs and to different sampling time points and experimental conditions.

Our predicted asRNAs were all repetitively seen in at least three samples, and majority of them were reutilized under different culture conditions, and hence are likely to be authentic asRNAs because it is highly unlikely that the same transcriptional noise can repeat itself exactly in many different conditions. Furthermore, the high overlap of our predicted asRNA with the earlier reported asRNAs further supports the authenticity of our predicted asRNAs. Our findings of the high dependency of the transcription of the predicted asRNAs on the environmental changes, strongly indicate that these asRNAs may have important biological functions.

We classified the transcriptional events of a gene in three possible modes according to its relative asRNA level to the mRNA level. The proportions of genes in these modes were highly variable, depending on the growth phases/time points and culture conditions, and many genes changed their transcriptional modes at the different time-points under specific culture conditions. The transcriptional modes that a gene adopted at different time points and culture conditions can be well explained by the known functions of the gene. All these results suggest that asRNAs may play a crucial role during the bacterium's growth and adaptation to environmental changes.

It has been shown that asRNAs can either down- or up-regulate the expression of the cognate genes (Georg and Hess 2011; Thomason and Storz 2010) that involved in important processes such as DNA replication, stress responses and iron transport (Brantl 2007). Our results support these earlier observations as we found that for some genes such as *sulA* (Figure 3.7A), relatively elevated antisense transcription was accompanied with decreased expression of the protein, while for some other genes such as *uspF*, relatively elevated antisense transcription was concomitant with increased expression of the protein (Figure 3.7B). Moreover, our finding that the correlation between protein levels and mRNA levels for genes in the antisense-dominant mode disappeared suggests that asRNAs may participate in gene expressional regulation. Several mechanisms have been proposed to explain how asRNA can affect gene expression, including transcriptional interference, alteration of mRNA stability, and modulation of translation (Brantl 2007; Georg and Hess 2011; Lasa et al. 2012; Lavorgna et al. 2004; Thomason and Storz 2010). Although the detailed molecular mechanisms remain to be elucidated, our results are in excellent agreement with the threshold linear response model for the stoichiometric interaction between asRNAs and cognate mRNAs (Legewie et al. 2008; Levine et al. 2007). According to this model, the formation of a duplex between an mRNA and its cognate asRNA will either decrease or increase the transcription, translation or stability of the mRNA, therefore an increase in the relative level of an asRNA disrupts otherwise strong correlation between the protein and mRNA levels. This model also is consistent with our finding that the lengths of asRNAs are irrelevant to their functions, because only a short asRNA is need to exert its function regardless of the length of the gene.

3.4 Methods

3.4.1 Bacteria Culture and Sample Collections

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°C and 250 rpm. The cells were then transferred to fresh LB medium in a flask by 1:100 dilutions, and cultured at 37°C and 250 rpm. When the cells grew to an optical density at 600 nm (OD_{600}) of 1.0, they were spun down at 3,200g for 25 min. For the minimal medium MOPS culture (MOPS), the cell pellets were resuspended in the same volume of MOPS medium (100ml 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°C and 250 rpm. For phosphorus-starvation treatment (M-P), the cell pellets were resuspended in the MOPS medium without KH_2PO_4. For carbon-starvation treatment (M-C), the cell pellets were resuspended in the MOPS medium without glucose. Each culture was done in three replicates in parallel, and an equal volume of cells was collected from each replicate to minimize experimental variations during sampling. Three milliliters of such pooled cell suspension were collected in a test 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 and 4hrs; MOPS, M-C and M-P: 1hr, 2hrs, 4hrs, 6hrs; M-C: 1hr, 2hrs, 4hrs, 6hrs). For the LB culture, samples were collected when the cell suspension OD_{600} reached to 0.5, 1.0 and 3. Cells were spun down at 6,000g for 8 mins (-4°C), and the pellets were resuspended in 1.5 ml of RNAlater. The samples were stored at -80°C until use.

For proteome and immunoblotting analysis, the collected cells were washed twice in the same volume of 0.9% NaCl by centrifugation at 6,000g for 8 min (-4°C), and the pellet was suspended in 1 ml of protein extraction buffer (100mM TrisCl, 500mM NaCl, 1mM sodium EDTA, 2mM DTT and 0.1% TX 100, protease inhibitor cocktail (Roch, 1 tablet in 50ml) PH 7.4). Cells were disrupted by sonication in an ice bath at 4% power, 10% pulse for 10 min, followed by centrifugation at 12,000g for 30min. The supernatant was quantified for protein levels using the Bradford method and stored at -80°C until use. Most cultures were done at least twice (two biological replicates).

3.4.2 Isolation and Enrichment of mRNA

Total RNA was isolated from the cells using RiboPure^{TM} -Bacteria Kit (Ambion) following the manufacturer's instructions. Once isolated, ~10μg 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 bp fragment of the *crp* gene (5'-primer:AGCATATTTCGGCAATCCAG; 3'-primer:TACAGCGTTTCCGCTTTTTC). rRNAs were depleted from the total RNA using a MICROBExpress kit (Ambion) to enrich mRNAs.

3.4.3 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 using Illumina Small RNA Sample Prep Kit following the vendor's instruction 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 10mM 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 the inserted RNA fragments. 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. 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 (Invitrogen) 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. Libraries for some samples (M-C1h and M-C2h) were prepared by the two methods and sequenced on the two different platforms.

3.4.4 Reads Preprocessing, Mapping and Transcript Assembling

The genome sequence of *E. coli* K12 substr. MG1655 (NC_000913.2) was downloaded from NCBI. The gene annotation file and the experimentally verified operons in the bacterium were downloaded from RegulonDB (version 8.6) (Salgado et al. 2013) (http://regulondb.ccg.unam.mx/). A total of 4,567 annotated genes (also including pseudo genes) are included in this analysis. As the RNA-seq reads were not size-selected during the library construction to capture short RNAs, we trimmed the 3' adapters attached to some short insertions using Trimmomatic (Bolger, Lohse, and Usadel 2014) with standard parameters (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:35). The trimmed reads were mapped to the *E. coli* K12 genome using Bowtie 2.0 (Langmead and Salzberg 2012) with --very-sensitive parameter. Only uniquely mapped reads were used for further analysis.

We predicted full length asRNAs in two steps. First, for each sample we stitched overlapping antisense reads for a tandem string of ORFs to form the longest antisense transcripts in the sample. Second, if the stitched antisense transcripts from a minimum number (3) of samples have TSSs within a 10bp window and a minimum expression level threshold (5 raw read counts) in all the samples, we stitched all the overlapping antisense transcripts to from a predicted asRNA. Raw expression levels were calculated using the Bioconductor package (Lawrence et al. 2013) in R (R Core Team 2015). Number of reads mapped to the sense strand of a gene was normalized in TPM (total reads per million) (Li et al. 2010), and number of reads mapped to the antisense strand was normalized by the length of assembled antisense transcript. The ratio of antisense to sense for each gene, γ, is calculated as the $\log_2$(normalized asRNA level/normalized mRNA level). In addition,

we used DAVID 6.7 (Huang, Sherman, and Lempicki 2009a, 2009b) to analyze functional enrichment for groups of genes.

3.4.5 UPLC and Tandem Mass Spectrometry Analysis

Fifty μg of total protein from each sample were separated on 10% Bis-Tris NuPAGE gels (Invitrogen, Carlsbad, CA, USA) with 6X sample buffer containing 300mM Tris-HCl, 0.01% (w/v) bromophenol blue, 15% (v/v) glycerol, 6% (w/v) SDS and 1% (v/v) β-mercaptoethanol after denaturation at 95°C for 5 minutes. Gels were stained with the GelCode® Blue stain reagent (Thermo Scientific, Rockford, IL, USA) after fixation using 50% methanol (v/v) with 7% acetic acid (v/v) for 5 min. After destaining with water, each gel lane was excised into twenty slices which were put into in-gel tryptic digestion and peptide extraction according to the method reported previously (Lee et al. 2013). The dried residues were resuspended in 25μL of 10% ACN (v/v) with 3% formic acid (v/v) for LC-MS/MS analysis.

The LC-MS/MS system used consisted of an LTQ/Orbitrap-XL mass spectrometer (Thermo Scientific, Rockford, IL, USA) equipped with Nanoacquity UPLC system (Waters, Milford, MA, USA). Peptides were separated on a reversed phase analytical column (Nanoacquity BEH C_{18}, 1.7μm, 150mm, Waters, Milford, MA, USA) combined with trap column (Nanoacquity, Waters, Milford, MA, USA). Good chromatographic separation was observed with an 80 min linear gradient consisting of mobile phases solvent A (0.1% formic acid in water) and solvent B (0.1% formic acid in ACN) where the gradient was from 5% B at 0 min to 40% B at 65 min at 0.35μL/min of flow rate. MS spectra were acquired by data dependent scans consisting of MS/MS scans

of the eight most intense ions from the full MS scan with dynamic exclusion of 30 seconds.

3.4.6 Protein Database Search and Data Compiling

The *Escherichia coli* str. K12 substr. MG1655 proteome file was downloaded from NCBI, and was used as the database to identify proteins using the SEQUEST algorithm (SRF v.5) in the Bioworks software v.3.3.1sp1. Search parameters were as follows: parent mass tolerance of 10 ppm, fragment mass tolerance of 0.5Da (monoisotopic), variable modification on methionine of 16 Da (oxidation) and maximum missed cleavage of two sites assuming the digestion enzyme trypsin. Search results were compiled using the Scaffold software (v3.6.3, Proteome Software, Portland, OR, USA) which provided spectral counts for data comparison under the following filter criteria: protein identifications were made at 95% peptide probability and 99% protein probability with at least two identified peptides. Shared and partial-tryptic peptides were excluded from spectral counts. Protein probability and redundancy were assigned by the Protein Prophet algorithm. Proteins that contained similar peptides, which could not be differentiated based on MS/MS spectra, were grouped into primarily assigned proteins. Spectral counts from duplicated analyses were compared using the Power Law Global Error Model (PLGEM) in order to identify the significance of the protein changes (Pavelka et al. 2004). Proteins are quantified by the number of peptides per hundred amino acids identified for the protein: NPPH=100n/L, where n is the number of identified peptides of the protein, and L the length of the protein.

3.4.7 Immunoblotting Analysis

Twenty μg protein from each sample were separated by 8% sodium dodecyl sulfate (SDS)-polyacrylamide gel (SDS-PAGE) and transferred to nitrocellulose membranes. The membrane was blocked in 7% fat-free dry milk in TBS containing 0.2% Tween-20, and probed with antibodies against interest proteins. The antibodies used in this study included rabbit anti-*Escherichia coli* antibodies for UspF (Mybiosource, San Diego, CA, USA), SulA (Mybiosource, San Diego, CA, USA) and GAPDH (abcam, Cambridge, MA, USA). The GAPDH protein is known to be stably expressed under different conditions (Wu et al. 2012) and was used as the control for the western blot experiments. These primary antibodies were applied to the membranes in a dilution of 1:2000. Following extensively washing, the membranes were incubated with a HRP-labeled secondary antibody. The blots were then developed using the ECL western blotting substrate (Thermo scientific, Rockford, IL, USA). The levels of protein expression were semi-quantified by optical densitometry using Image J Software version 1.46. The ratio between the net intensity of each sample to that of the GAPDH internal control was calculated and served as an index of relative expression of an interest protein.

CHAPTER 4: CONCLUSIONS

In this dissertation, we aimed to use the genomics, transcriptomics and proteomics data to gain a better understanding of genome-wide gene regulation, transcription and annotation in prokaryotes and specifically in *E. coli* K12. Our results presented in Chapters 1~3 indicate that we have largely achieved the goals.

