UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SINGLE-NUCLEOTIDE RESOLUTION VIEW OF GENE EXPRESSION IN *ESCHERICHIA COLI* K-12 UNDER VARIOUS PHYSIOLOGICAL CONDITIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JAMES P. CREECY
Norman, Oklahoma
2015

SINGLE-NUCLEOTIDE RESOLUTION VIEW OF GENE EXPRESSION IN *ESCHERICHIA COLI* K-12 UNDER VARIOUS PHYSIOLOGICAL CONDITIONS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MICROBIOLOGY AND PLANT BIOLOGY

BY

_____
Dr. Tyrrell Conway, Chair

_____
Dr. Anne K. Dunn

_____
Dr. Elizabeth Karr

_____
Dr. Michael J. McInerney

_____
Dr. Cecelia M. Brown

This dissertation is dedicated to all of the exceptional people that have made achieving my goals possible. Without the love, support and contributions provided me by my family, friends, and colleagues none of my accomplishments over the last four years would have been achievable.

To Claire- the sacrifices you have made over the course of this effort have been great, and I cannot begin to thank you enough for all that you have done for me. When times were hard and the odds seemed insurmountable, I took strength from the confidence you had in me.

To Rowan and Greyson- you are my greatest achievement, and all that I do is for each of you. For me obtaining this degree was never about personal achievement, it was always to provide the best life I could for the two you.

Acknowledgements

My time at the University of Oklahoma has been one of the most educational experiences of my life, and I have to acknowledge all of those that play a vital role in my education and the science contained within this document. Foremost, I must acknowledge Dr. Tyrrell Conway. His commitment not only to my academic and professional success has been unmatched and without his guidance and support I would have never earned this degree. Next I must acknowledge my committee members: Drs. Cecelia Brown, Anne Dunn, Elizabeth Karr, and Michael McInerney. In an afternoon in September I learned what it was to be a scientist, and each of my committee members were instrumental in that process. Finally, I would like to recognize Drs. Dwight Adams, Gloria Caddell, Jenna Hellack, and Charlotte Simmons for giving me the freedom to complete my education while maintaining the job that I love.

Table of Contents

List of Tables

List of Figures

Abstract

We analyzed the transcriptome of Escherichia *coli* K-12 by strand-specific RNA sequencing at single-nucleotide resolution during logarithmic- growth and upon entry into stationary phase under carbon, nitrogen, and phosphate starvation conditions. To generate high-resolution transcriptome maps, we developed a quantitative method for first annotating and then calculating the three features that define an operon: the promoter, terminator, and deep RNA sequence read coverage to connect the two transcript ends. Based upon the annotation of transcription features we were able to calculate relative promoter activities, terminator efficiencies, and transcription unit activities for 2,122 promoters, 1,774 terminators, and 1,510 operons, respectively. Our analyses revealed an unprecedented view of *E. coli* operon architecture. A large proportion (36%) of operons are complex with internal promoters or terminators that generate multiple transcription units. We found that 276 of 370 convergent operons terminate inefficiently, generating complementary 3' transcript ends which overlap on average by 286 nucleotides, and 136 of 388 divergent operons have promoters arranged such that their 5' ends overlap on average by 168 nucleotides. We found 89 antisense transcripts of 397-nucleotide average length, 7 unannotated transcripts within intergenic regions, and 18 sense transcripts that completely overlap operons on the opposite strand. Of 519 overlapping transcripts, 75% correspond to sequences that are highly conserved in *E. coli* (>50 genomes). Additionally, we sought to identify and characterize RpoS-dependent operons, genes and promoters under carbon, phosphate and nitrogen starvation. RpoS-dependency was identified using DEseq software. Following differential expression analysis by DEseq, only transcription units, genes and promoters

that were statistically significant (p-value ≤ 0.05) and demonstrated a 4-fold or greater change in expression were classified. As a result of our analysis 315 operons, 317 genes, and 278 promoters were classified as being RpoS-dependent. It was observed that RpoS-dependency was most impactful when the culture was starved for carbon, accounting for two-times more differentially regulated transcription units than nitrogen or phosphate starvation. Significant differences in the structure of RpoS-dependent transcripts were observed when compared to RpoS-independent transcripts. It was determined that most RpoS-dependent operons are monocistronic and are approximately half the size of RpoS-independent operons. Analysis of the -10 regions of the 278 putative RpoS-dependent promoters determined that the most abundant nucleotide sequence was CTACGCTTAA, a significant deviation from the consensus motif (CTATAATTAA). We hypothesize that the presence of guanine and cytosine nucleotides (CGC) at base locations -8 through -10 results in the preferential binding of RpoS to these promoter regions, whereas the vegetative sigma factor RpoD would not bind. Additionally, four new RpoS-dependent transcripts were identified within the intergenic regions of the *E. coli* genome. These results and conclusions describe RpoS-dependency at the operon, gene, and promoter levels, and elucidate the "core" of the RpoS regulon under three different starvation conditions.

Chapter 1: Literature Review of Global Gene Expression Analysis of Bacterial Transcriptomes by RNA-seq

The work presented in this dissertation focuses on genome-scaled investigations of bacterial gene expression and regulation, and utilizes RNA-seq strategies to provide a level of detail of the *Escherichia coli* transcriptome that has never before been observed. Over the last twenty years, advances in the technology used to study complex biological phenomena have propelled the field of molecular biology to its current state of prominence (1-3). Within the field we have become captivated by the power of these advances, and benefited greatly from what they allow us to achieve (4). When utilized correctly, these technologies have and will continue to provide researchers with the insight needed to more rapidly propel science forward. High-throughput sequencing has emerged as one of the most popular technologies by which biological disciplines investigate essential questions(5-8). The tremendous amount of data obtained from high-throughput sequencing techniques has pushed the frontier of our understanding in the fields of personalized medicine (9), whole genome analysis (10), metagenomics (11), and RNA-seq (12). For the first time in history, an entire bacterial transcriptome can be analyzed by directly sequencing the total RNA present, a tool by which we as investigators can use to refine our understanding of bacterial transcription (13-15).

It is well recognized that bacteria regulate the expression of their genes based on the environmental conditions they encounter (16). Bacterial cells exposed to suitable growth conditions will react by rapidly dividing, while under less favorable conditions the cell will halt division, decrease protein production, and in bacteria species where it

is possible they will sporulate (17). The genes responsible for transitioning between these variations in environment do so by restructuring cellular physiology and metabolism, and are under direct regulation by a number of mechanisms, including relative promoter activity, repressor status, transcript secondary folding, and antisense RNA activity. Of great importance to the fields of microbiology and medicine, gene expression studies have begun to explain how pathogens cause disease and elucidate the relationship between the host and pathogens (15). The study of gene regulation and expression has a long and rich history consisting of landmark discoveries, from the *lac* operon to localization of gene products by analyzing gene fusions to green fluorescent protein (18, 19). Until recently, the majority of gene expression mapping was achieved using laborious single-operon analysis techniques such as S1 protection (20), primer extension (21), or 5'-RACE (22). By 1995 single operon techniques were replaced by DNA microarray analysis, and for the first time it was possible to investigate gene expression on a genome-wide scale (23).

History of Bacterial Transcriptomic Studies by DNA Microarray Analysis

Few technological breakthroughs have advanced both biology and medicine more than DNA microarray analysis. At its apex, DNA microarray technology was applied to the investigation of gene identification, alternative splicing events (24), single nucleotide polymorphisms (25), protein-DNA interactions (26), bacterial community ecology (27), and gene expression (28). As influential as DNA microarray analysis has been to both applied and hypothesis based sciences, its origin was humble. In principal, DNA microarray technology was a logical extension of Southern blot

analysis, described in 1975 by Dr. Edwin Southern (29). The truly innovative feature of DNA microarrays was the use of a solid surface (e.g. glass) as a binding surface for oligonucleotide probes. The advantage of a solid surface was apparent. A non-porous surface would allow for a denser configuration of probes, and therefore a greater amount of data could be analyzed.

The first paper on the use of DNA microarray technology for the analysis of gene expression was published in 1995 by Dr. Patrick Brown's group at Stanford University, but the story of the development of this technology starts three years prior. The first grant for the development of DNA microarray for gene expression analysis was submitted to the National Institute of Health (NIH) in November of 1992, and consisted of three aims, 1) develop a DNA microarray system, 2) develop a statistical tool for interpreting the data, and 3) evaluate a genomic infrastructure of human population using the new technology and statistical model (30). The proposal was not well received, and earned a priority score of 344 (3.4 on today's scale). Upon resubmission the grant was dramatically scaled back, consisting of a single aim, and was funded at a substantially diminished level. The struggle to have DNA microarray technology recognized did not end there. In the summer of 1994 at a conference in Holland, Mark Schena (a PhD student in Dr. Brown's lab) presented the microarray concept, and by his own admission was laughed off the stage (31). Undaunted, the development of this important technology continued in spite of the marginal financial support and skepticism from peers. Often scientific progress is thought of as a linear process, moving from one small advance to the next. However, on rare occasions a field experiences a quantum leap, and in those instances true scientific genius emerges.

Starting in the mid-1990s with the publication by Schena *et. al.* (1995) on *Arabidopsis* gene expression, the sentiment of the scientific community towards microarray technology dramatically changed (23). Presented with the overwhelming evidence and the elegant simplicity of the experimental design, the development of DNA microarrays reshaped the manner in which biology was studied. By the late-1990s the cost of manufacturing custom DNA microarrays declined so dramatically that large-scale studies on the entirety of a bacterial transcriptome were achievable (32). Companies like Affymetrix quickly mainstreamed the production of gene expression and tiling arrays for a number of model organisms, including *E coli*. It was at this time, viewing the totality of bacterial transcription, in which our "simplistic" concept of bacterial transcription began to change (33).

DNA microarray technology was instrumental in the development of our understanding of the genetics and physiology of *E. coli*. *E. coli* was the first bacterium analyzed by DNA microarray technology (34, 35), and because of this technology *E. coli* became one of the best understood organisms on Earth (36). Starting in 1999, studies employing *E. coli* DNA microarrays aided in the discovery of unknown genes, identification of pathogenic strains, response to environmental stresses, refining metabolic pathways, and introducing the "modular unit" concept of E. coli transcriptional organization (33). During this period of time the field of microbiology also witnessed the emergence of high quality community resources like GenoBase (37), RegulonDB (38), EcoCyc (39), and GenExpDB. These community resources made the sharing of data accurate, rapid and freely available. On the heels of microarray technology, the field of microbiology has once again been presented with the

opportunity to develop a new technology, RNA-seq, to better understand microbial systems. Aided by the foundational work published by those that implemented DNA microarray technology, the development of RNA-seq is poised to usher in a new era of bacterial transcriptome analysis that promises to offer a view of transcription that was not possible until now.

Short History of RNA-seq Analysis of Bacterial Transcriptomes

Historically, the bacterial transcriptome has been viewed as simplistic, and until recently much of the available evidence would support this conclusion. When comparing the transcriptome of eukaryotes to that of prokaryotes it would be logical to observe the single chromosome, containing minimal intergenic DNA, lacking introns, and organized in discreetly transcribed units and infer that prokaryotic transcription was simplistic. While admittedly bacterial transcriptomes are less complex than those of eukaryotes, it would be amiss to view this as a disadvantage. Instead, bacteria have evolved a number of elegant strategies for orchestrating transcript abundance that allow them to respond to the environmental conditions in the most efficient manner possible. In an effort to better understand the phenomena of bacterial transcription, emerging RNA-seq technologies have been employed to study the entirety of transcription under controlled physiological conditions.