In chapter 1, we presented PorthoMCL, a fast tool for finding orthologous genes among a very large number of genomes. PorthoMCL can be run on a single machine or in parallel on computer clusters. We have demonstrated PorthoMCL's capability not only by showing it's faster than current orthology finding tools, but also by identifying orthologs in 2,758 prokaryotic genomes. PorthoMCL will facilitate comparative genomics analysis with increasing number of available genomes thanks to the rapidly evolving sequencing technologies.

In Chapter 2, we analyzed alternative operon utilizations in *E. coli* K12 at multiple time-points in three different stress conditions based on TUs assembled using RockHopper. In addition, we analyzed dynamic transcription of TUs in the samples by adopting a model that has been successfully used in modeling the expression levels alternate splicing isoforms in eukaryotes. We found that this model can accurately reflect the extent of dynamic transcriptions. Our results show that 22% of operons have alternative TU transcriptions, and up to 36% of TUs display dynamic transcription in response to environmental changes for adaptation.

In Chapter 3, we determined the transcriptomes and proteomes of *E. coli* K12 at multiple time points in five culture conditions using strand-specific RNA-seq techniques and a quantitative mass spectrometry method and identified a total of 1,629 asRNAs, which were generally short, largely condition dependent, and overlapped with the previously published asRNAs. We found that the proportions of the genes which had asRNAs were highly variable (0.8%–34.6%) based on the time points and culture conditions. We classified the transcriptional activities of the genes in three distinct transcriptional modes according to their relative levels of asRNA to mRNA: sense-dominant, antisense-dominant and silent modes. We found that many (19%–27%) of genes changed their transcriptional modes at different time points of the culture conditions, and that such transitions can be described in a well-defined manner. Intriguingly, the protein levels and mRNA levels of genes in the sense-dominant mode were strongly correlated, but the same was not true for genes the antisense-dominant mode. These observations were further validated by western blot analyses on candidate genes, where an increase in asRNA transcription diminished gene expression in one case and enhanced it in another. These results suggest that asRNAs may directly or indirectly regulate translation by forming duplexes with cognate mRNAs. Thus, asRNAs may play an important role in the bacterium's responses to environmental changes during growth and adaption to different environments.

To summarize, we developed a fast and scalable tool predicting orthologs in a large number of prokaryotic genomes. Additionally, we proposed to use a eukaryotic alternative splicing isoforms model to investigate alternative and dynamic operon transcription in *E. Coli* K12. This model yields a better understanding of the patterns of

transcriptional changes in bacteria in response to environmental changes. Furthermore, we elucidated yet another aspect of complex gene regulation in *E. coli* K12, antisense transcription, and explored the effects and roles of asRNAs in bacteria's response to environmental changes.

REFERENCES

Albrecht, Marco, Cynthia M. Sharma, Richard Reinhardt, Jörg Vogel, and Thomas Rudel. 2010. "Deep Sequencing-Based Discovery of the Chlamydia Trachomatis Transcriptome." *Nucleic acids research* 38(3):868–77.

Alexeyenko, Andrey, Julia Lindberg, Asa Pérez-Bercoff, and Erik L. L. Sonnhammer. 2006. "Overview and Comparison of Ortholog Databases." *Drug discovery today. Technologies* 3(2):137–43.

ALTSCHUL, S., W. GISH, W. MILLER, E. MYERS, and D. LIPMAN. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215(3):403–10.

Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. "Detecting Differential Usage of Exons from RNA-Seq Data." *Genome research* 22(10):2008–17. Retrieved November 6, 2014

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics (Oxford, England)* 30(15):2114–20.

Brantl, Sabine. 2007. "Regulatory Mechanisms Employed by Cis-Encoded Antisense RNAs." *Current opinion in microbiology* 10(2):102–9.

Brouwer, Rutger W. W., Oscar P. Kuipers, and Sacha A. F. T. van Hijum. 2008. "The Relative Value of Operon Predictions." *Briefings in bioinformatics* 9(5):367–75.

Chetal, Kashish and Sarath Chandra Janga. 2015. "OperomeDB: A Database of Condition-Specific Transcription Units in Prokaryotic Genomes." *BioMed research international* 2015:318217.

Cho, Byung-Kwan et al. 2009. "The Transcription Unit Architecture of the Escherichia Coli Genome." *Nature biotechnology* 27(11):1043–49.

Chou, Wen-Chi et al. 2015. "Analysis of Strand-Specific RNA-Seq Data Using Machine Learning Reveals the Structures of Transcription Units in Clostridium Thermocellum." *Nucleic acids research* 43(10):e67.

Chuang, Li-Yeh, Hsueh-Wei Chang, Jui-Hung Tsai, and Cheng-Hong Yang. 2012. "Features for Computational Operon Prediction in Prokaryotes." *Briefings in functional genomics* 11(4):291–99.

Conway, Tyrrell et al. 2014. "Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing." *mBio* 5(4):e01442-14.

Cox, D. R. and N. Reid. 1987. "Parameter Orthogonality and Approximate Conditional

Inference." *Journal of the Royal Statistical Society. Series B (Methodological) Journal of the Royal Statistical Society. Series BMethodological) J. R. Statist. Soc. B* 49(1):1–39.

Dillies, Marie-Agnès et al. 2012. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in bioinformatics* bbs046.

Dongen, S. 2000. "Graph Clustering by Flow Simulation."

Dornenburg, James E., Anne M. Devita, Michael J. Palumbo, and Joseph T. Wade. 2010. "Widespread Antisense Transcription in Escherichia Coli." *mBio* 1(1):e00024-10.

Dühring, Ulf, Ilka M. Axmann, Wolfgang R. Hess, and Annegret Wilde. 2006. "An Internal Antisense RNA Regulates Expression of the Photosynthesis Gene isiA." *Proceedings of the National Academy of Sciences of the United States of America* 103(18):7054–58.

van Duin, M. et al. 1989. "Conserved Pattern of Antisense Overlapping Transcription in the Homologous Human ERCC-1 and Yeast RAD10 DNA Repair Gene Regions." *Molecular and Cellular Biology* 9(4):1794–98.

Ekseth, Ole Kristian, Martin Kuiper, and Vladimir Mironov. 2014. "orthAgogue: An Agile Tool for the Rapid Prediction of Orthology Relations." *Bioinformatics (Oxford, England)* 30(5):734–36.

Enright, A. J., S. V. Dongen, and C. A. Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." *Nucleic Acids Research* 30(7):1575–84.

Faghihi, Mohammad Ali et al. 2010. "Evidence for Natural Antisense Transcript-Mediated Inhibition of microRNA Function." *Genome biology* 11(5):R56.

Filiatrault, Melanie J. et al. 2010. "Transcriptome Analysis of Pseudomonas Syringae Identifies New Genes, Noncoding RNAs, and Antisense Activity." *Journal of bacteriology* 192(9):2359–72.

Fitzgerald, Devon M., Richard P. Bonocora, and Joseph T. Wade. 2014. "Comprehensive Mapping of the Escherichia Coli Flagellar Regulatory Network." *PLoS genetics* 10(10):e1004649.

Fortino, Vittorio, Olli-Pekka Smolander, Petri Auvinen, Roberto Tagliaferri, and Dario Greco. 2014. "Transcriptome Dynamics-Based Operon Prediction in Prokaryotes." *BMC bioinformatics* 15:145.

Gabaldón, Toni and Eugene V. Koonin. 2013. "Functional and Evolutionary Implications of Gene Orthology." *Nature Reviews Genetics* 14(5):360–66.

Gál, József, Attila Szvetnik, Róbert Schnell, and Miklós Kálmán. 2002. "The metD D-

Methionine Transporter Locus of Escherichia Coli Is an ABC Transporter Gene Cluster." *Journal of bacteriology* 184(17):4930–32.

Gama-Castro, Socorro et al. 2011. "RegulonDB Version 7.0: Transcriptional Regulation of Escherichia Coli K-12 Integrated within Genetic Sensory Response Units (Gensor Units)." *Nucleic acids research* 39(Database issue):D98-105.

Gama-Castro, Socorro et al. 2015. "RegulonDB Version 9.0: High-Level Integration of Gene Regulation, Coexpression, Motif Clustering and beyond." *Nucleic Acids Research* 44(D1):gkv1156.

Georg, Jens et al. 2009. "Evidence for a Major Role of Antisense RNAs in Cyanobacterial Gene Regulation." *Molecular systems biology* 5:305.

Georg, Jens and Wolfgang R. Hess. 2011. "Cis-Antisense RNA, Another Level of Gene Regulation in Bacteria." *Microbiology and molecular biology reviews : MMBR* 75(2):286–300.

George, J., M. Castellazzi, and G. Buttin. 1975. "Prophage Induction and Cell Division in E. Coli. III. Mutations sfiA and sfiB Restore Division in Tif and Lon Strains and Permit the Expression of Mutator Properties of Tif." *Molecular & general genetics : MGG* 140(4):309–32.

Goodman, Aaron J., Evan R. Daugharthy, and Junhyong Kim. 2013. "Pervasive Antisense Transcription Is Evolutionarily Conserved in Budding Yeast." *Molecular biology and evolution* 30(2):409–21.

Gower, J. C. 1966. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika* 53(3–4):325–38.

Graham, RL, TS Woodall, and JM Squyres. 2005. "Open MPI: A Flexible High Performance MPI." *Parallel Processing and Applied*

Güell, Marc et al. 2009. "Transcriptome Complexity in a Genome-Reduced Bacterium." *Science* 326(5957):1268–71.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009a. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." *Nucleic acids research* 37(1):1–13.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009b. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature protocols* 4(1):44–57.

Jacob, F., D. Perrin, C. Sanchez, and J. Monod. 1960. "[Operon: A Group of Genes with the Expression Coordinated by an Operator]." *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 250:1727–29.

Kawano, Mitsuoki, L. Aravind, and Gisela Storz. 2007. "An Antisense RNA Controls Synthesis of an SOS-Induced Toxin Evolved from an Antitoxin." *Molecular microbiology* 64(3):738–54.

Kawano, Mitsuoki, April A. Reynolds, Juan Miranda-Rios, and Gisela Storz. 2005. "Detection of 5'- and 3'-UTR-Derived Small RNAs and Cis-Encoded Antisense RNAs in Escherichia Coli." *Nucleic acids research* 33(3):1040–50.

Kröger, Carsten et al. 2012. "The Transcriptional Landscape and Small RNAs of Salmonella Enterica Serovar Typhimurium." *Proceedings of the National Academy of Sciences of the United States of America* 109(20):E1277-86.

Kuzniar, Arnold, Roeland C. H. J. van Ham, Sándor Pongor, and Jack A. M. Leunissen. 2008. "The Quest for Orthologs: Finding the Corresponding Gene across Genomes." *Trends in Genetics* 24(11):539–51.

Langmead, Ben and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature methods* 9(4):357–59.

Lasa, Iñigo et al. 2011. "Genome-Wide Antisense Transcription Drives mRNA Processing in Bacteria." *Proceedings of the National Academy of Sciences of the United States of America* 108(50):20172–77.

Lasa, Iñigo, Alejandro Toledo-Arana, and Thomas R. Gingeras. 2012. "An Effort to Make Sense of Antisense Transcription in Bacteria." *RNA biology* 9(8):1039–44.

Lavorgna, Giovanni et al. 2004. "In Search of Antisense." *Trends in biochemical sciences* 29(2):88–94.

Lawrence, Michael et al. 2013. "Software for Computing and Annotating Genomic Ranges." *PLoS computational biology* 9(8):e1003118.

Lee, Jin-Gyun, Kimberly Q. McKinney, Jean-Luc Mougeot, Herbert L. Bonkovsky, and Sun-Il Hwang. 2013. "Proteomic Strategy for Probing Complementary Lethality of Kinase Inhibitors against Pancreatic Cancer." *Proteomics* 13(23–24):3554–62.

Legewie, Stefan, Dennis Dienst, Annegret Wilde, Hanspeter Herzel, and Ilka M. Axmann. 2008. "Small RNAs Establish Delays and Temporal Thresholds in Gene Expression." *Biophysical journal* 95(7):3232–38.

Levine, Erel, Zhongge Zhang, Thomas Kuhlman, and Terence Hwa. 2007. "Quantitative Characteristics of Gene Regulation by Small RNA." *PLoS biology* 5(9):e229.

Li, Bo, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. 2010. "RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty." *Bioinformatics (Oxford, England)* 26(4):493–500.

Li, Li, Christian J. Stoeckert, and David S. Roos. 2003. "OrthoMCL: Identification of

Ortholog Groups for Eukaryotic Genomes." *Genome research* 13(9):2178–89.

Li, Shan, Xia Dong, and Zhengchang Su. 2013. "Directional RNA-Seq Reveals Highly Complex Condition-Dependent Transcriptomes in E. Coli K12 through Accurate Full-Length Transcripts Assembling." *BMC Genomics* 14(1):520.

Ling, Maurice H. T., Yuguang Ban, Hongxiu Wen, San Ming Wang, and Steven X. Ge. 2013. "Conserved Expression of Natural Antisense Transcripts in Mammals." *BMC genomics* 14(1):243.

Lioliou, Efthimia, Cédric Romilly, Pascale Romby, and Pierre Fechter. 2010. "RNA-Mediated Regulation in Bacteria: From Natural to Artificial Systems." *New biotechnology* 27(3):222–35.

Lybecker, Meghan, Bob Zimmermann, Ivana Bilusic, Nadezda Tukhtubaeva, and Renée Schroeder. 2014. "The Double-Stranded Transcriptome of Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 111(8):3134–39.

Al Mamun, Abu Amar M. et al. 2012. "Identity and Function of a Large Gene Network Underlying Mutagenic Repair of DNA Breaks." *Science (New York, N.Y.)* 338(6112):1344–48.

Mao, Xizeng et al. 2014. "DOOR 2.0: Presenting Operons and Their Functions through Dynamic and Integrated Views." *Nucleic acids research* 42(Database issue):D654-9.

Mazin, P. V. et al. 2014. "Transcriptome Analysis Reveals Novel Regulatory Mechanisms in a Genome-Reduced Bacterium." *Nucleic Acids Research* 42(21):13254–68.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic acids research* 40(10):4288–97.

McClure, Ryan et al. 2013. "Computational Analysis of Bacterial RNA-Seq Data." *Nucleic acids research* 41(14):e140.

Merlin, Christophe, Gregory Gardiner, Sylvain Durand, and Millicent Masters. 2002. "The Escherichia Coli metD Locus Encodes an ABC Transporter Which Includes Abc (MetN), YaeE (MetI), and YaeC (MetQ)." *Journal of bacteriology* 184(19):5513–17.

Murtagh, Fionn and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification* 31(3):274–95.

Nachin, Laurence, Ulf Nannmark, and Thomas Nyström. 2005. "Differential Roles of the Universal Stress Proteins of Escherichia Coli in Oxidative Stress Resistance,

Adhesion, and Motility." *Journal of bacteriology* 187(18):6265–72.

Nicolas, Pierre et al. 2012. "Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus Subtilis." *Science (New York, N.Y.)* 335(6072):1103–6.

Nonaka, Gen, Matthew Blankschien, Christophe Herman, Carol A. Gross, and Virgil A. Rhodius. 2006. "Regulon and Promoter Analysis of the E. Coli Heat-Shock Factor, sigma32, Reveals a Multifaceted Cellular Response to Heat Stress." *Genes & development* 20(13):1776–89.

Okuda, Shujiro et al. 2007. "Characterization of Relationships between Transcriptional Units and Operon Structures in Bacillus Subtilis and Escherichia Coli." *BMC genomics* 8:48.

Oliver, Haley F. et al. 2009. "Deep RNA Sequencing of L. Monocytogenes Reveals Overlapping and Extensive Stationary Phase and Sigma B-Dependent Transcriptomes, Including Multiple Highly Transcribed Noncoding RNAs." *BMC genomics* 10:641.

Opdyke, Jason A., Ju-Gyeong Kang, and Gisela Storz. 2004. "GadY, a Small-RNA Regulator of Acid Response Genes in Escherichia Coli." *Journal of bacteriology* 186(20):6698–6705.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. 1999. "The Use of Gene Clusters to Infer Functional Coupling." *Proceedings of the National Academy of Sciences* 96(6):2896–2901.

Passalacqua, Karla D. et al. 2009. "Structure and Complexity of a Bacterial Transcriptome." *Journal of bacteriology* 191(10):3203–11.

Passalacqua, Karla D. et al. 2012. "Strand-Specific RNA-Seq Reveals Ordered Patterns of Sense and Antisense Transcription in Bacillus Anthracis." *PloS one* 7(8):e43350.

Pavelka, Norman et al. 2004. "A Power Law Global Error Model for the Identification of Differentially Expressed Genes in Microarray Data." *BMC bioinformatics* 5:203.

Perkins, Timothy T. et al. 2009. "A Strand-Specific RNA-Seq Analysis of the Transcriptome of the Typhoid Bacillus Salmonella Typhi." *PLoS genetics* 5(7):e1000569.

Pflaum, K., E. R. Tulman, J. Beaudet, X. Liao, and S. J. Geary. 2015. "Global Changes in Mycoplasma Gallisepticum Phase-Variable Lipoprotein Gene vlhA Expression during In Vivo Infection of the Natural Chicken Host." *Infection and immunity* 84(1):351–55.

Qiu, Yu et al. 2010. "Structural and Operational Complexity of the Geobacter Sulfurreducens Genome." *Genome research* 20(9):1304–11.

R Core Team. 2015. “R: A Language and Environment for Statistical Computing.”

Raghavan, Rahul, Daniel B. Sloan, and Howard Ochman. 2012. “Antisense Transcription Is Pervasive but Rarely Conserved in Enteric Bacteria.” *mBio* 3(4):e00156-12-e00156-12.

Raman, Babu, M. P. Nandakumar, Vignesh Muthuvijayan, and Mark R. Marten. 2005. “Proteome Analysis to Assess Physiological Changes in Escherichia Coli Grown under Glucose-Limited Fed-Batch Conditions.” *Biotechnology and bioengineering* 92(3):384–92.

Rasmussen, Simon, Henrik Bjørn Nielsen, and Hanne Jarmer. 2009. “The Transcriptionally Active Regions in the Genome of Bacillus Subtilis.” *Molecular Microbiology* 73(6):1043–57.

Remm, M., C. E. Storm, and E. L. Sonnhammer. 2001. “Automatic Clustering of Orthologs and in-Paralogs from Pairwise Species Comparisons.” *Journal of molecular biology* 314(5):1041–52.

Rossignol, Fabrice et al. 2004. “Natural Antisense Transcripts of HIF-1alpha Are Conserved in Rodents.” *Gene* 339:121–30.

Saadeh, Bashir et al. 2015. “Transcriptome-Wide Identification of Hfq-Associated RNAs in Brucella Suis by Deep Sequencing.” *Journal of bacteriology* 198(3):427–35.

Salgado, H., G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. 2000. “Operons in Escherichia Coli: Genomic Analyses and Predictions.” *Proceedings of the National Academy of Sciences of the United States of America* 97(12):6652–57.

Salgado, Heladia et al. 2013. “RegulonDB v8.0: Omics Data Sets, Evolutionary Conservation, Regulatory Phrases, Cross-Validated Gold Standards and More.” *Nucleic acids research* 41(Database issue):D203-13.

Saveanu, Cosmin et al. 2002. “Structural and Nucleotide-Binding Properties of YajQ and YnaF, Two Escherichia Coli Proteins of Unknown Function.” *Protein science : a publication of the Protein Society* 11(11):2551–60.

Selinger, D. W. et al. 2000. “RNA Expression Analysis Using a 30 Base Pair Resolution Escherichia Coli Genome Array.” *Nature biotechnology* 18(12):1262–68.

Sharma, Cynthia M. et al. 2010. “The Primary Transcriptome of the Major Human Pathogen Helicobacter Pylori.” *Nature* 464(7286):250–55.

Siefert, J. L., K. A. Martin, F. Abdi, W. R. Widger, and G. E. Fox. 1997. “Conserved Gene Clusters in Bacterial Genomes Provide Further Support for the Primacy of RNA.” *Journal of molecular evolution* 45(5):467–72.

Siqueira, Franciele Maboni, Augusto Schrank, and Irene Silveira Schrank. 2011.

"Mycoplasma Hyopneumoniae Transcription Unit Organization: Genome Survey and Prediction." *DNA research : an international journal for rapid publication of reports on genes and genomes* 18(6):413–22.

Slonczewski, Joan L. 2010. "Concerns about Recently Identified Widespread Antisense Transcription in Escherichia Coli." *mBio* 1(2).

Sonnhammer, Erik L. .. and Eugene V Koonin. 2002. "Orthology, Paralogy and Proposed Classification for Paralog Subtypes." *Trends in Genetics* 18(12):619–20.

Sorek, Rotem and Pascale Cossart. 2010. "Prokaryotic Transcriptomics: A New View on Regulation, Physiology and Pathogenicity." *Nature reviews. Genetics* 11(1):9–16.

Swamy, Krishna B. S., Chih-Hsu Lin, Ming-Ren Yen, Chuen-Yi Wang, and Daryi Wang. 2014. "Examining the Condition-Specific Antisense Transcription in S. Cerevisiae and S. Paradoxus." *BMC Genomics* 15(1):521.

Tatusov, Roman L. et al. 2003. "The COG Database: An Updated Version Includes Eukaryotes." *BMC bioinformatics* 4:41.

Thomason, Maureen et al. 2015. "Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia Coli." *Journal of bacteriology* 197(1):18–28.

Thomason, Maureen Kiley and Gisela Storz. 2010. "Bacterial Antisense RNAs: How Many Are There, and What Are They Doing? *." *Annual Review of Genetics* 44(1):167–88.

Toledo-Arana, Alejandro et al. 2009. "The Listeria Transcriptional Landscape from Saprophytism to Virulence." *Nature* 459(7249):950–56.

Toledo-Arana, Alejandro and Cristina Solano. 2010. "Deciphering the Physiological Blueprint of a Bacterial Cell: Revelations of Unanticipated Complexity in Transcriptome and Proteome." *BioEssays : news and reviews in molecular, cellular and developmental biology* 32(6):461–67.

Vasil'eva, S. V and E. V Makhova. 2003. "[Heat Shock Inhibits the Induced Expression of the SOS Genes and SoxRS Regulons in Escherichia Coli]." *Genetika* 39(8):1033–38.

Vijayan, Vikram, Isha H. Jain, and Erin K. O'Shea. 2011. "A High Resolution Map of a Cyanobacterial Transcriptome." *Genome biology* 12(5):R47.

Wade, Joseph T. and David C. Grainger. 2014. "Pervasive Transcription: Illuminating the Dark Matter of Bacterial Transcriptomes." *Nature Reviews Microbiology* 12(9):647–53.