Principles of RNA-seq

Currently, the preferred method for the global-analysis of bacterial transcription is RNA-seq. RNA-seq relies on the massively parallel analysis of complementarity

DNA (cDNA) by high-throughput sequencing. When compared to hybridization-based methods (quantitative reverse transcription PCR or microarray chips) used previously, data obtained from RNA-seq provides the investigator with a number of benefits including: reduction in cost, single-nucleotide resolution, increased sensitivity, and greater robustness. Each of these advantages originates from the fact that transcriptome analysis by RNA-seq functions by a fundamentally different principle than previous methods. Rather than hybridizing to a complementary DNA probe, RNA-seq data are aligned to the nucleotide sequence of a reference genome. The lack of a hybridization probe and the direct sequencing of total RNA results in the interrogation of all transcripts with minimal experimental bias (40). In the absence of such a bias, previously unknown genetic features such as untranslated regions, regulatory small RNAs, operon structure, alternative promoters, and terminators have been identified at an unprecedented rate (41-43). By avoiding the use of hybridization techniques, the data obtained by RNA-seq are more precise and quantifiable(44). The increase in resolution between microarray chip and RNA-seq methods has been dramatic. The length of the probe (~25-50 nucleotides) determines the resolution of a microarray chip, while RNA-seq data can be resolved to a single nucleotide (45). In addition, the lack of non-specific binding between probe and cDNA means that the incidence of false positives becomes virtually nonexistent. The advantages of RNA-seq are numerous, but the development of bacterial specific RNA-seq methods was not without challenges.

RNA-seq Strategies

It should be noted that while protocols for the sequencing of bacterial RNA are now prevalent in the literature, until recently this was not the case. The first RNA-seq studies were conducted on eukaryotic organisms, and were focused on the medical applications of this new technology (46, 47). Guided by the literature available, biotechnology companies manufactured RNA-seq chemistries designed for the preparation and analysis of eukaryotic RNA. These RNA-seq kits relied on genetic features that are unique to eukaryotic organisms (5' cap and 3'-poly-adenylated tail) for transcript isolation and analysis, and were not suitable for the analysis of bacterial transcription. As such, those of us who were studying bacterial transcriptomics resorted to developing novel approaches that were better suited to bacterial transcriptomes (48).

The preparation of cDNA sequencing libraries from total RNA is an essential element of every bacterial transcriptome study. As mentioned above, bacterial transcripts have a number of properties that are unique to the domain bacteria; consequently accurate transcriptome analysis requires an understanding of these characteristics and how they affect library preparation. The first consideration should be the RNA extraction method. Over the last five years the field has undergone a shift in the RNA extraction techniques used. Commercially available extraction kits were often utilized because of the ease of use and brevity of the protocols. Unfortunately, the majority of these kit chemistries used column separation that excluded small RNAs, resulting in a bias (45). An example of the differences between extraction methods can be observed when comparing data obtained from a membrane based method verses an organic solvent method. Membrane filter purification techniques function by capturing

all nucleic acids greater than the pore size of the membrane, so RNA fragments less than 50 nucleotides are often lost during this process. Alternatively, organic purification methods retain the total RNA population because the purification is chemical and not physical and performed in a single tube. While it could be argued that the contribution made by these very small transcripts is insignificant to the overall transcriptome, a growing body of literature indicates that microRNA (~22 nucleotides) and small RNA (50-250 nucleotides) are vital for understanding bacterial gene expression and regulation(49). It therefore becomes necessary that RNA be extracted using a simple hot-phenol method and purified using ethanol precipitation in order to obtain the totality of biologically significant transcripts. (50).

The next consideration is the challenge that ribosomal RNA presents to the construction of a valuable sequencing library. Approximately 90% of all RNA within a bacterial cell is ribosomal(51). While this level of ribosomal RNA (rRNA) is essential for maintaining the health of the bacterial cell, it presents a complication when faced with analyzing the totality of a transcriptome. Due to the disproportional abundance of rRNA compared to all other forms of RNA, the vast majority of sequence data obtained will be from the seven rRNA genes (at least on the *E. coli* transcriptome). In an effort to improve the ratio between rRNA and the remaining 10% of transcripts, two strategies were developed, 1) 5'-dependent terminator exonuclease (TEX) treatment for the degradation of 5' monophosphate RNA and 2) rRNA specific depletion by hybridization.

Ribosomal RNA removal by TEX treatment can be performed on the total RNA sample and in principle provides a method for the enrichment of newly transcribed

primary transcripts (i.e., not processed). The abundance of rRNA inside a cell can be attributed to both the rate of transcription and the stability of the rRNA (52). Stability is best defined as the resistance to degradation, and it is well established that rRNA is one of the most stable and highly transcribed RNA (53). The stability of rRNA is achieved by chemical bonding with ribosomal proteins and a substantial amount of self-annealing to form complex folded structures (54). However, this does not mean that all forms of RNA degradation are prevented. As rRNA begins to degrade one of the first alterations that occurs is a modification to the 5' end of the transcript, 5'-triphosphate ends are converted to 5'-monophosphates (55). As a result of the primary nucleotide triphosphate not forming a phosphodiester linkage with an upstream nucleotide, newly synthesized bacterial transcripts possess a 5'-triphosphate end. As RNA degrades, the conversion to 5'-monophosphate ends functions to mark the RNA for turnover by RNase activities. Because rRNA is so stable, the majority of the 5' ends are monophosphate, while the remainder of the transcripts will persist unaffected and functional. Following RNA extraction and purification, rRNA is no longer stabilized by ribosomal proteins or folding. Treatment with TEX will therefore degrade rRNA preferentially, and greatly increase the probability of sequencing non ribosomal RNA.

Alternatively, rRNA can be selectively hybridized and removed from a sample containing total RNA using a form of affinity chromatography. The genes that encode for rRNA are resilient to genetic mutation. As such, there are regions of the rRNA genes in which the nucleotide sequences are conserved within and between phyla. To enrich for all other forms of RNA, oligonucleotides complementary to the conserved regions of rRNA are synthesized and then bound to silica beads. As the total RNA

sample is exposed to the silica bead column, rRNA becomes bound to the silica beads, while the remainder of the RNA (i.e., mRNA, asRNA, etc.) passes through the column and into the eluent where it becomes available for cDNA synthesis (56).

While both terminator exonuclease treatment and rRNA depletion are effective for the enrichment of non-ribosomal RNA, use of these techniques also introduces an experimental bias (57). It has been determined that the half-life of a given RNA transcript varies depending on the function of that transcript (58). Naturally, this would mean some non-ribosomal RNA transcripts would form stabilization complexes, persist in the cell, and undergo a similar 5' end conversion as that seen in rRNA. The *ompA* mRNA in *E. coli* is an example of this principle in practice. *ompA* transcripts are highly stable due to the abundance of secondary folding associated with the 5′-untranslated region (5′-UTR) of the transcript (59, 60). The single-stranded regions in between the hairpin loops of the 5′-UTR contain RNase E digestion sites. However, RNase E is prevented from accessing these digestion sites while the ribosome is bound to the ribosome-binding site of *ompA* mRNA (61). Because of this, it can be concluded that terminator exonuclease treatment will result in transcripts with the longest half-lives being underrepresented in the sequence data. Similarly, rRNA depletion by hybridization is not exclusively selective for rRNA. Non-specific hybridization has been shown to occur, and as a result these transcripts are never analyzed. While some form of rRNA depletion was essential for transcriptome sequencing studies only three years ago, the practice has been abandoned because of the bias that it introduces. Due to the rapid development of high-throughput sequencing methods and the significant increase in the amount of sequence data that can be obtained, rRNA depletion is no

longer required (62). Instead, rRNA sequences are now managed bioinformatically during the sequence alignment stage of analysis (63). As a result of this new approach, more than 90% of the sequence data is mapped to the seven ribosomal genes, but what remains stands to be the most accurate view of a bacterial transcriptome obtainable by modern science.

Bacterial transcription is distinctive among the domains of life. Common to bacterial transcription is the ability for RNA transcripts to be transcribed from both strands of the genome, which generates interactions between converging and diverging operons that is not common in eukaryotic organisms (44). As such, it is essential that sequencing library preparation is strand-specific, and first-generation commercially available kits were not designed to do this. Many of the early methods for sequencing RNA were not concerned with preserving the strandedness of the transcript. Following RNA extraction and DNA digestion, the total RNA was converted into cDNA through the use of random hexamer-primed reverse transcription. The use of random priming reverse transcription remains a straightforward and rapid method for the conversion of RNA to cDNA, however this process fails to retain strandedness that remains critical for analyzing bacterial transcriptomes accurately. The random nature of primer binding results in the accumulation of sequence reads from the middle of genes, and underestimates the abundance of the 5’ and 3’ ends of transcripts (45). Because bacterial genes are organized in operons, the use of a random priming approach would result in the overestimation of transcripts corresponding to the genes internal to operons. An additional drawback with random priming methods occurs during the PCR amplification of the second strand of cDNA. Following second strand synthesis by

PCR, it becomes impossible to determine which strand of the genome the RNA was transcribed from. In short, random primer reverse transcription is not a viable option for the analysis of bacterial transcriptomics because it results in an inability to identify the strand. Another consequence of random priming is that the majority of RNA sequence reads pile up in the middle of transcripts, diminishing the ability to map small RNAs, promoters, and terminators. Therefore the most logical approach for the generation of an accurate sequencing library is to employ a ligation based strategy that eliminates the need for random priming.

Ligation-based methods for cDNA library construction are strand-specific. An oligonucleotide adapter with a known primer-binding site can be ligated to the 5'-end of the RNA molecules, thus creating DNA-RNA hybrids. The adapter then becomes utilized to prime second strand synthesis by a reverse transcription reaction. Following the creation of a cDNA library, all cDNA is sequenced using one of many available sequencing platforms and sequence data is obtained for subsequent analysis. While cDNA synthesis by ligation is more costly and time intensive, the sequence data can be aligned to the appropriate strand of the genome and the lack of primer binding bias means that the data can be quantified with greater accuracy (45).

Differential RNA-seq

To date, the most impactful contribution to the field of bacterial transcriptomics studies has been the development of the *differential RNA sequencing* (dRNA-seq) methodology by Sharma *et al* in 2010 (48). The objective of any bacterial RNA-seq study is to directly analyze the totality of RNA transcription for a population of cells,

often within a single pure culture under a prescribed physiological condition. Data obtained from whole transcriptome analysis by RNA-seq has allowed investigators to study bacterial gene expression and regulation and improve the annotation of transcriptional features at a rate never before experienced (64). Because RNA-seq results in the direct analysis of RNA rather than by hybridization, transcription features like transcription start sites, untranslated regions (UTRs), and unannotated genes are more readily identified and annotated (65). The criticism remains, that sequence data may not represent the current state of transcription, and instead results from the accumulation of RNA degradation products over the lifespan of the cells. The development of the dRNA-seq method has resolved this dilemma by enriching for and selectively sequencing only *de novo* transcribed, functionally active transcripts (48).