Wang, Yejun, Keith D. MacKenzie, and Aaron P. White. 2015. "An Empirical Strategy

to Detect Bacterial Transcript Structure from Directional RNA-Seq Transcriptome Data." *BMC genomics* 16:359.

Wolf, Y. I. 2001. "Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context." *Genome Research* 11(3):356–72.

Wu, Yonghong et al. 2012. "Glyceraldehyde-3-Phosphate Dehydrogenase: A Universal Internal Control for Western Blots in Prokaryotic and Eukaryotic Cells." *Analytical biochemistry* 423(1):15–22.

Wurtzel, Omri et al. 2010. "A Single-Base Resolution Map of an Archaeal Transcriptome." *Genome research* 20(1):133–41.

Wurtzel, Omri et al. 2012. "Comparative Transcriptomics of Pathogenic and Non-Pathogenic Listeria Species." *Molecular systems biology* 8:583.

Yassour, Moran et al. 2010. "Strand-Specific RNA Sequencing Reveals Extensive Regulated Long Antisense Transcripts That Are Conserved across Yeast Species." *Genome biology* 11(8):R87.

Yoder-Himes, D. R. et al. 2009. "Mapping the Burkholderia Cenocepacia Niche Response via High-Throughput Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 106(10):3976–81.

APPENDIX A: SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER FOUR

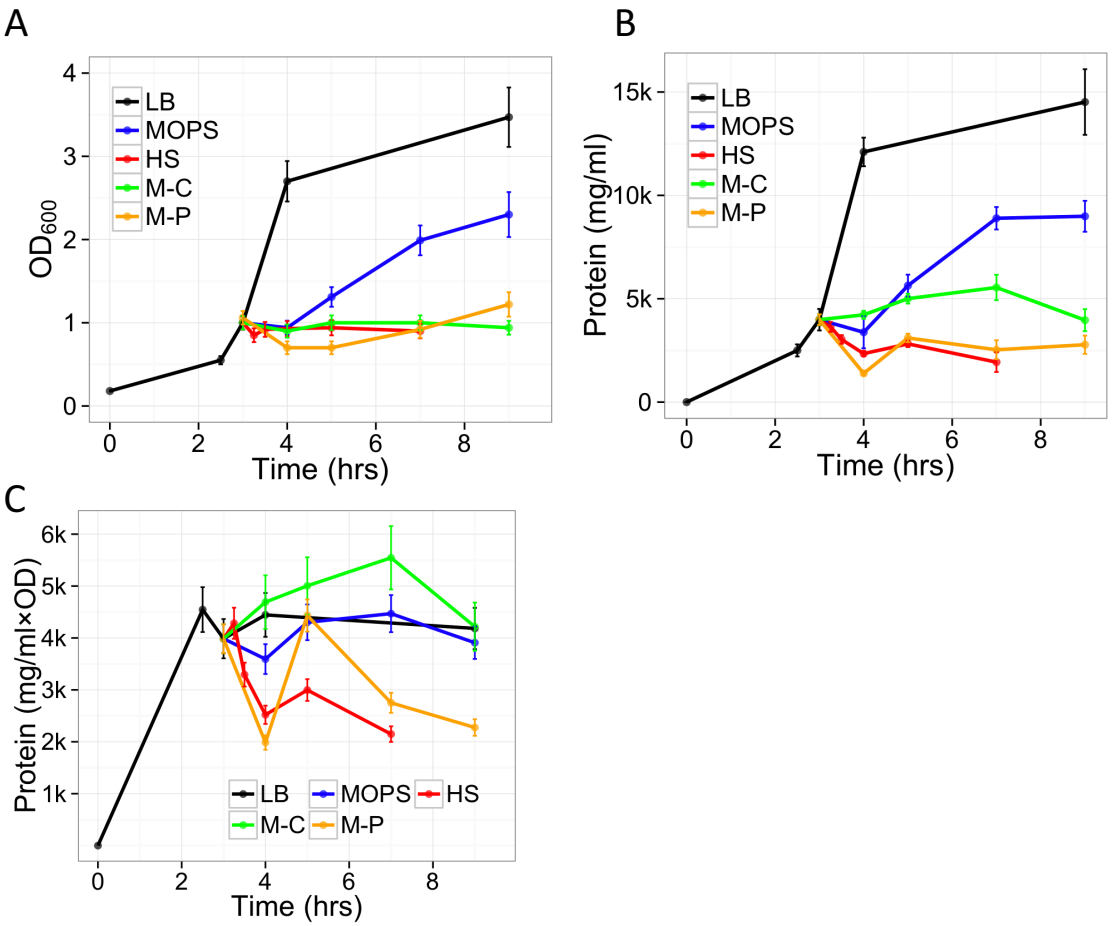


Figure A1: Cell growth and protein concentration at the indicated time points under the five culture conditions.
(A) Average optical density of the cells measured at 600nm. (B) Average protein concentration of the cells. (C) Average protein concentration normalized by the OD_{600} values of the samples.

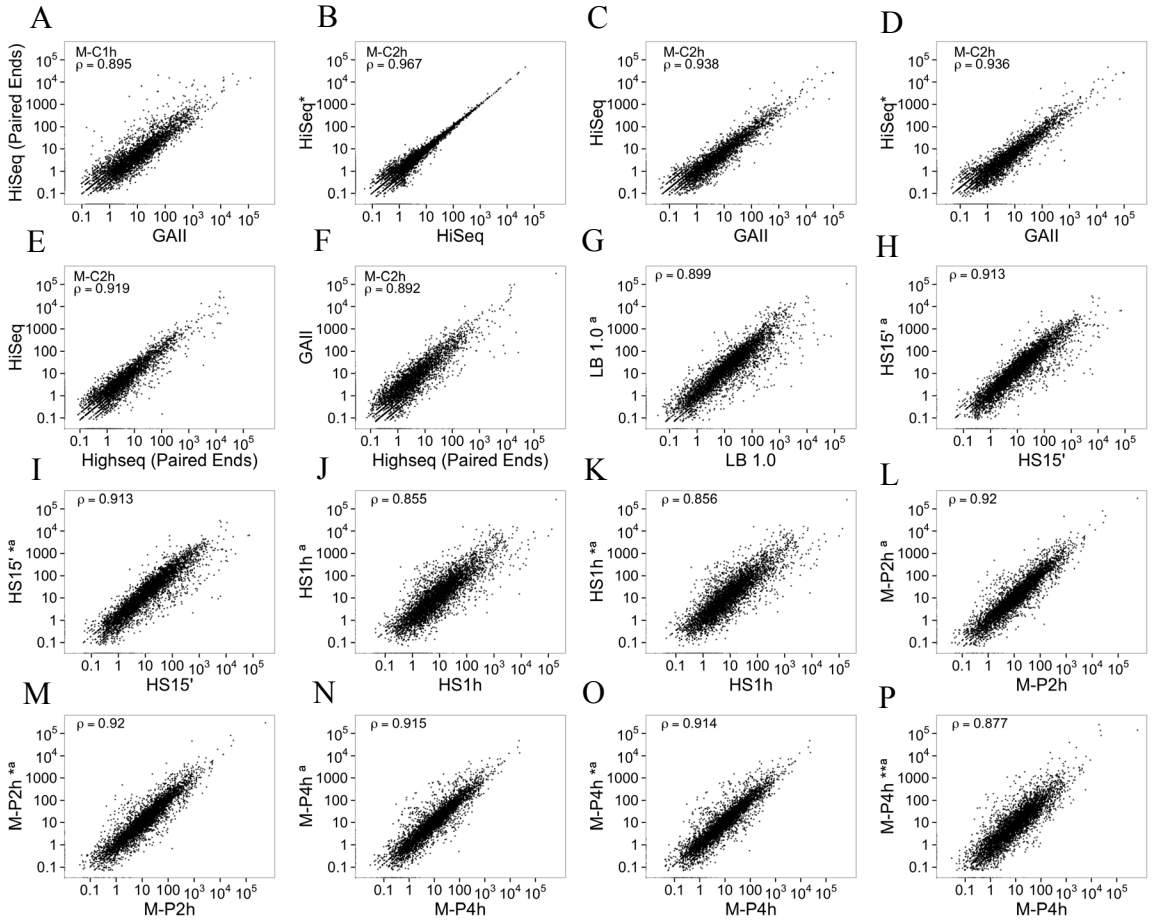


Figure A2: Correlation of mRNA levels of genes between any two replicates for the samples.

Each dot represents a gene. The expression levels are in TPM values. The Pearson correlation (ρ) of the expression levels is shown in each plot. **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, HiSeq reads and HiSeq* reads. **C**. Correlation of expression levels for M-C2h between two biological replicates sequenced on GAII reads and HiSeq reads. **D**. Correlation of expression levels for M-C2h between GAII reads and HiSeq* reads. **E**. Correlation of expression levels for M-C2h between two biological replicates sequenced by paired-ends HiSeq and HiSeq. **F**. Correlation of expression levels for M-C2h between two biological replicates sequenced by paired-ends HiSeq reads and GAII reads. **G**. Correlation of expression levels for LB 1.0 between two biological replicates. **H** and **I**. Correlation of expression levels for HS 15' between three biological replicates. **J** and **K**. Correlation of expression levels for HS1h between three biological replicates. **L** and **M**. Correlation of expression levels for M-P2h between three biological replicates. **N**, **O** and **P**. Correlation of expression levels for M-P4h between four biological replicates.

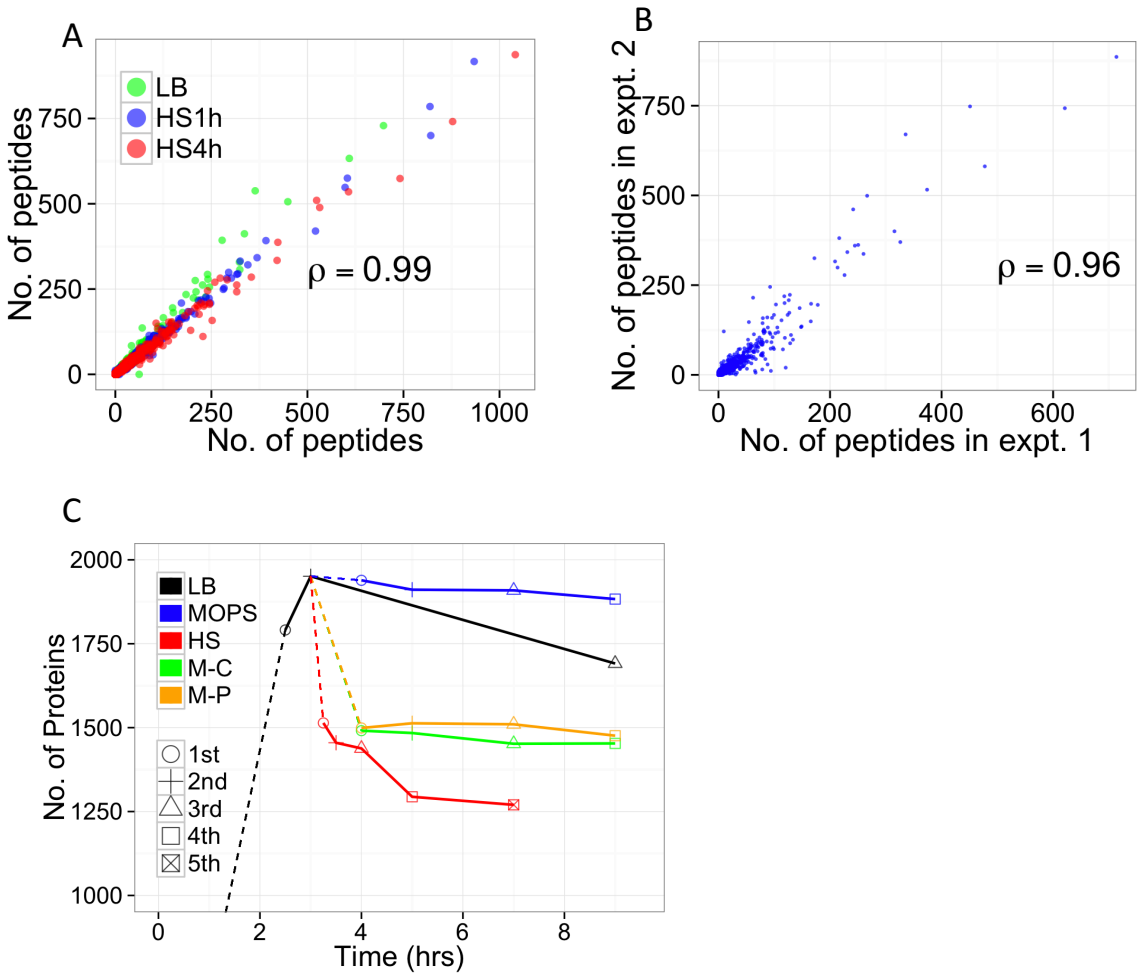


Figure A3: Preliminary proteomics analysis.
(A) Correlation of the number of peptides detected for proteins between two technical replicates for HS1h, HS4h and LB1.0. (B) Correlation of the number of peptides detected for proteins by two biological replicates for LB1.0. Each dot in (A) and (B) represents a protein. (C) Number of proteins detected in each samples taken at the indicated time points under the five culture conditions.

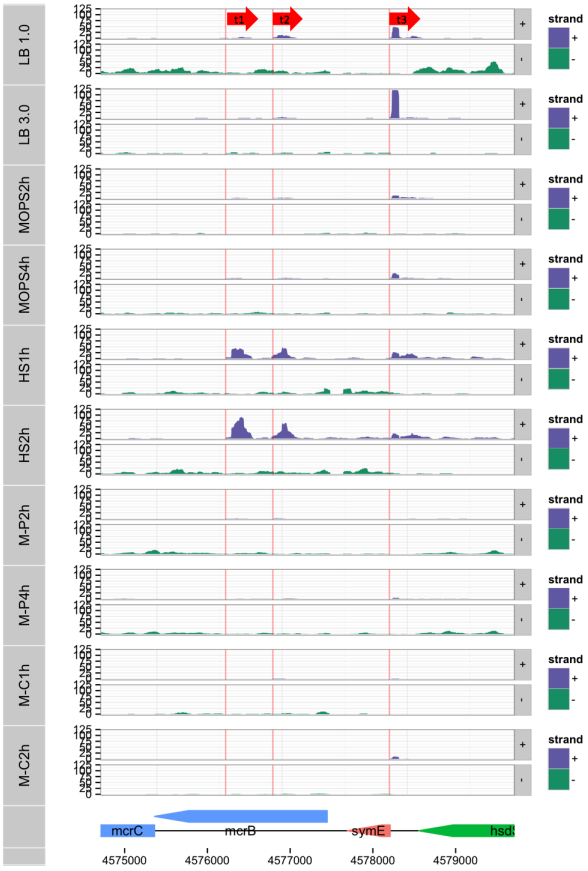


Figure A4: An example of predicted antisense TSSs.
The asRNAs initiating at t_1 and t_2 overlap with two genes, *mcrB* and *symE*, whereas the asRNA initiating at t_3 (annotated as *symR*) only overlaps one gene. t_3 is observed in LB, MOPS and HS growth conditions, but t_1 and t_2 only occurs under heat shock growth condition. Expression levels are shown in five growth condition. Mapped reads to the sense strand is shown in green, while reads mapped to antisense is shown in purple. Predicted antisense TSS are marked by vertical lines. Genes of the same color are in the same annotated operon.

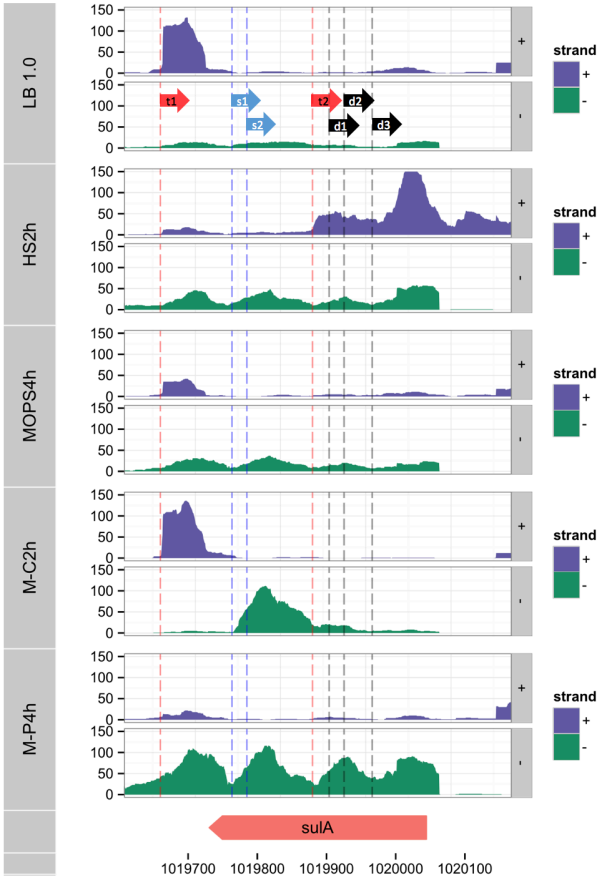
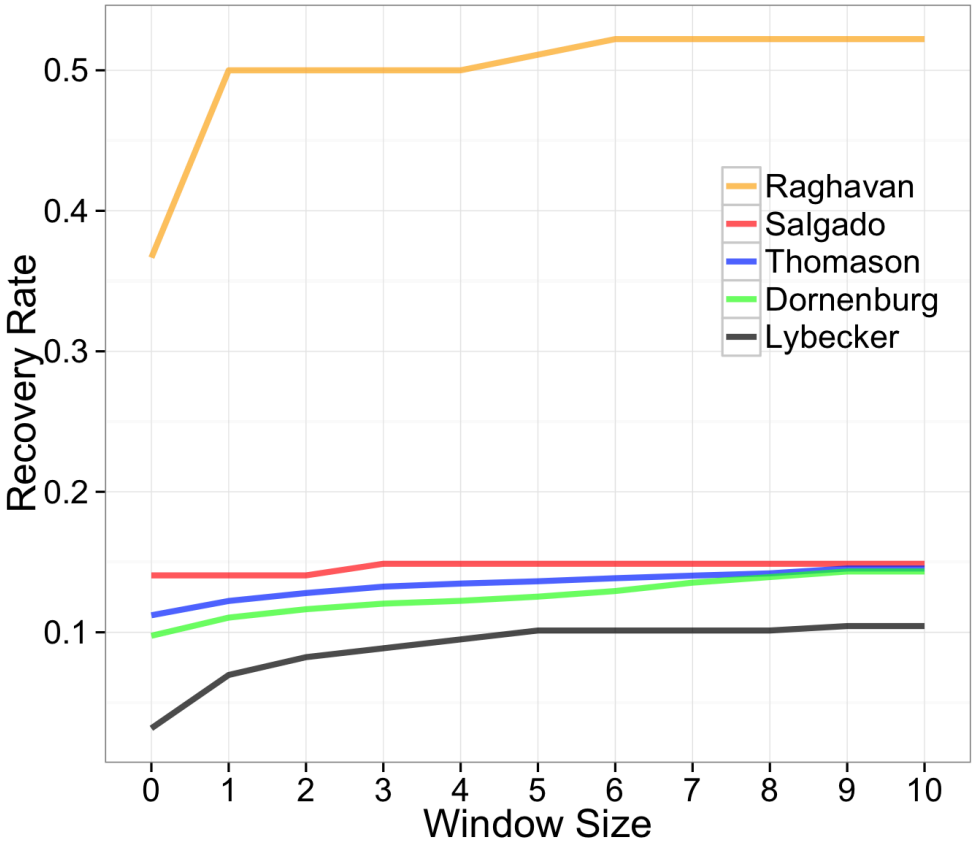


Figure A5: Expression levels of the gene *sulA* in five growth condition.
Mapped reads to the sense strand is shown in green, while reads mapped to antisense is shown in purple. Predicted antisense TSS are marked by vertical lines. Our predicted antisense TSSs are marked in red marked t_1 and t_2, blue arrows marked s_1 and s_2 were from Thomason and Storz 2014, and black arrows marked d_1, d_2 and d_3 were from Dornenburg 2010. Both t_1 and t_2 were used under all the culture conditions, but t_2 is more preferred under heat shock, while t_1 is more preferred under the other conditions.



FigureA6: Recovery rate of antisense TSSs for each dataset by our predicted ones as a function of the cutoff of distance between the two TSSs.

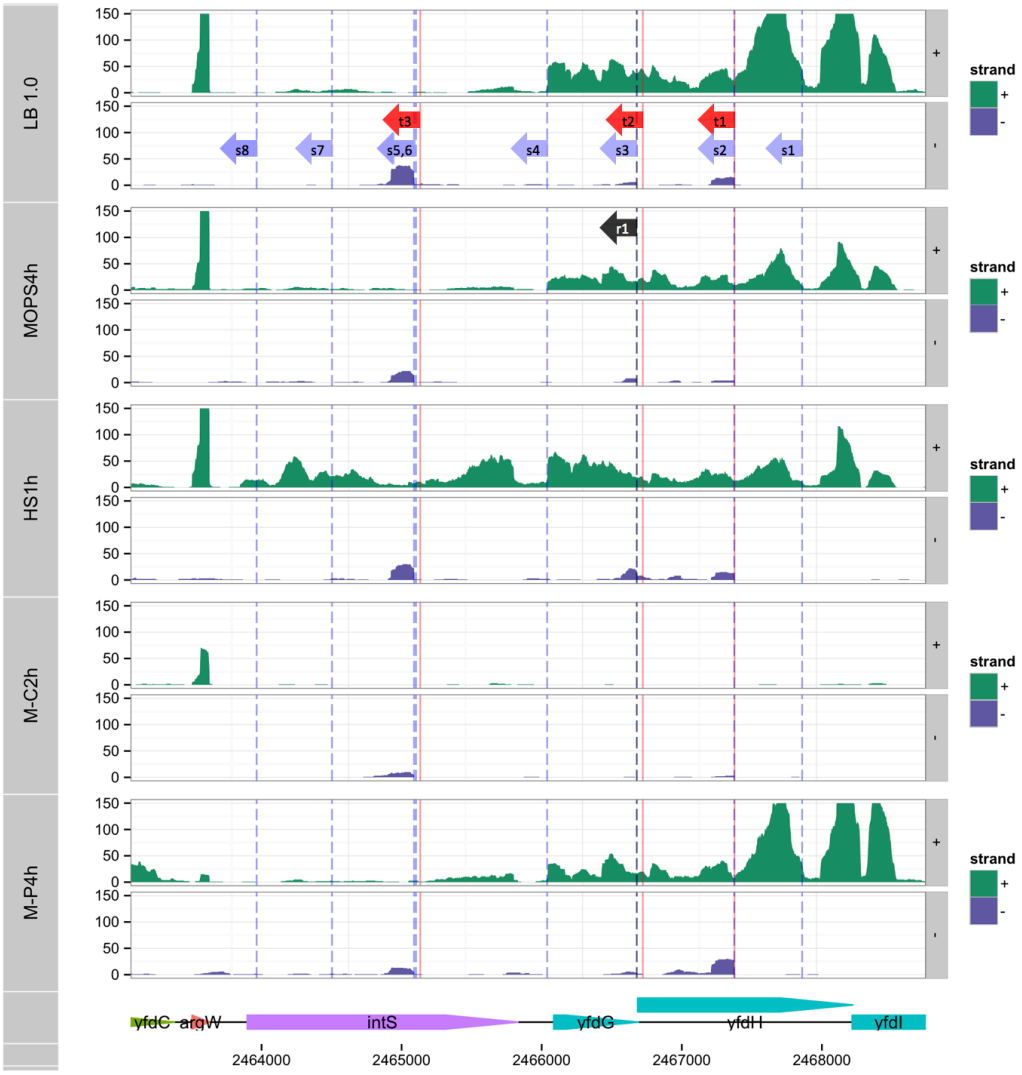


Figure A7: Expression levels of region of genome with predicted asRNA in five growth condition.
Mapped reads to the sense strand is shown in green, while reads mapped to antisense is shown in purple. Predicted antisense TSS are marked by vertical lines. Genes of the same color are in the same annotated operon. Our prediction (red: marked t_1, t_2 and t_3) versus other studies; Thomason and Storz 2014 in blue (marked s_1 through s_9), and Salgado 2010 in black (marked r_1); t_1, s_2 mark the same TSS; t_2, s_3 and r_1 mark the same TSSs; t_3, s_5 and s_6 also mark the same antisense TSS.