The logic behind the development of the dRNA-seq protocol lies in the nature of the pool of RNA within the bacterial cell. Bacterial RNA is either newly synthesized or undergoing decay. Functionally active transcripts are discriminated from those undergoing degradation based on the phosphorylation status of the 5' end of RNA. RNAs that are either processed or being degraded possess a 5'-monophosphate. On the other hand, functionally active transcripts carry a 5'-triphosphate end that is generated by *de novo* transcription initiation. TEX enriches 5'-triphosphate ends.

As seen in figure 1-1, analysis by dRNA-seq requires the construction of two sequencing libraries originating from the same RNA sample. One library is constructed without alteration as described in the section "*RNA-seq strategies,"* while the other is treated with TEX to degrade 5'-monophosphate containing RNA. Following treatment with TEX the resulting RNA pool will primarily consist of transcripts possessing 5'-

triphosphates. The sample is then treated with tobacco acid pyrophosphatase (TAP) to remove pyrophosphate from the 5' end of all transcripts prior to ligation of sequencing adaptors. Transcripts containing the sequencing adaptor are then poly-A tailed and cDNA is synthesized. The cDNA library is subsequently sequenced at a depth dependent on the size of the organism's genome. As few as 2 million reads is sufficient to annotate the primary transcriptome of the typical prokaryotic organism (64).

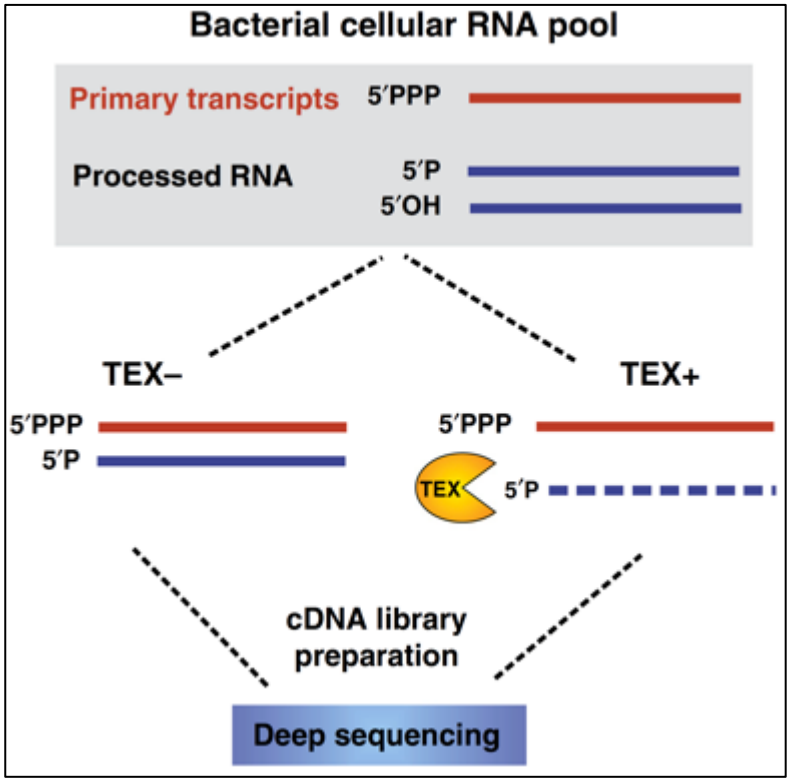


Figure 1-1: Differential RNA sequencing method. 5'-dependent terminator exonuclease (TEX) enriches for functionally active transcripts. The pool of RNA obtained from a bacterial cell consists of primary transcripts with a 5'-triphosphates and processed RNAs with a 5'-monophosphates. To construct dRNA-seq libraries, each RNA sample is divided into two parts. One half is untreated (TEX-), while the other half is treated with TEX (TEX+). TEX specifically degrades RNAs with a 5'-monophosphates, thereby enriching for functionally active transcripts containing a 5'-triphosphate. The TEX treated and untreated samples are converted to a cDNA library and analyzed by high-throughput sequencing. [Courtesy of CM Sharma and J Vogel Current Opinion in Microbiology 2014, 19:97–105]

The resulting sequencing data obtained by dRNA-seq possesses a characteristic pattern that is advantageous when annotating a bacterial transcriptome. As seen in figure 1-2, data obtained from the TEX treated sample contains fewer reads, but the majority of them align to the 5'-end of transcripts. It becomes logical to conclude that dRNA-seq analysis is essential for the discovery of 5'-untranslated regions, promoter location, operon structure, pervasive transcription, and antisense RNAs (66).

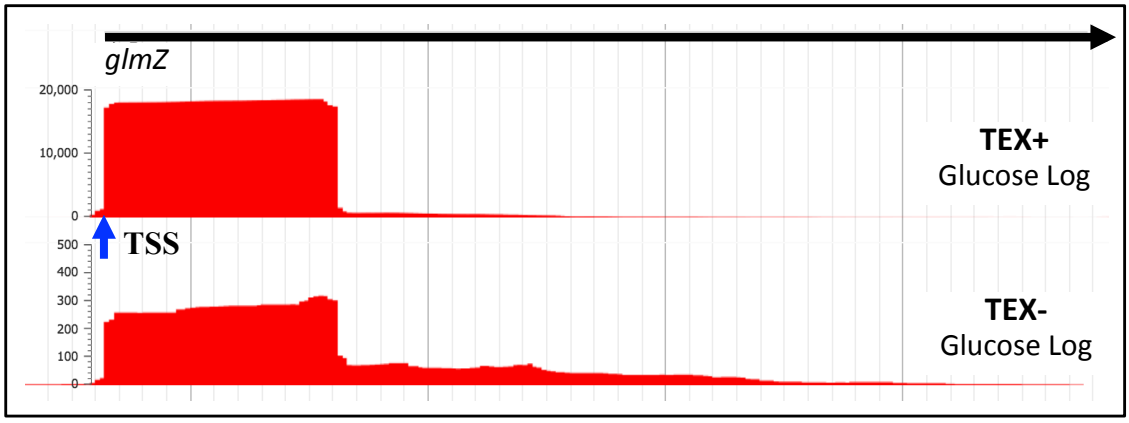


Figure 1-2: Differential RNA sequence data showing enrichment of the 5' transcripts ends associated with the *glmZ* gene. Treatment with TEX (TEX+) results in the accumulation of sequence reads at the 5'-end of the gene, which indicates the location of the transcription start site (TSS; blue arrow). The untreated sample (TEX-) has greater sequence coverage across the entire gene. Note the differences in scale. [Figure obtained from my unpublished data]

Biological Insights Gained from RNA-seq

The advantage of RNA-seq methods over previous technologies like DNA microarray, 5' RACE, and Southern blots remains abundantly clear. However, the true value of any new technology can only be measured by the biological insights gained over previous methods. Over the last five years the microbiology community has utilized RNA-seq and dRNA-seq to investigate biological phenomena at a scale that was not previously possible (67, 68). As a result, high-quality and well-annotated

transcriptomes for pathogenic and non-model bacteria have been frequently published. This has provided a greater understanding of gene expression and regulation across the domain bacteria, and facilitated the characterization of new classes of regulatory RNA in bacteria (69). This is exemplified by a recent study that investigated quorum sensing in *Vibrio cholera* by dRNA-seq and identified 7,240 transcriptional start sites of which 47% were in the antisense direction (70). This highlights the role that antisense RNA may play in the pathogenicity of *Vibrio cholera*. As the sophistication of experimental design employing RNA-seq technology advances, it can be assumed that biological insights will rival those contributed by DNA microarray technology.

In addition to transcriptome annotation, RNA-seq has enabled the investigation of biological hypotheses that were previously unimaginable. For the first time the gene expression of intracellular bacterial pathogens can be analyzed *in vivo*, providing an insight into the physiology of these pathogens that was unobtainable by any other means (15). In a recent study of *Moraxella catarrhalis*, a major nasopharyngeal pathogen of the human respiratory track, researchers investigated the medical observation that *M. catarrhalis* infections are more frequent and severe in the winter (71). In this study, the investigators evaluated the transcriptional response of *M. catarrhalis* to cold-shock. RNA-seq analysis of *M. catarrhalis* grown at 37°C and 26°C (similar to breathing in cold winter air) was conducted and differences in gene expression were analyzed. It was observed that a 26°C cold shock induces the expression of genes related to virulence. Genes involved in high affinity phosphate transport, iron acquisition, and nitrogen metabolism was strongly induced. The

investigators concluded that when exposed to cold shock, *M. catarrhalis* orchestrates a series of adaptive responses that appear to enhance colonization and virulence (71).

The value of RNA-seq methods for the investigation of bacterial transcription can be measured by the volume of publications produced, the value of the knowledge gained, and the variety of biological questions that now can be investigated. Bacterial transcriptome analysis by RNA-seq is still in its infancy, yet insights obtained by this method have made it evident that bacterial gene expression and regulation is more complex than previously assumed. As a scientist this excites me, as it appears that every new RNA-seq study matures into a multitude of novel hypotheses. What follows is my exploration of a series of hypotheses concentrated on providing meaningful biological insights in the area of *E. coli* transcriptomics.

References

1. **Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM.** 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature **452:**872-876.

2. **Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM.** 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods **5:**613-619.

3. **Hao da C, Ge G, Xiao P, Zhang Y, Yang L.** 2011. The first insight into the tissue specific taxus transcriptome via Illumina second generation sequencing. PLoS One **6:**e21220.

4. **Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ.** 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol **10:**599-606.

5. **Buitrago DH, Patnaik SK, Kadota K, Kannisto E, Jones DR, Adusumilli PS.** 2015. Small RNA Sequencing for Profiling MicroRNAs in Long-Term Preserved Formalin-Fixed and Paraffin-Embedded Non-Small Cell Lung Cancer Tumor Specimens. PLoS One **10:**e0121521.

6. **Groves RA, Hagel JM, Zhang Y, Kilpatrick K, Levy A, Marsolais F, Lewinsohn E, Sensen CW, Facchini PJ.** 2015. Transcriptome Profiling of Khat (Catha edulis) and Ephedra sinica Reveals Gene Candidates Potentially Involved in Amphetamine-Type Alkaloid Biosynthesis. PLoS One **10:**e0119701.

7. **Cheng T, Fu B, Wu Y, Long R, Liu C, Xia Q.** 2015. Transcriptome Sequencing and Positive Selected Genes Analysis of Bombyx mandarina. PLoS One **10:**e0122837.

8. **Li G, Zhao Y, Liu Z, Gao C, Yan F, Liu B, Feng J.** 2015. De novo Assembly and Characterization of the Spleen Transcriptome of Common Carp (Cyprinus carpio) Using Illumina Paired-End Sequencing. Fish Shellfish Immunol doi:10.1016/j.fsi.2015.03.014.

9. **Hamburg MA, Collins FS.** 2010. The path to personalized medicine. N Engl J Med **363:**301-304.

10. **Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldon T, Capella-Gutierrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MP, Prosdocimi F, Samaniego JA, et al.** 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science **346:**1320-1331.

11. **Elshahed MS, Najar FZ, Aycock M, Qu C, Roe BA, Krumholz LR.** 2005. Metagenomic analysis of the microbial community at Zodletone Spring (Oklahoma): insights into the genome of a member of the novel candidate division OD1. Appl Environ Microbiol **71:**7598-7602.

12. **Innocenti N, Golumbeanu M, D'Heroue LA, Lacoux C, Bonnin RA, Kennedy SP, Wessner F, Serror P, Bouloc P, Repoila F, Aurell E.** 2015. Whole-genome mapping of 5' RNA ends in bacteria by tagged sequencing: a comprehensive view in Enterococcus faecalis. RNA doi:10.1261/rna.048470.114.

13. **Zoephel J, Randau L.** 2013. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. Biochem Soc Trans **41:**1459-1463.

14. **Forde BM, O'Toole PW.** 2013. Next-generation sequencing technologies and their impact on microbial genomics. Brief Funct Genomics **12:**440-453.

15. **Westermann AJ, Gorski SA, Vogel J.** 2012. Dual RNA-seq of pathogen and host. Nat Rev Microbiol **10:**618-630.

16. **Liebeke M, Lalk M.** 2014. Staphylococcus aureus metabolic response to changing environmental conditions - a metabolomics perspective. Int J Med Microbiol **304:**222-229.

17. **Fimlaid KA, Shen A.** 2015. Diverse mechanisms regulate sporulation sigma factor activity in the Firmicutes. Curr Opin Microbiol **24:**88-95.

18. **Jacob F, Monod J.** 1961. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol **3:**318-356.

19. **Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC.** 1994. Green fluorescent protein as a marker for gene expression. Science **263:**802-805.

20. **Berk AJ, Sharp PA.** 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. Cell **12:**721-732.

21. **Thompson JA, Radonovich MF, Salzman NP.** 1979. Characterization of the 5'-terminal structure of simian virus 40 early mRNA's. J Virol **31:**437-446.

22. **Bensing BA, Meyer BJ, Dunny GM.** 1996. Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of Enterococcus faecalis. Proc Natl Acad Sci U S A **93:**7794-7799.

23. **Schena M, Shalon D, Davis RW, Brown PO.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:**467-470.

24. **Lee C, Roy M.** 2004. Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol **5:**231.

25. **Syvanen AC.** 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet **2:**930-942.

26. **Bulyk ML.** 2006. DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol **17:**422-430.

27. **He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, Criddle C, Zhou J.** 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J **1:**67-77.

28. **Conway T, Schoolnik GK.** 2003. Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Mol Microbiol **47:**879-889.

29. **Southern EM.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol **98:**503-517.

30. **Brown P.** 2009. You say you want a revolution: an interview with Pat Brown. Interview by Jane Gitschier. PLoS Genet **5:**e1000560.

31. **Schena M.** 2003. Microarray analysis. Wiley-Liss, Hoboken, NJ.

32. **Goldmann T, Gonzalez JS.** 2000. DNA-printing: utilization of a standard inkjet printer for the transfer of nucleic acids to solid supports. J Biochem Biophys Methods **42:**105-110.

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

34. **Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR.** 1999. Genome-wide expression profiling in Escherichia coli K-12. Nucleic Acids Res **27:**3821-3835.

35. **Tao H, Bausch C, Richmond C, Blattner FR, Conway T.** 1999. Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. J Bacteriol **181:**6425-6440.

36. **Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, 3rd, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL.** 2006. Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res **34:**1-9.

37. **Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H.** 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2:**2006 0008.

38. **Huerta AM, Salgado H, Thieffry D, Collado-Vides J.** 1998. RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res **26:**55-59.

39. **Karp PD, Riley M, Paley SM, Pelligrini-Toole A.** 1996. EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res **24:**32-39.

40. **Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Labaj PP, Kreil DP, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian HR, Lancashire LJ, Bessarabova M, Nikolsky Y, Furlanello C, Chierici M, Albanese D, Jurman G, Riccadonna S, Filosi M, Visintainer R, Zhang KK, Li J, Hsieh JH, Svoboda DL, Fuscoe JC, Deng Y, Shi L, Paules RS, Auerbach SS, Tong W.** 2014. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol **32:**926-932.

41. **Miyakoshi M, Chao Y, Vogel J.** 2015. Regulatory small RNAs from the 3' regions of bacterial mRNAs. Curr Opin Microbiol **24:**132-139.

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

43. **Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E.** 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. PLoS One **4:**e7526.

44. **Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL.**

2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio **5:**e01442-01414.

45. **Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J.** 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics **13:**734.

46. **Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M.** 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45:**81-94.

47. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:**621-628.

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

49. **Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC.** 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A **109:**E1277-1286.

50. **Ares M.** 2012. Bacterial RNA isolation. Cold Spring Harb Protoc **2012:**1024-1027.

51. **Chen Z, Duan X.** 2011. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol Biol **733:**93-103.

52. **Deutscher MP.** 2006. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic Acids Res **34:**659-666.

53. **Neidhardt FC.** 1964. The regulation RNA synthesis in bacteria. Prog Nucleic Acid Res Mol Biol **3:**145-181.

54. **Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius J, Gutell R, Hogan JJ, Noller HF.** 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res **8:**2275-2293.

55. **Deana A, Celesnik H, Belasco JG.** 2008. The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. Nature **451:**355-358.

56. **Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, Livny J, Earl AM, Gevers D, Ward DV, Nusbaum C, Birren BW, Gnirke A.** 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol **13:**R23.

57. **Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, Grant GR, Hogenesch JB.** 2014. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol **15:**R86.

58. **Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C.** 2003. Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. Genome Res **13:**216-223.

59. **Emory SA, Belasco JG.** 1990. The ompA 5' untranslated RNA segment functions in Escherichia coli as a growth-rate-regulated mRNA stabilizer whose activity is unrelated to translational efficiency. J Bacteriol **172:**4472-4481.

60. **Emory SA, Bouvet P, Belasco JG.** 1992. A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. Genes Dev **6:**135-148.

61. **Hansen MJ, Chen LH, Fejzo ML, Belasco JG.** 1994. The ompA 5' untranslated region impedes a major pathway for mRNA degradation in Escherichia coli. Mol Microbiol **12:**707-716.

62. **Sharma CM, Vogel J.** 2014. Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin Microbiol **19:**97-105.

63. **Creecy JP, Conway T.** 2015. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol **23:**133-140.

64. **Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC.** 2013. An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. Cell Host Microbe **14:**683-695.

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

66. **Wade JT, Grainger DC.** 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol **12:**647-653.

67. **Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R, Young DB.** 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. Cell Rep **5:**1121-1131.

68. **Voigt K, Sharma CM, Mitschke J, Lambrecht SJ, Voss B, Hess WR, Steglich C.** 2014. Comparative transcriptomics of two environmentally relevant cyanobacteria reveals unexpected transcriptome diversity. ISME J **8:**2056-2068.

69. **Wang H, Ayala JC, Benitez JA, Silva AJ.** 2015. RNA-Seq Analysis Identifies New Genes Regulated by the Histone-Like Nucleoid Structuring Protein (H-NS) Affecting Vibrio cholerae Virulence, Stress Response and Chemotaxis. PLoS One **10:**e0118295.

70. **Papenfort K, Forstner KU, Cong JP, Sharma CM, Bassler BL.** 2015. Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. Proc Natl Acad Sci U S A **112:**E766-775.

71. **Spaniol V, Wyder S, Aebi C.** 2013. RNA-Seq-based analysis of the physiologic cold shock-induced changes in Moraxella catarrhalis gene expression. PLoS One **8:**e68298.

Chapter 2: RNA-seq Generated Transcriptome Map of *Escherichia coli K-12* Reveals the Complexity of Bacterial Operon Structure

Chapter Summary

The majority of the content described within chapter 2 was published in the open-access journal mBio on July 8th 2014. As mentioned in the preface, the focus of this chapter was on the methods and data analysis strategies developed to analyze RNA-seq generated transcription data for the mapping of *Escherichia coli K-12* promoters, terminators, antisense RNA, operons etc. It is important to mention that there are a number of notable differences in the content presented here and the original manuscript. Foremost, the original draft of the manuscript was written primarily by Dr. Conway and rightfully remains his intellectual property. I did contribute to the writing and editing of portions of the manuscript, so some of this text appears as it did in published form. However, to avoid infringing upon the work of Dr. Conway and to highlight the contributions that I have made to this project, I have adapted the concepts, ideas, and strategies described in the original publication to describe a parallel narrative that I view as uniquely my own. As such, only the content that I was directly responsible for generating will be discussed within this chapter, and all content that I was not involved with was intentionally omitted. Finally, the narrative of this chapter emphasizes the complexity of bacterial operon structure and presents evidence that point to the need for modernizing the 55-year-old "operon concept". No longer is it accurate to describe all of bacterial transcription as the result of a single promoter driving the transcription of

polycistronic mRNAs. When observed globally, bacterial transcription is best described as unexpectedly complex.

Abstract

We analyzed the transcriptome of Escherichia *coli* K-12 by strand-specific RNA sequencing at single-nucleotide resolution during steady-state (logarithmic-phase) growth and upon entry into stationary phase in glucose minimal medium. To generate high-resolution transcriptome maps, we developed an organizational schema which showed that in practice only three features are required to define operon architecture: the promoter, terminator, and deep RNA sequence read coverage to connect the two ends. We precisely annotated 2,122 promoters and 1,774 terminators, defining 1,510 operons with an average of 1.98 genes per operon. Our analyses revealed an unprecedented view of *E. coli* operon architecture. A large proportion (36%) of operons are complex with internal promoters or terminators that generate multiple transcription units. For 43% of operons, we observed differential expression of polycistronic genes, despite being in the same operons, indicating that E. *coli* operon architecture allows fine-tuning of gene expression. We found that 276 of 370 convergent operons terminate inefficiently, generating complementary 3' transcript ends which overlap on average by 286 nucleotides, and 136 of 388 divergent operons have promoters arranged such that their 5' ends overlap on average by 168 nucleotides. We found 89 antisense transcripts of 397-nucleotide average length, 7 unannotated transcripts within intergenic regions, and 18 sense transcripts that completely overlap operons on the opposite strand. Of 519 overlapping transcripts, 75% correspond to sequences that are highly conserved in *E.*

coli (>50 genomes). Our data extend recent studies showing unexpected transcriptome complexity in several bacteria species and suggest that antisense RNA regulation is widespread.

Importance

The precise location of the 5' and 3' ends of RNA transcripts were mapped across the *E. coli* K-12 genome by using a single- nucleotide analytical approach commonly referred to as differential RNA sequencing (dRNA-seq). The resulting high-resolution transcriptome maps show that approximately one-third of *E. coli* operons are complex, with internal promoters and terminators generating multiple transcription units and allowing differential gene expression within these operons. Extensive antisense transcription was also discovered. Greater than 500 operons, which fully overlap or extensively overlap adjacent divergent or convergent operons. The genomic regions corresponding to these antisense transcripts are highly conserved in *E. coli* (including *Shigella* species), however it remains to be demonstrated whether or not these antisense transcripts are functional. The expansive number of annotated features unearthed by single-nucleotide transcriptome mapping suggest that deeper layers of transcriptional regulation in bacteria exists that are not fully explained by Monod's 'operon concept,' and are likely to be revealed and fully characterized in the future.