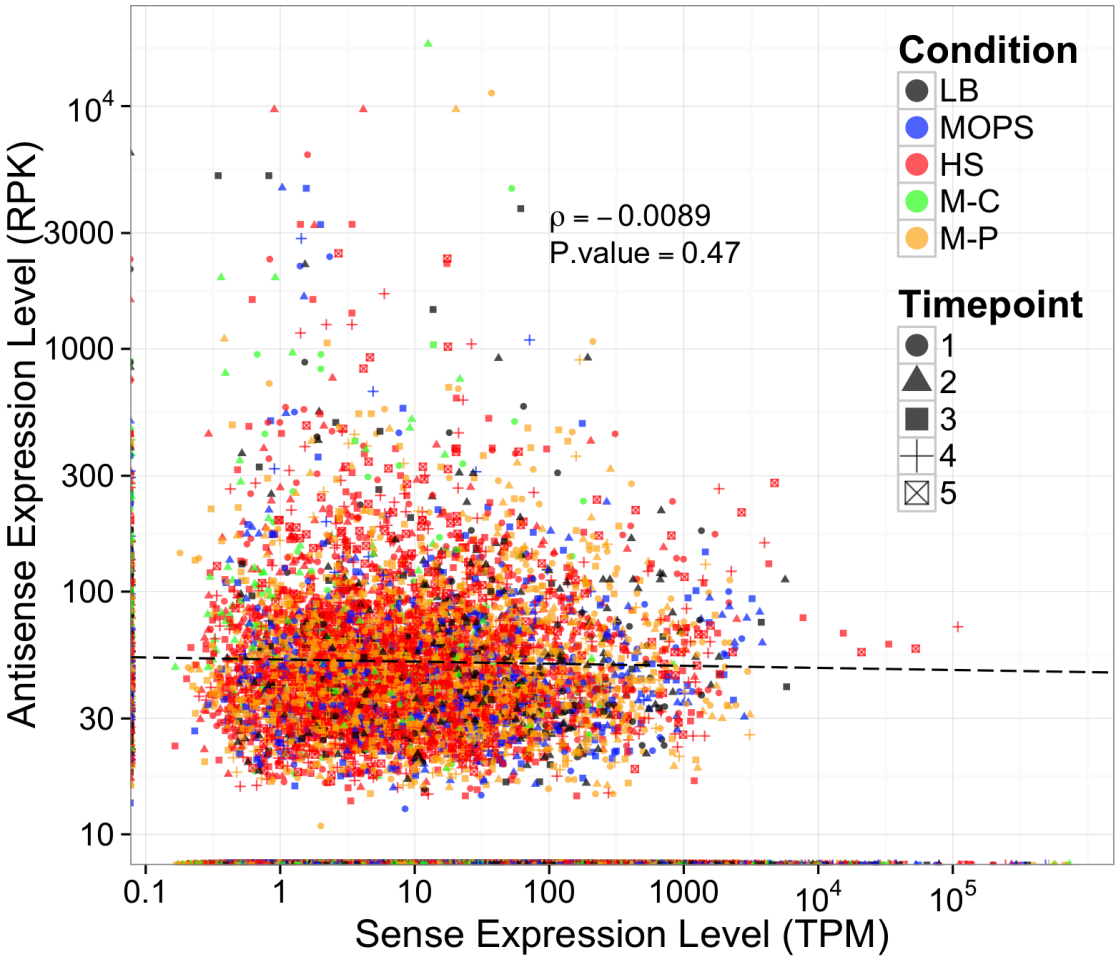


Figure A8: Relationship between mRNA levels and asRNA levels. Spearman correlation coefficient (ρ) and the p-value is plotted on the graph.

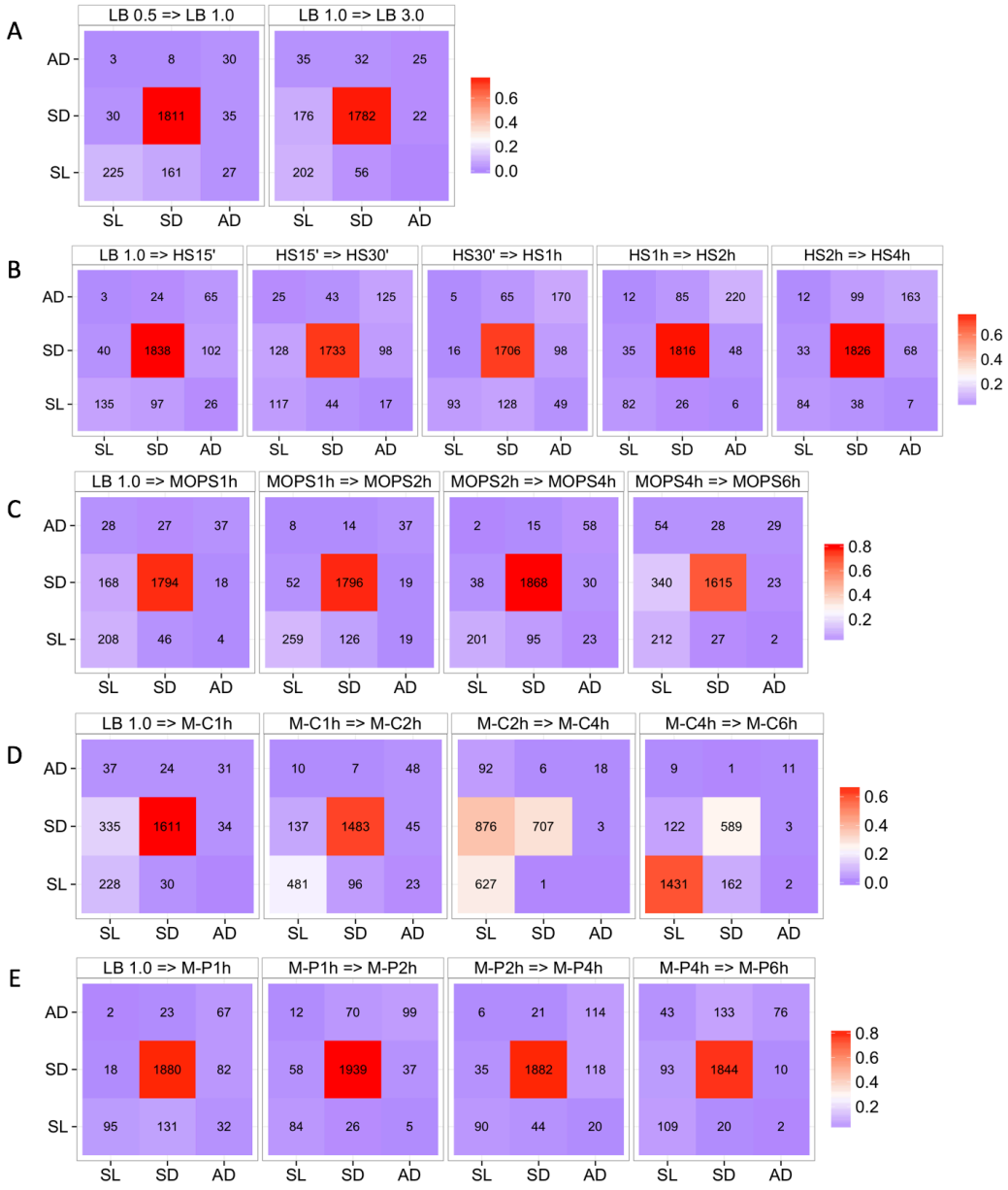


Figure A9: The probability and number of genes that change their transcriptional modes between two adjacent sampling time points in each growth condition.
In each heat map y-axis labels are the starting mode, and x-axis labels are the ending modes. SD, AD and SL stand for sense-dominant, antisense-dominant and silent transcriptional modes, respectively.

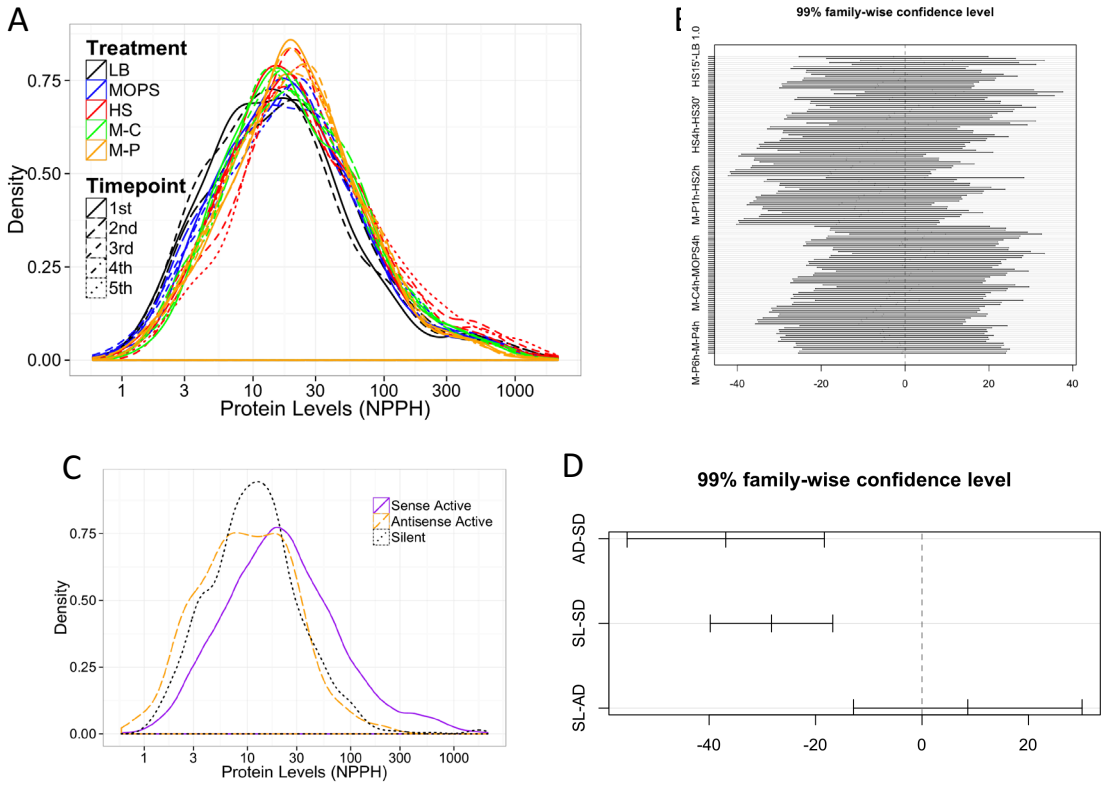


Figure A10: Protein level distributions
(A) Distribution of protein expression levels in NPPH (number of peptides per hundred amino acids) for each time point. (accepted ANOVA null hypothesis p-value=.029>0.01). (B) Tukey honest significant difference of pairwise comparisons shows that all samples are similarly distributed. (C) Distribution of protein expression levels (NPPH) for genes in different transcriptional modes. Data are pooled from all the samples. (rejected ANOVA null hypothesis p-value=2.2e-16<0.01), and (D) Tukey honest significant difference of pairwise comparisons shows that the distribution for antisense-dominant and silent transcriptional modes are similar, but they are differently from that for sense-dominant mode.