Introduction

Escherichia coli emerged as the premier model organism for studying molecular biology in 1961 when Francois Jacob and Jacque Monod described the expression of

the *lac* operon and proposed the operon model as the manner in which genes are regulated in bacteria (1). Since then, *E. coli* has been an integral tool for scientific discovery, and has played a role in research that has resulted in at least ten Nobel prizes (1–10). In addition, the *E. coli* K-12 genome was one of the first full genomes of any organism to be sequenced (11). Of great importance and unique to *E. coli*, biochemical and/or genetic evidence exists for the functions of approximately 75% of its known genes, making it one of the best understood organisms (12), exceeding even humans. Investigation of the genome sequence of *E. coli* confirmed Monod's presumption, that in many instances genes of related function are arranged in operons (13–15). However, the established operon model does not account for the high degree of variability observed within modern transcriptome and proteome datasets (16–20).

Soon after the discovery and characterization of the *lac* operon, it became clear that not all operons are simply transcribed as sets of genes neatly arranged end-to-end on the genome. It was first recognized that regions of phage lambda are transcribed on complementary strands (21). This form of transcription was later termed '*cis*-antisense'. Over the next 50 years, operons were studied individually or as small sets, based on similarity of function. Restricted by the technology available and unable to evaluate bacterial transcription at the global level, the majority of single operon studies supported a simplistic view of the operon. On occasion, indications of transcriptional complexity were documented, such as overlapping, divergent (22, 23) and convergent operons (24, 25), but these occurrences were often regarded as rare idiosyncrasies of transcription and not prevalent or impactful. The scientific community's view of transcriptome complexity was forever changed when it was determined that one or more

antisense transcription start sites (TSSs) are associated with nearly one-half of *Helicobacter pylori* genes (26). The prevalence of antisense transcripts was incontrovertible, and was quickly determined not to be exclusive to *Helicobacter* species. Later discoveries concluded that substantial amounts of antisense transcription also occurs in *E. coli* (27–29).

Some investigators have suggested that the majority of observed antisense transcription is a "by-product" of the transcription machinery, largely because antisense transcripts did not appear to be conserved in enteric bacteria (30). Others hold the alternative view that antisense RNA has an important role in transcriptional regulation (31–36, 18). What is conclusive is that antisense transcription is not a rare occurrence within the domain bacteria, and that not enough is known about the impact antisense transcripts may have on gene regulation and cellular physiology. Recently 316 potentially functional double-stranded RNAs in *E. coli* were identified by antibody binding (laboratory evolved antibody specific for dsRNA) and sequenced in an effort toward resolving this dispute (37). The "excludon concept" of antisense RNA control has emerged as the most plausible means by which divergent operons regulate one another via interaction between overlapping and complementary transcripts (38). A recent study of *Staphylococcus aureus* suggests that antisense transcripts drive RNase III-mediated RNA processing, although a comparison of the antisense RNA content of selected bacteria led the authors to infer that the mechanism is prevalent in Gram positives but absent in Gram negatives (34). It remains unclear if this inference will stand the test of additional scientific evaluation. Due to the increasing amount of evidence for transcriptional complexity in bacteria and the insight that antisense

transcripts are prevalent in bacteria, a comprehensive strategy for the analysis of the *E. coli* transcriptome was developed using RNA- and dRNA-seq.

Current high-throughput sequencing methods for sequencing total RNA offers tremendous resolution power for transcriptome analysis. However, the fullness of its potential has yet to be completely realized for *E. coli*. In all previous studies of the *E. coli* transcriptome the investigators failed to annotate both the 5' and 3' transcript ends, therefore operons were not precisely mapped but rather inferred. Dr. Conway and I therefore developed an organizational structure to precisely map and quantify RNA-seq data across the entirety of operons. This organizational structure centered on annotating operons, which resulted from the identification of both the 5' and 3' transcript ends and sufficient RNA sequence read coverage to connect the ends together. Though others have used tiling microarray technology to study bacterial transcriptome organization (33, 39), tiling microarrays lack the resolving power needed to define transcript ends to the nucleotide or to elucidate operons with multiple promoters. The limitation of tiling microarrays comes from the large size of the probes on the microarray and the inability of the investigator to know where on the probe hybridization took place.

Recent transcriptome mapping studies of *E. coli* have relied on a modified Rapid Amplification of 5' Complementary DNA Ends (5' RACE) protocol followed by high-throughput sequencing to identify TSSs (40, 41). However, critical examination of these data sets has revealed extensive discrepancies that call into question many candidate TSSs, and points to the need for alternative promoter-mapping strategies (42). While conceptually this methodology should accurately identify TSSs to the nucleotide, it appears that in the process of modifying 5' RACE for high-throughput sequencing the

accuracy of the original method was dramatically diminished. One explanation for the lost of accuracy may be the use of random hexamer primers for the amplification of 5' ends. Recent endeavors into RNA-seq analyses of *E. coli* were also unfortunately not designed to map transcript ends accurately. In one study, sequencing library preparation was performed using randomly primed cDNA synthesis (43). While randomly primed library preparation is less expensive and time intensive, it has been well documented that this method causes a bias toward reads in the middle of transcripts and hence the 3' and 5' transcript ends are lost (43). In another study, low sequence read coverage resulted in a resolution of only about 50 nucleotides (44), similar to that of tiling microarray. The recent development of differential RNA sequencing (dRNA-Seq) techniques has dramatically improved the quality and resolving power of RNA-seq data, thereby improving the reliability of identified TSSs annotated by this method. Differential RNA-seq allowed the global mapping of TSSs in *Helicobacter pylori* (26) and *Salmonella enterica* (18, 45); however, the operon architecture of these organisms was not determined because the 3' transcript ends were not mapped. In evaluating the approaches of all four of these studies, it was recognized that the identification of both 5' and 3' transcript ends was essential for the precise mapping of operons and their associated transcriptional regulatory features.

Considering the fundamental role that the operon concept has played in advancing the field of molecular biology, high-resolution RNA-seq analysis of *E. coli* provides the opportunity to investigate transcription on a global level, and with a level of detail unmatched by other forms of global analysis. The advantage of RNA-seq is that it is remarkably precise and provides the investigator with the opportunity to

simultaneously study the regulation of all operons under a single physiological condition. To annotate as many operons as possible and to characterize *E. coli's* response to carbon starvation, we obtained a time series of RNA samples from wild-type *E. coli* K-12 BW38028 cultures grown to stationary phase on chemically defined, glucose-limited minimal medium (46). Logarithmic growth and carbon starvation conditions were selected because they are intrinsic to the physiology that allows *E. coli* to colonize the mammalian intestine and also survive in the environment until encountering a new host and, in the case of *E. coli* pathogens, cause disease (47). Together Dr. Conway and I analyzed all RNA samples by high-throughput sequencing using a strand-specific RNA ligation approach (48). Sequencing library preparation by strand-specific RNA ligation ensured sufficient read coverage (*i.e.* sequencing depth) and precise mapping of both the 5' and 3' transcript ends. In practice, only three transcriptional features were needed to define operon architecture, regardless of its complexity. These are the 5' ends (promoters), the 3' ends (terminators), and sufficient RNA-seq read coverage to connect the ends, which together define operons (Fig. 2-1). Both our RNA-seq and analytical strategies were well suited for obtaining and annotating transcriptome data in an understandable and quantitative manner. Detailed analyses revealed an unprecedented high-resolution view of *E. coli* operon architecture. In addition, the analytical approach employed allowed us to test the hypothesis that bacterial operon structure accommodates substantial transcriptional complexity.

Materials and Methods

Bacterial Strains and Growth Conditions

To annotate operons and characterize their response to carbon starvation, wild-type *E. coli* BW38028 and *E. coli* BW39452 (Δ*rpoS*::*cat*) were grown in 1 liter of morpholinepropanesulfonic acid (MOPS) minimal medium (46) containing 0.2% glucose in a fermenter at 37°C with constant pH and aeration by Dr. Scott Maddox. MOPS medium solutions were modified as described by Wilmes-Riesenberg and Wanner (49), which permits preparation of 40X "M" stock solution and the same final medium chemistry as in the original publication (46). All cultures were sampled at 10 time points during the logarithmic growth phase of *E. coli* BW38028 and at five time points for *E. coli* BW39452, as shown in Fig. 2-S1 in the supplemental material. Logarithmic- and stationary-phase samples were duplicated from replicate cultures.

RNA Sequencing

Total RNA was extracted and purified by Dr. Scott Maddox using an RNeasy kit (Qiagen, USA). In subsequent studies (Chapter 4), the kit-based approach was replaced with hot-phenol extractions because most of he small RNAs in the sample were lost during column purification, and therefore underrepresented in the datasets employed in this study. Biological replicates of logarithmic- and stationary-phase RNA were treated with Terminator 5'-phosphate-dependent exonuclease (TEX) (Epicenter, USA), an enzyme that selectively degrades 5'-monophosphate ends over 5'- triphosphate, to enrich the 5'- triphosphate mRNA fragments for transcription start site mapping. RNA

sequencing libraries (see Table 2-S1) were prepared by Dr. Phillip San Miguel at the Purdue Genomics Facility using a strand-specific, ligation- based approach to SOLiD Total RNA sequencing. Paired-end sequencing was performed using a SOLiD 4 Genetic Analyzer at the Purdue Genomics Facility.

Raw Data Processing

The resulting sequence data were aligned to the *E. coli* MG1655 reference genome (U00096.2) by Dr. Conway, Joe Grissom, and myself using Bowtie version 1.8 (50). In order to maximize the amount of data that aligned to the reference genome, a multiple pass approach was employed. On the first pass, paired-end color space mapping was used with a cutoff distance of 350 bases between read mates. A window of less than 350 bases assured that chimeras of sequences from distant locations were excluded from the analysis. Bowtie parameters were set to include only perfect matches and retained only one alignment where a read mapped to more than one genome location. In practice, it was found that the efficiency of paired-end mapping was between 3 and 10%, meaning that more than 90% of the data sequencing did not align perfectly to the reference genome. To improve the overall alignment, a second and third pass strategy was utilized. The 5'- and 3'-end orphan reads, the data that aligned to the reference genome at one end but not the other, were mapped with Bowtie (one pass for the 5' reads and another pass for the 3' reads). The output of the three passes through Bowtie was three SAM files for each sample. Overall, 40 to 60% mapping efficiency was achieved using the three-pass strategy. The SAMtools (51) utilities were then used to convert SAM files to BAM format and to sort and index them. The binary read

alignment (BAM) files are a binary version of the original SAM file and is substantially smaller yet still readable by most bioinformatics software. The BAM files were displayed in Integrated Genome Viewer (IGV version 2) for primary analysis and quality control.

The BAM files were then converted to base count (WIG) files using an in-house script, written by Joe Grissom, to extract strand-specific base count data (outputs were separated into positive- and negative-strand WIG files). First, the in-house solidbam2wig.pl script read in the paired-end BAM file and counted the nucleotides spanning inserts between the mated 5' and 3' reads. Next, the script brought in the orphan 5' and 3' data from the respective BAM files and incremented the base counts at each base location without duplicating the reads already obtained from the paired-end data. Base count data were then normalized based on the assumption that reads were randomly distributed across the genome and that if sequencing was sufficiently deep, all expressed transcripts would be represented in the data set (43). Another in-house script, normWIG.pl, analyzes the raw WIG files and normalizes based on a simple global normalization approach. The count at each base location was multiplied by 1 billion and the resulting value was divided by the sum of the base counts at all base locations in the file. This normalization strategy is analogous to the total count approach used for normalizing gene-specific read alignments (52). In this way, the base counts are expressed as parts per billion. In practice, SOLiD sequencing did not generate data sets in which the lowest- abundance transcripts were fully covered by contiguous reads. In addition, inefficient ribo-depletion can bias the number of reads that map to non-rRNA genes (53). The normalization strategy employed accounted for both of these factors by

maximizing transcription unit coverage and removing rRNA reads during data processing. For visualization in JBrowse (54), the normalized WIG files were converted to BIGWIG files again using SAMtools software (51). Dr. Conway and I analyzed all of the RNA-seq data manually using a graphic user interface linking JBrowse (54) for data visualization to an Oracle database for recording the annotation data.