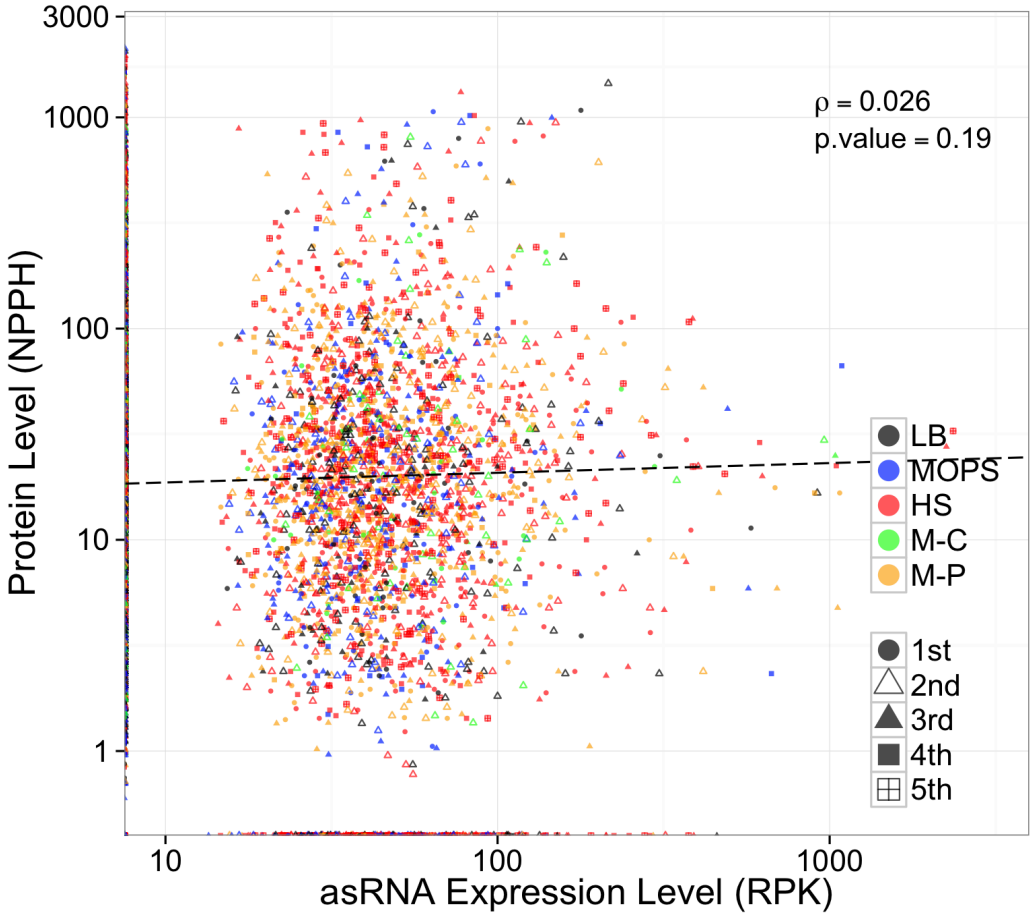


Figure A11: Relationship between antisense (asRNA) transcription levels and protein levels.

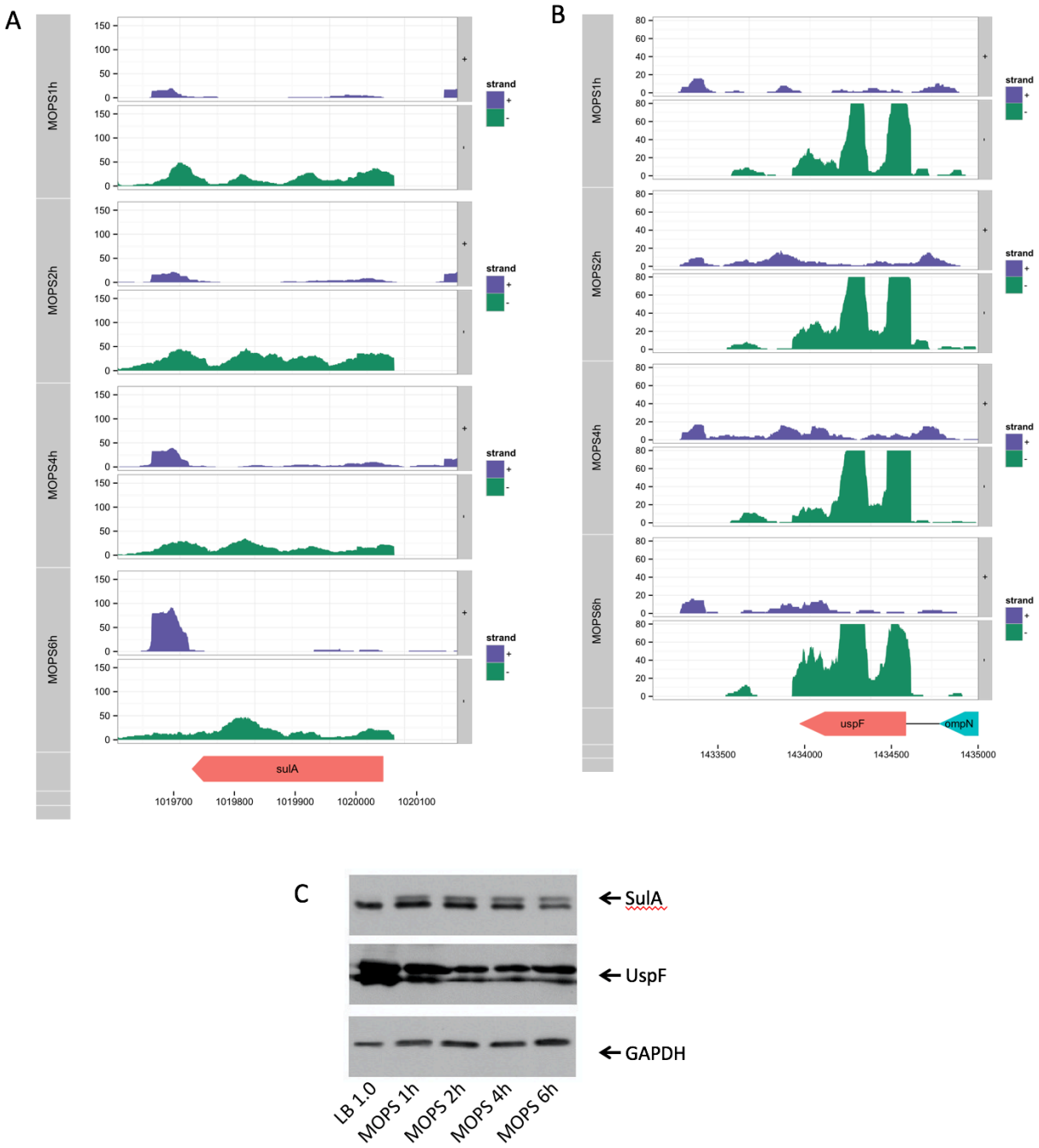


Figure A12: Genes uspF and sulA under MOPS culture condition.
The mRNA and asRNA stayed the same for *sulA* (A) and *uspF* (B). (C) The protein level for both genes stayed the same in MOPS culture.

Table A1: Summary of mapping results of reads for each library and their replicates. Paired-end RNA-seq reads (HiSeq PE) are counted as two. The replicates of the same sampling time point are differentiated by *s. The samples marked with [a] are published with GEO accession GSE48151, while others are published with GEO accession GSE64021.

Sample	Platform	Total Reads	Uniquely mapped reads		Multiple mapped reads		Reads failed to map	
LB 0.5	HiSeq	34,383,742	3,115,619	9.06%	21,868,656	63.60%	9,399,467	27.34%
LB 1.0	HiSeq	43,445,694	4,687,978	10.79%	27,178,069	62.56%	11,579,647	26.65%
	HiSeq [a]	30,342,253	12,957,049	42.70%	15,358,653	50.62%	2,026,551	6.68%
LB 3.0	HiSeq	38,390,801	3,301,474	8.60%	32,783,268	85.39%	2,306,059	6.01%
MOPS1h	HiSeq	27,484,298	5,037,313	18.33%	21,470,283	78.12%	976,702	3.55%
MOPS2h	HiSeq	33,953,509	5,692,231	16.76%	26,916,847	79.28%	1,344,431	3.96%
MOPS4h	HiSeq	31,689,012	6,390,362	20.17%	24,089,233	76.02%	1,209,417	3.82%
MOPS6h	HiSeq	42,082,261	3,207,604	7.62%	36,862,412	87.60%	2,012,245	4.78%
HS15min	HiSeq	38,931,920	6,515,642	16.74%	30,683,557	78.81%	1,732,721	4.45%
	HiSeq [a]	12,893,957	4,064,877	31.53%	7,772,036	60.28%	1,057,044	8.20%
	HiSeq* [a]	13,519,002	4,244,490	31.40%	8,126,717	60.11%	1,147,795	8.49%
HS30min	HiSeq	33,843,406	5,191,738	15.34%	26,361,107	77.89%	2,290,561	6.77%
	HiSeq [a]	12,259,543	4,764,392	38.86%	6,883,572	56.15%	611,579	4.99%
	HiSeq* [a]	12,072,961	4,663,825	38.63%	6,750,012	55.91%	659,124	5.46%
HS1h	HiSeq	42,655,417	6,576,614	15.42%	34,329,913	80.48%	1,748,890	4.10%
	HiSeq [a]	9,862,082	3,153,208	31.97%	6,161,768	62.48%	547,106	5.55%
	HiSeq* [a]	9,900,249	2,837,305	28.66%	5,344,090	53.98%	1,718,854	17.36%
HS2h	HiSeq	38,110,690	4,930,994	12.94%	31,576,875	82.86%	1,602,821	4.21%
HS4h	HiSeq	42,283,291	3,882,423	9.18%	36,292,249	85.83%	2,108,619	4.99%
M-C1h	GAII	32,860,275	2,121,263	6.46%	28,431,917	86.52%	2,307,095	7.02%
	HiSeq PE	52,569,894	1,866,957	3.55%	23,453,919	44.61%	27,249,018	51.83%
M-C2h	GAII	34,227,127	2,911,138	8.51%	29,873,432	87.28%	1,442,557	4.21%
	HiSeq	20,626,282	2,565,836	12.44%	16,687,789	80.91%	1,372,657	6.65%
	HiSeq*	19,938,937	2,497,399	12.53%	16,199,627	81.25%	1,241,911	6.23%
	HiSeq PE	51,499,346	1,336,844	2.60%	23,517,300	45.67%	26,645,202	51.74%
M-C4h	HiSeq PE	48,347,366	1,336,794	2.76%	21,853,322	45.20%	25,157,250	52.03%
M-C6h	HiSeq PE	47,885,922	1,150,280	2.40%	22,027,709	46.00%	24,707,933	51.60%
M-P1h	HiSeq	25,761,215	9,573,378	37.16%	14,882,741	57.77%	1,305,096	5.07%
M-P2h	HiSeq	29,160,891	7,798,985	26.74%	20,118,173	68.99%	1,243,733	4.27%
	HiSeq [a]	25,756,744	11,924,090	46.30%	12,810,332	49.74%	1,022,322	3.97%
	HiSeq* [a]	30,991,286	14,243,021	45.96%	15,350,582	49.53%	1,397,683	4.51%
M-P4h	HiSeq	26,468,056	10,019,459	37.85%	14,854,915	56.12%	1,593,682	6.02%
	HiSeq [a]	17,950,016	7,124,864	39.69%	9,692,405	54.00%	1,132,747	6.31%
	HiSeq* [a]	19,726,630	7,771,208	39.39%	10,629,750	53.89%	1,325,672	6.72%
	HiSeq** [a]	30,291,626	10,003,921	33.03%	18,734,784	61.85%	1,552,921	5.13%
M-P6h	HiSeq	23,830,918	4,725,480	19.83%	17,889,369	75.07%	1,216,069	5.10%