Nucleotide Sequence Accession Number

RNA sequencing data and curated results were deposited at Gene Expression Omnibus, accession no. GSE52059. We offer our annotated *E. coli* K-12 operon map as a community resource upon which others can participate in annotating additional transcriptional regulatory features.

Results and Discussion

Single-nucleotide Resolved RNA-seq Data Sets

Escherichia coli K-12 has served as an important model organism for molecular biology for more than a 50 years and was the first bacterium analyzed by DNA microarray technology (55, 56), making it a logical chose for RNA-seq analysis. While several other bacteria now have been analyzed by RNA-seq (26, 31, 33, 35, 45, 57–59), the limited number of RNA-seq studies performed on *E. coli* have not provided the quality of data needed to make meaningful conclusions about global transcription (43, 44). As mention previously, a strand-specific RNA ligation-based RNA-seq strategy was used, in tandem with a robust analytical approach, allowed for transcriptional

features to be define across the entirety of the *E. coli* genome at single-nucleotide resolution. RNA samples from a time series was obtained on duplicate cultures of *E. coli* K-12 BW38028 and its isogenic *rpoS* mutant BW39452 during logarithmic- and stationary-phase growth on glucose- limited minimal medium (see Fig. 2-S1 in the supplemental material). In total, 26 RNA samples were sequenced to generate a data set of 72.1 million uniquely mapped sequence reads corresponding to more than 5.5 gigabases of RNA-seq data (see Table 2-S1 in the supplemental material). As a method for verifying that the time series samples were collected at the correct time points, the temporal expression of *bolA*, a known glucose starvation-inducible gene (60), was analyzed. It was confirmed that the RNA-seq data obtained from the time series correctly represented the growth conditions described in the supplemental methods (Fig. 2-1). The correlation between replicate cultures was greater than 0.96 (see Fig. 2-S1), so it was concluded that this level of biological replication provided a reliable view of the *E. coli* K-12 transcriptome under logarithmic growth and carbon starvation physiological conditions (Fig. 2-2).

An in-house computational tool was developed to convert the binary read alignment (BAM) files to base count (WIG) files to facilitate single-nucleotide resolution analyses. While similar tools are now widely available (61–63), during the initial stages of this study no such algorithm was in place that faithful performed this conversion. Base count data were normalized using a strategy analogous to the total count approach (52) for normalizing gene-specific read alignments. Normalization of RNA-seq data continues to be an area of great debate. For this study the total count approach was selected because of its simplicity and precedence originating from

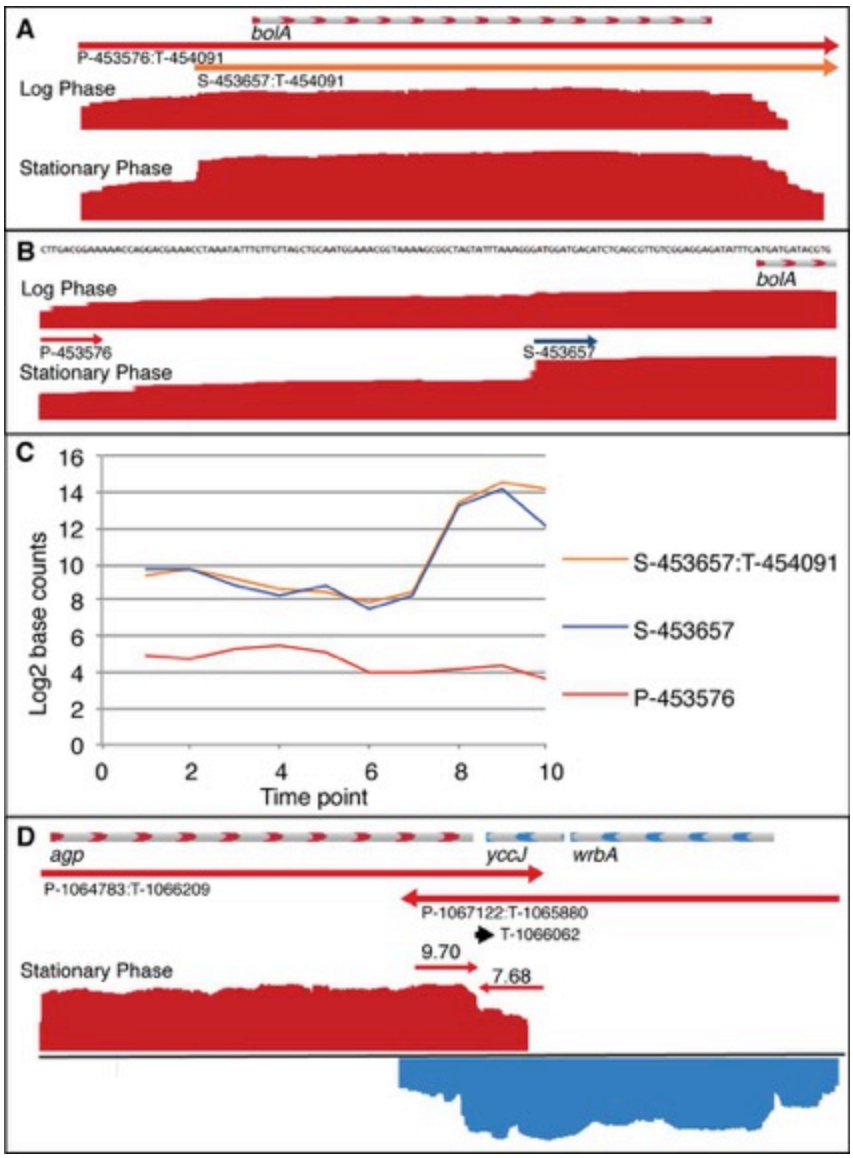


Figure 2-1: Single-nucleotide resolution of promoters and terminators in example complex operons. (A) The *bolA* operon contains transcription units (TUs) P-453657:T-454091 (red arrow) and S-453688:T-454091 (orange arrow). RNA-Seq data are shown in a JBrowse visualization of positive-strand (red) transcription in logarithmic- and stationary-phase samples (average from three replicates). The base count data were normalized and log2 transformed such that track heights in JBrowse are directly comparable. (B) *bolA* promoter region showing primary promoter P-453576 and secondary promoter S-453658 at single-nucleotide resolution (drawn to scale). (C) Plot of promoter strength (average count of 10 bases beginning at TSS) and TU usage (avg. count of bases within TU) for 10 growth curve time points showing *bolA* induction upon entry into stationary phase (see Fig. 2-S1 for growth curve). (D) Terminator efficiency (avg. counts of 10 bases preceding and following terminator) is shown for T-1066062, which is shared by converging operons *agp* on positive strand (red) and *wrbA-yccJ* on negative strand (blue). (Adopted T. Conway, 2014)

microarray analysis. Following conversion and normalization, all of the resulting WIG files contained only the base location, obtained from the reference genome, and the number of times each base was sequenced. Importantly, all WIG files were more than 100 times smaller than the corresponding read alignment (SAM) files. Advantages of this simple base count approach are several-fold: first, the data are inherently more computable; second, normalization of base count data makes all samples directly comparable and eliminates transcription unit length bias; third, the base counts of individual features can be computed and queried at any desired resolution from single nucleotide to an entire operon.

It is my opinion that the greatest advance in transcriptome research is the ease by which RNA-seq reads can be digitized and computed upon, to produce measurable and quantifiable data. Because all analysis of the RNA-seq data was performed on base counts, the normalized base count values for any region of the transcriptome could be easily averaged across any range of bases to calculate the relative activity of transcriptional features, including promoters, terminators, transcription units, and operons (Fig. 2-1). The number of bases used to calculate promoter strength were empirically determined by comparing the single base count value at the transcription start site to the 3-, 5-, 10-, and 20-base averages, each starting at the transcription start site obtained from the sequencing data. When evaluated, the shorter base count lengths were highly variable, presumably because of single base variability at the start locations that are occasionally observed in primer extension experiments (64) and were frequently observed in the RNA-seq data sets presented here. On the other hand, the average of 20-base-count length was too long to allow discrimination of closely spaced

promoters. It was therefore determined that the use of 10-base average counts for quantifying promoter strength was best suited from subsequent analysis (Fig. 2-1). The same 10-base average was empirically determined to be best suited for calculating the efficiency of terminators by comparing the 10-base average counts before and after the manually annotated termination site (Fig. 2-1 and 2-3). These average base count values were used to calculate the activity of individual transcription features no matter their location in a given operon. In addition, the same average base count strategy was used to quantify the impact of operon structure on relative transcription unit and gene expression.

Promoter Mapping

Essential to the annotation of operons on a transcriptome is the identification of promoters corresponding to mapped transcription start sites. The search for promoters was driven by the manual mapping of putative transcription start sites on the basis of three criteria: (i) sequencing read enrichment facilitated by terminator exonuclease (TEX); (ii) promoter motif analysis; and (iii) consensus among replicate data sets. The three criteria listed above were important for promoter identification because: (a) treatment with TEX preferentially degrades RNA molecules with 5'-monophosphate ends and enriches mRNA with 5'- triphosphate ends corresponding to the nucleotide initiated de novo by RNA polymerase; (b) promoter sequencing motif analysis; and (c) repetition between datasets instills confidence that the putative promoter is correctly associated with the physiological conditions being investigated. None of these approaches alone are comprehensive, and each can gives rise to false-positive results or

fail to find legitimate transcription start sites (25). An example of this is TEX treatment; not all transcription start sites enrich when treated with TEX. In some instances RNA 5'-pyrophosphohydrolase activity removes the 5'-triphosphates from newly synthesized RNA, and as such these transcripts are not enriched (65). Additionally, not all promoters have a prototypical consensus motif that can be identified by computer algorithms (66), and promoter motif searches are prone to reporting false positives (67).

To facilitate accurate mapping of promoters and hasten the process, an algorithm was written to search and report only changes in base count values that exceeded 2-fold. Using this algorithm, all of the minor variations observed in the data were rendered nominal, and only the pronounced transcription start sites were indicated. The transcription start sites of highly expressed genes were apparent in all 14 replicates (n = 14, wild-type and *rpoS* culture samples from logarithmic and stationary phase). However, since the 14 samples represented logarithmic- and stationary-phase samples, expression of some promoters and their respective transcription start sites were observed to be condition specific. Proper consideration was given to these condition specific transcription start sites, and all were included for downstream analysis. In order to generate a transcriptome map that was condition independent by which annotating the response to multiple conditions could be accomplished in the future, consensus of only three replicates, of either logarithmic- or stationary- phase samples was considered significant. This strategy revealed 11,329 putative transcription start sites, a finding that is similar to the number of promoters found in a recent study by Thomason and Storz (68), and includes known promoters of even weakly expressed genes. This value exceeds the expected promoter density on the *E. coli* genome, thus exemplifying the

need to use a multifaceted approach to confirm promoters. While many scientists who study bacterial transcriptomics are not surprised by the large number of transcription start sites discovered in *E. coli,* it is worth pointing out that ours was the first study to produce measurable evidence that supports what many have assumed, "that bacterial transcription is more complex than previously postulated."

Next, I used a bioinformatics approach to search for known promoter motifs within the 50-nucleotide sequences immediately upstream of the 12,583 putative transcription start sites using Find Individual Motif Occurrences (FIMO) software (69). To screen these 50-nucleotide sequences, a library of known *E. coli* promoter motifs was assembled using the resources at DPInteract (70). I found it was necessary to modify the RpoD promoter library according to the characterization of 554 promoters by Mitchell et al. (71), which demonstrated that the RpoD consensus promoter has -10 and -35 regions with spacing of 14 to 20 bases between promoter elements. The search output was restricted to promoter sequences correctly positioned within ±3 bases of the transcription start site, with E-values corresponding to P values of <0.02. This three-facet approach of enrichment, consensus, and promoter motif searching resulted in the locating of 5,653 putative RpoD-dependent promoters, which were evaluated further by direct visual observation and manual annotation.

A JBrowse (54) visual graphic environment interface was used to interact with and write annotation data to an Oracle database, which facilitated the documentation of the transcriptome. From the list of candidate promoters obtained using the strategy described above, a JBrowse track was created at the corresponding base locations along the reference genome, each displayed a "clickable" URL call to the database that

automatically recorded the base location and allowed manual entry of metadata, including the type of promoter, regulatory information supported by differential expression analysis, and comments. Only promoters that could be experimentally associated with operons were annotated, by using RNA-seq data as described in the next section. This strategy reduced the number of putative promoters from 5,653 to 2,122 (Fig. 2-2), which more than doubled the 811 individually characterized *E. coli* promoters annotated and cataloged at RegulonDB. In addition, it calls into question the several thousand candidate promoters that were identified by less reliable high-throughput strategies (39, 42). The promoter data set was dominated by primary promoters, defined as the furthest upstream promoter in an operon (66.3%), with significantly fewer promoters falling into alternative categorizes: secondary promoters that were intergenic and downstream of primary promoters (19.6%), internal promoters that were intragenic (9.8%), and finally antisense promoters (4.2%). (see Table S2 in T. Conway, 2014) Upon further evaluation of promoter type, it was determined that all possible arrangements and orientations exist, and no discernable pattern was determined. Collectively, this high degree of promoter variation within operons generates extensive complexity within the *E. coli* transcriptome (Fig. 2-2).

It is well known that promoter strength, i.e., quality, varies greatly from promoter to promoter (71), and that variability is reflected in the transcriptome data presented here as well. In an effort to quantify promoter strength, we scored the three criteria (metrics) used to map candidate promoters. The promoter strength score was calculated by applying a weighted matrix on the basis of a 10 points scale, where TEX enrichment was assigned a weight of 5, the promoter motif score carries a weight of 3,

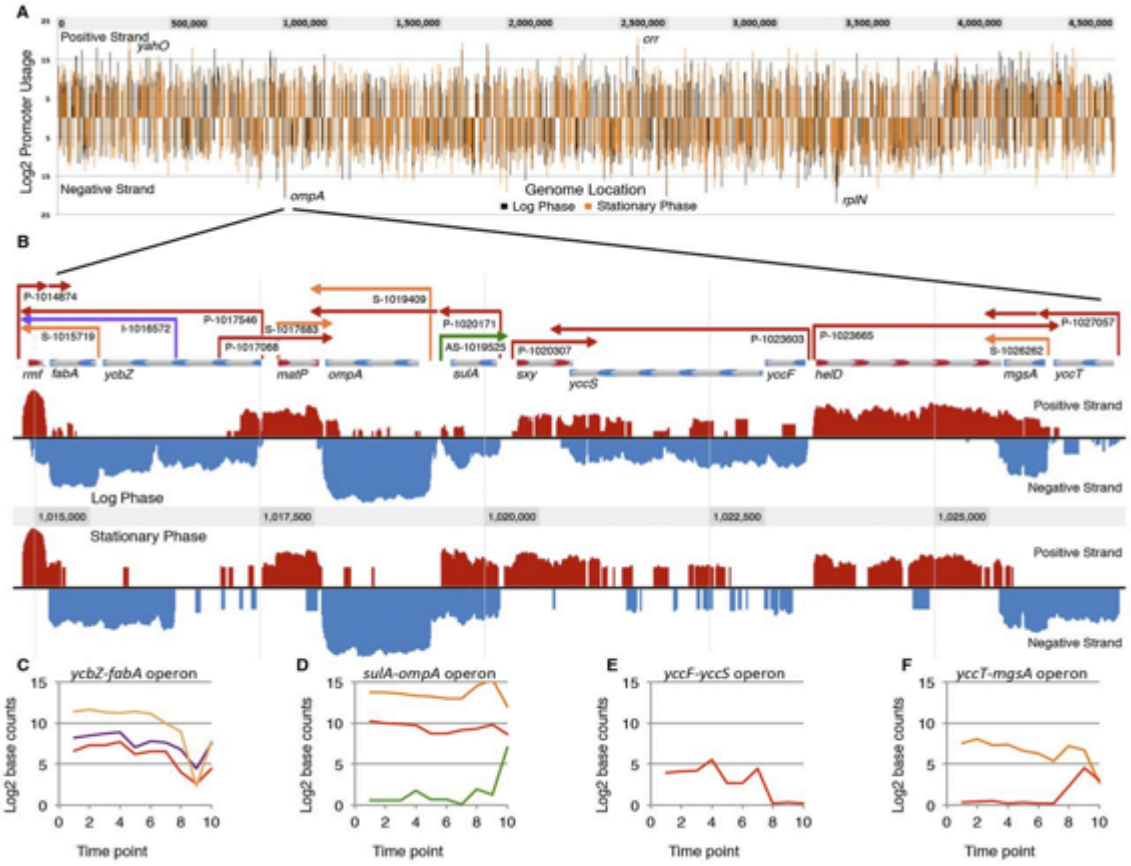


Figure 2-2: Genome-wide promoter locations and annotated transcriptome map of a selected region. (A) Promoters aligned by genome location. Line heights correspond to normalized, TEX-enriched promoter usage values, shown for logarithmic phase (black) and stationary phase (orange). (B) Annotated regulatory features of a selected region of the genome. Positive-strand RNA-Seq data (red) and negative-strand data (blue) were normalized for comparison between logarithmic- and stationary-phase samples. Primary promoters and corresponding TUs (red) are indicated by arrows extending from promoter to terminator, as are secondary promoters (orange), internal promoters (purple), and AS promoters (green). Beginning on the left, *rmf* is transcribed from a primary promoter and depending on growth conditions terminates either before or within the *ycbZ-fabA* operon, which has a primary promoter upstream of *ycbZ*, an internal promoter within *ycbZ*, and a secondary promoter upstream of *fabA*. *matP* is transcribed from primary and secondary promoters. *ompA* is transcribed from a secondary promoter in log phase and is cotranscribed from the primary promoter of the *sulA-ompA* operon during stationary phase. An AS TU that overlaps the *sulA* sense transcript is turned on in stationary phase. The *sxy* and *yccF-yccS* operons converge. Finally, *mgsA* is transcribed as an independent TU from a secondary promoter in log phase and also is expressed in the *yccT-mgsA* operon from a promoter that is active only in stationary phase. (C) Plot of TU base counts for *ycbZ-fabA* operon, colorized according to color scheme in panel B; (D) TU plot of *sulA-ompA* operon; (E) TU plot of *yccFS* operon; (F) TU plot of *yccT-mgsA* operon. (Adopted from T. Conway, 2014)

and the transcription start site consensus (between replicates) score carries a weight of 2. It should be understood that these were assigned based on my examination of the data and the greatest weight was given to the data that best supported the identification of previously characterized promoters (e.g., *bolA*). The resulting analyses yielded promoters scored on a scale of 0 to 10. The TEX enrichment metric reflects the number of instances among four TEX replicates in which the ratio of TEX-treated versus non-TEX-treated base counts (10- base-count average beginning at the transcription start site) for a sample exceeded 2-fold. The promoter motif scores obtained from FIMO analysis were calculated by dividing the entire data set into quartiles of E-values for RpoD-dependent promoter motifs. The final metric, the transcription start site consensus score was calculated as the number of occurrences of a transcription start site at a precise base location divided by the total number of samples evaluated (n = 14). The 2,122 promoters ranged in score from 10 to 0.14, with the top 10% of promoters scoring above 7.8, the bottom 10% scoring below 2.9, and the average promoter scoring 5.5. This wide variation in promoter score highlight the dynamic nature of bacterial transcription, and supports the hypothesis that abundance of a transcript can be scaled up or down based on the strength of the upstream promoter region.

It is important to note that I found no strong correlation between promoter usage (average count of first 10 transcribed bases after a transcription start site) and promoter confidence scores or promoter motif scores (see supplemental material Fig. 2-S2), which stands in agreement with an earlier report (71), but in conflict with intuition. However, a weak correlation between promoter usage and transcription unit usage (average count of bases from promoter to terminator) was observed (see supplemental

material Fig. 2-S2). It was also confirmed that transcription unit usage and RNA half-life (72) (measured under similar conditions) did not correspond, again as noted previously (ref needed). Nevertheless, promoter and transcription unit usage values do appear to reflect the physiologically relevant transcript level at a given time point, because the RNA concentration in the cell is determined both by the frequency of transcription initiation and the rate of RNA decay, which vary substantially for different transcripts (72). In short, even though the abundance of a given transcript could not be directly explained by any of the promoter metrics evaluated, it was possible to view biologically signification changes in transcript abundance as the culture transitioned from logarithmic- to stationary- phase.

Operon Mapping

To annotate operons, it was also necessary to map the 3' ends of transcripts. Doing so allowed documentation of relationships between promoters and the corresponding downstream terminators (Fig. 2-1). The criteria for operon annotation employed for this analysis were (i) the primary promoter must be followed by sufficient sequence read coverage across the entire operon, (ii) the mapped 3' ends must extend beyond the stop codon of the last gene in the operon, (iii) downstream of secondary or internal promoters there must be a corresponding increase in sequence reads in the coverage sample, and (iv) internal terminators must result in the decline of sequencing data in the coverage sample for downstream bases without interrupting contiguous coverage by read-through transcripts. Analysis of 3' transcript ends that could be associated with an annotated promoter(s) led to the mapping of 1,774 candidate

terminators (see Table S3 in T. Conway, 2014). Of these terminators, 264 were located within operons and permitted partial read-through transcription of downstream genes, as demonstrated for the *sdh*CDAB-*suc*ABCD operon (Fig. 2-3). The 1,774 putative terminators were evaluated by using the TransTermHP software (73). TransTermHP analysis confirmed that 623 (35%) putative terminators had sequence characteristics indicative of intrinsic (Rho-independednt) terminators. This extends the number of annotated *E. coli* terminators previously annotated, 227 (42), by nearly 8-fold. Alternatively, it has been predicted that about one-half of terminators are intrinsic (74). The remaining 1,151 terminators that were not confirmed by TransTermHP are therefore candidates for terminators that require wither Rho or another protein factor for termination. The data in Table S3 (see Conway, 2014 in Appendix A) represent one of the most extensive genome-wide predictions of nonintrinsic terminators in *E. coli*.