Table A2: Summary of expressed genes and genes with asRNA transcription in each sample.
Sense coverage is the portion of the coding regions covered by reads. Sense depth is the average number of reads covering a nucleotide. Antisense coverage is the portion of the antisense strands of coding regions covered by reads. AS Depth indicates the average number of reads covering the transcribed regions.

Sample	Expressed Genes		Sense Coverage	Sense Depth	Genes with reads mapped to AS strand		Antisense Coverage	AS Depth
LB 0.5	4,289	93.91%	69.42%	49	3,036	66.48%	12.61%	3
LB 1.0	4,379	95.88%	77.68%	91	3,448	75.50%	18.78%	4
LB 3.0	4,280	93.72%	68.60%	66	2,800	61.31%	10.90%	3
MOPS1h	4,317	94.53%	69.10%	78	3,127	68.47%	13.43%	3
MOPS2h	4,368	95.64%	74.37%	100	3,334	73.00%	16.61%	3
MOPS4h	4,418	96.74%	78.06%	110	3,570	78.17%	20.78%	3
MOPS6h	4,149	90.85%	58.43%	79	2,543	55.68%	8.42%	2
HS15min	4,432	97.04%	77.76%	119	3,791	83.01%	24.14%	4
HS30min	4,497	98.47%	77.71%	87	4,080	89.34%	31.73%	2
HS1h	4,487	98.25%	77.73%	100	4,003	87.65%	30.47%	4
HS2h	4,426	96.91%	68.48%	96	3,715	81.34%	20.71%	5
HS4h	4,497	98.47%	76.50%	58	3,945	86.38%	28.83%	4
M-C1h	4,130	90.43%	59.93%	107	2,344	51.32%	7.67%	3
M-C2h	4,130	90.43%	57.66%	189	2,500	54.74%	8.25%	4
M-C4h	2,941	64.40%	20.43%	112	787	17.23%	1.44%	2
M-C6h	2,953	64.66%	21.23%	92	797	17.45%	1.51%	2
M-P1h	4,465	97.77%	85.56%	179	3,925	85.94%	33.72%	5
M-P2h	4,419	96.76%	81.86%	151	3,752	82.15%	25.43%	5
M-P4h	4,457	97.59%	81.10%	182	3,909	85.59%	29.70%	6
M-P6h	4,374	95.77%	75.99%	102	3,440	75.32%	17.81%	4

Table A3: Distribution of the uniquely mapped nucleotides (nt) on the coding regions (sense and antisense) and intergenic regions.

Sample	Total nt counts	Sense nt		Antisense nt		Intergenic nt	
LB 0.5	158,568,439	142,302,847	89.74%	1,408,627	0.89%	14,856,965	9.37%
LB 1.0	333,543,506	293,455,221	87.98%	2,902,903	0.87%	37,185,382	11.15%
LB 3.0	206,532,997	187,324,150	90.70%	1,243,750	0.60%	17,965,097	8.70%
MOPS1h	249,619,397	225,407,534	90.30%	1,680,717	0.67%	22,531,146	9.03%
MOPS2h	343,401,281	309,008,431	89.98%	2,005,533	0.58%	32,387,317	9.43%
MOPS4h	398,865,535	358,481,753	89.88%	2,856,358	0.72%	37,527,424	9.41%
MOPS6h	204,339,350	190,681,235	93.32%	872,050	0.43%	12,786,065	6.26%
HS15min	436,263,306	385,908,445	88.46%	3,857,057	0.88%	46,497,804	10.66%
HS30min	321,685,797	282,336,463	87.77%	3,139,685	0.98%	36,209,649	11.26%
HS1h	374,378,661	324,391,262	86.65%	5,280,883	1.41%	44,706,516	11.94%
HS2h	313,464,390	273,685,966	87.31%	4,615,799	1.47%	35,162,625	11.22%
HS4h	217,783,942	182,981,872	84.02%	4,387,387	2.01%	30,414,683	13.97%
M-C1h	300,393,399	265,757,262	88.47%	949,627	0.32%	33,686,510	11.21%
M-C2h	503,350,565	452,392,612	89.88%	1,300,119	0.26%	49,657,834	9.87%
M-C4h	98,718,814	94,853,901	96.08%	146,095	0.15%	3,718,818	3.77%
M-C6h	84,333,157	81,451,011	96.58%	137,509	0.16%	2,744,637	3.25%
M-P1h	710,099,396	635,528,110	89.50%	6,769,135	0.95%	67,802,151	9.55%
M-P2h	563,571,190	514,083,619	91.22%	5,658,292	1.00%	43,829,279	7.78%
M-P4h	663,534,418	614,435,857	92.60%	6,940,523	1.05%	42,158,038	6.35%
M-P6h	349,998,125	323,200,256	92.34%	2,892,909	0.83%	23,904,960	6.83%

Table A4: Summary of changes of transcriptional mode of genes under each growth condition.
Most two dominant transitions are shaded for each condition. HS and M-P show similar patterns. Also, M-C, MOPS show a similar pattern.

Mode Changes	LB		MOPS		HS		M-C		M-P	
Sense Dominant (SX)	1,697	72.83%	1,538	66.01%	1,578	67.73%	586	25.15%	1,781	76.44%
Antisense Dominant (AX)	16	0.69%	18	0.77%	75	3.22%	5	0.21%	52	2.23%
Silent (NO)	188	8.07%	164	7.04%	54	2.32%	477	20.47%	63	2.70%
No mode change	**1,901**	**81.59%**	**1,720**	**73.82%**	**1,707**	**73.26%**	**1,068**	**45.84%**	**1,896**	**81.37%**
SD ⇒ AD	45	1.93%	37	1.59%	219	9.40%	8	0.34%	111	4.76%
SD ⇒ SL	128	5.49%	341	14.64%	125	5.36%	1,032	44.29%	120	5.15%
AD ⇒ SD	10	0.43%	14	0.60%	85	3.65%	5	0.21%	87	3.73%
AD ⇒ SL	11	0.47%	26	1.12%	15	0.64%	50	2.15%	24	1.03%
SL ⇒ SD	195	8.37%	126	5.41%	76	3.26%	99	4.25%	43	1.85%
SL ⇒ AD	20	0.86%	21	0.90%	17	0.73%	24	1.03%	9	0.39%
One mode change	**409**	**17.55%**	**565**	**24.25%**	**537**	**23.05%**	**1,218**	**52.27%**	**394**	**16.91%**
Multiple mode changes	**20**	**0.86%**	**45**	**1.93%**	**86**	**3.69%**	**44**	**1.89%**	**40**	**1.72%**

Table A5: Spearman correlation coefficient between protein levels and mRNA levels for genes for each sample.

Sample	No. of Genes	Spearman Coef Correlation	P-Value
LB 0.5	920	0.63	7.27E-103
LB 1.0	1,019	0.60	2.53E-99
LB 3.0	867	0.55	7.22E-71
MOPS1h	985	0.58	9.20E-91
MOPS2h	981	0.54	3.13E-74
MOPS4h	971	0.61	1.92E-98
MOPS6h	932	0.46	2.40E-49
HS15'	802	0.52	6.63E-56
HS30'	764	0.45	2.60E-39
HS1h	764	0.43	5.42E-35
HS2h	694	0.42	9.03E-32
HS4h	683	0.47	6.93E-39
M-C1h	755	0.47	5.74E-43
M-C2h	746	0.41	9.68E-32
M-C4h	404	0.31	1.09E-10
M-C6h	421	0.23	1.40E-06
M-P1h	773	0.56	5.83E-66
M-P2h	782	0.59	3.74E-73
M-P4h	772	0.62	1.20E-81
M-P6h	748	0.62	1.26E-79

Table A6: Spearman correlation of sense transcription (mRNA) and protein levels in every sample in the sense-dominant and antisense dominant transcriptional modes. `No. of Genes` column list the number of genes in each sample/mode with both protein and mRNA values. Rows with less than 10 genes are excluded from correlation coefficient analysis. Only heat shock growth condition has enough data points for analysis at every time point.

Sample	Sense Dominant Genes			Antiense Dominant Genes		
	No. of Genes	Spearman Coef Correlation	P-Value	No. of Genes	Spearman Coef Correlation	P-Value
LB 0.5	920	0.630	7.27E-103	0	-	-
LB 1.0	1,017	0.595	2.57E-98	2	-	-
LB 3.0	863	0.552	6.17E-70	4	-	-
MOPS1h	982	0.581	1.32E-89	3	-	-
MOPS2h	975	0.538	3.29E-74	6	-	-
MOPS4h	960	0.607	1.55E-97	11	-0.506	1.13E-01
MOPS6h	926	0.454	3.33E-48	6	-	-
HS15'	780	0.514	6.53E-54	22	-0.199	3.74E-01
HS30'	727	0.438	2.39E-35	37	0.319	5.40E-02
HS1h	715	0.417	1.72E-31	49	0.120	4.12E-01
HS2h	664	0.415	4.69E-29	30	0.061	7.49E-01
HS4h	650	0.456	9.80E-35	33	0.351	4.49E-02
M-C1h	748	0.470	2.40E-42	7	-	-
M-C2h	728	0.408	1.28E-30	18	-0.354	1.50E-01
M-C4h	403	0.315	9.36E-11	1	-	-
M-C6h	421	0.233	1.40E-06	0	-	-
M-P1h	764	0.566	4.82E-66	9	-	-
M-P2h	772	0.584	1.09E-71	10	-	-
M-P4h	751	0.621	2.05E-81	21	0.173	4.54E-01
M-P6h	746	0.615	6.69E-79	2	-	-

APPENDIX B: SUPPLEMENTARY DATASETS

Supplementary datasets for this dissertation are as follows

DS2: functional enrichment analysis for the TUs expressed in all TPCs.

DS3A: List of detected asRNAs.

DS3B: List of the genes used for proteomics study.

DS3C: functional enrichment analysis.

These datasets are accessible in http://ehsun.me/go/phd.