The analyses discussed to this point consisted of only logarithmic- and stationary- phase samples and revealed a total of 6,463 regulatory features, including 2,122 promoters, 1,774 terminators, and 2,566 transcription units corresponding to 1,510 operons. The mapped reads from the sequencing data obtained for this project covered more than 90% of bases on the *E. coli* genome, and 90% of these reads were mapped to an annotated operons. The 1,510 operons cover 2,985 of 4,457 known *E. coli* genes (67%) annotated on the reference genome. As more datasets from different growth conditions are analyzed, the simple organizational schema described above will

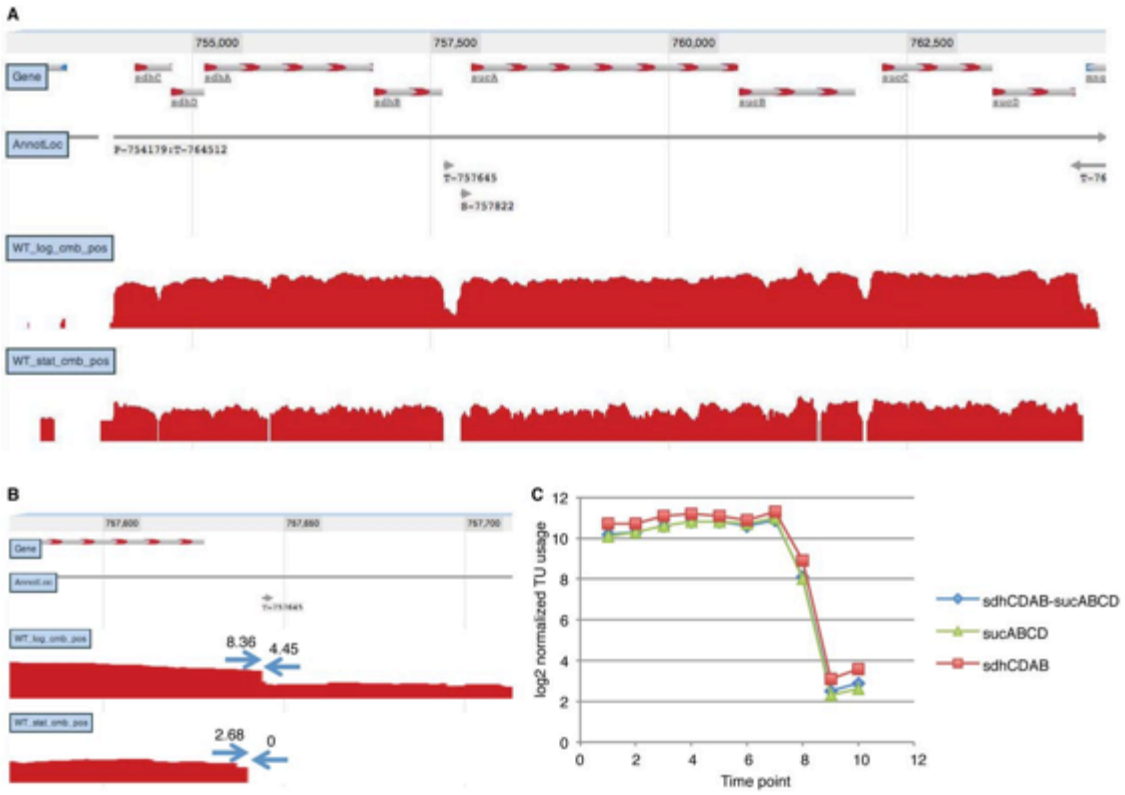


Figure 2-3: Balanced transcript coverage of the *sdhCDAB-sucABCD* operon achieved by complex interaction of internal terminator and secondary promoter. (A) JBrowse instance showing coverage data; (B) terminator usage in logarithmic (WT_log_cmb_pos) and stationary (WT_stat_cmb_pos) phase; (C) TU coverage time series. (Adopted from T.Conway, 2014)

be able to accommodate the addition of newly identified regulatory features to the *E. coli* K-12 transcriptome map. For ease of use by the scientific community, all annotation calls made on the data sets were converted to GenBank format using the terms "promoter," "terminator," and "operon" as feature keys (75). Converting the data into this format should allow annotation of any number of experimental parameters that affect the usage of these features. The entirety of the *E. coli* K-12 transcriptome annotation and the GenBank feature table discussed here can be obtained from the Gene Expression Omnibus (accession no. GSE52059) at the National Center for Biotechnology Information.

Operon organization examples

The data presented in Fig. 2-2 unequivocally confirm that the *E. coli* genome is organized in operons. However, the complexity of these operons varies dramatically from Monod's original conception of the operon consisting of a regulatory region with a single promoter. This single promoter was found to initiate transcription of a polycistronic mRNA covering all of the genes that make up the *lac* operon, ending at a single terminator. Indeed, many *E. coli* operons fit this model or are even simpler (64%), contain a single gene. Analyzed in its entirety, the *E. coli* transcriptome reveals densely packed regulatory features that could not have been discerned from the nucleotide sequence of the genome alone (Fig. 2-2). Complex operons accounted for 36% of annotated operons. Complex operons result from transcripts originating from secondary and internal promoters, as well as internal terminators. An example of this complexity is the *sulA* and *ompA* region of the genome. During logarithmic phase these genes are independently transcribed, with each gene having its own promoter and terminator. However, during stationary phase, the *sulA* transcription unit reads through a nonintrinsic *sulA* terminator to form a *sulA-ompA* transcript, driven by a secondary promoter that increases expression of the *ompA* transcription unit (Fig. 2-2). While there is no reason to think that the proteins produced by the *sulA* and *ompA* genes, cell-division inhibitor and outer membrane protein respectively, have any interaction with one another once translated, it is apparent from the transcriptome data that transcription regulation of this operon takes place at two different promoters. In addition, an antisense transcript that fully overlaps the 510-nucleotide *sulA* coding sequence is also turned on in stationary phase. This arrangement of the *sulA-ompA* operon and antisense

transcript was postulated as a means for posttranscriptional control of the synthesis of the cell division inhibitor SulA (76), which is further supported by our results showing differential expression of the antisense transcript. Our organizational schema makes the previously unannotated *sulA* antisense transcript and similar regulatory features readily apparent on the *sulA-ompA* transcriptome map (Fig. 2-2). Such differential expression of transcription units within operons can provide bacteria with the ability to modulate gene expression to cope with physiological complexity (33, 34, 38, 45).

Notably, Fig. 2-2 reveals the *E. coli* transcriptome for only two growth conditions, logarithmic- and stationary- phase due to carbon source limitation. Analyses of the data presented here showed that 29% of operons have more than one promoter, and 15% of operons have more than one terminator under these conditions (Fig. 2-4). Further, many operons were subject to multiple regulatory inputs (42) resulting from multiple promoters and terminators within a single operon. Adding additional complexity, differential mRNA decay has been shown to contribute to an additional layer of control within operons (72). No doubt, future RNA-seq studies on *E. coli* for the myriad of responses to numerous regulatory signals will likely reveal substantially more variation in operon architecture, as seen for *Salmonella* (45).

The intricacy of operons with internal promoters and terminators is readily apparent. An example of this is three promoters upstream of the *ahp*CF operon that contribute to the expression of the operon in an additive fashion (Fig. 2-5). This arrangement of promoters permits differential control of alkylhydroperoxidase production in response to stationary phase, osmotic stress, and oxidative stress (77). Likewise, three promoters hat contribute to *ybfE-fldA-uof-fur* operon expression during

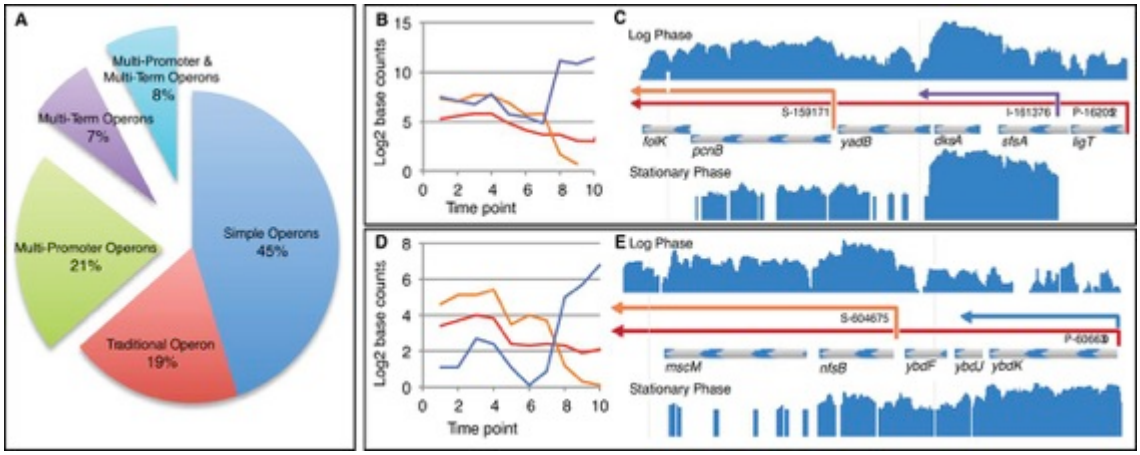


Figure 2-4: Computational analysis of single-nucleotide resolution data reveals complex operon architecture. (A) Operons organized by increasing complexity; (B) TU usage plot of ligT-sfsA-dksA-yadB-pcnB-floK operon. The primary TU corresponding to the entire operon is shown in red. The differentially expressed dksA-specific TU driven by promoter I-161376 is shown in purple. The pcnB-folK TU driven by S-159171 is shown in orange. Note that transcript levels of dksA increase upon entry into stationary phase, whereas pcnB-folK decreases. (C) JBrowse instance showing ligT-sfsA-dksA-yadB-pcnB-floK operon; (D) TU usage plot of ybdK-ybdJ-ybdF-nrsB-mbcM operon. Note the primary TU corresponding to the entire operon (red) decreases only slightly during transition from logarithmic phase into stationary phase, because it is comprised of two differentially expressed TUs, one of which increases and the other decreases during growth: the nfsB-mbcM-specific transcript (orange) essentially disappears in stationary phase, whereas the ybdK-specific transcript (blue) is induced in stationary phase. (E) JBrowse instance of ybdK-ybdJ-ybdF-nrsB-mbcM operon. (Adopted from T. Conway, 2014)

logarithmic phase, allowing for continuation of the *uof-fur* transcription unit expression, decline of *fldA* expression, and completely turning off the expression of *ybfE* in stationary phase (Fig. 2-5). Although cotranscription of the complex *ybfE-fldA-uof-fur* operon was not previously recognized (78), it is reasonable to think that *uof-fur* should be transcribed independently of *ybfE-fldA* under certain conditions, because *fur* encodes a negative regulator of genes for iron uptake. Furthermore, *uof* expression is controlled indirectly by the trans-acting noncoding RNA RhyB, which is itself Fur regulated, thus forming a negative feedback loop in responsive to iron limitation (78).

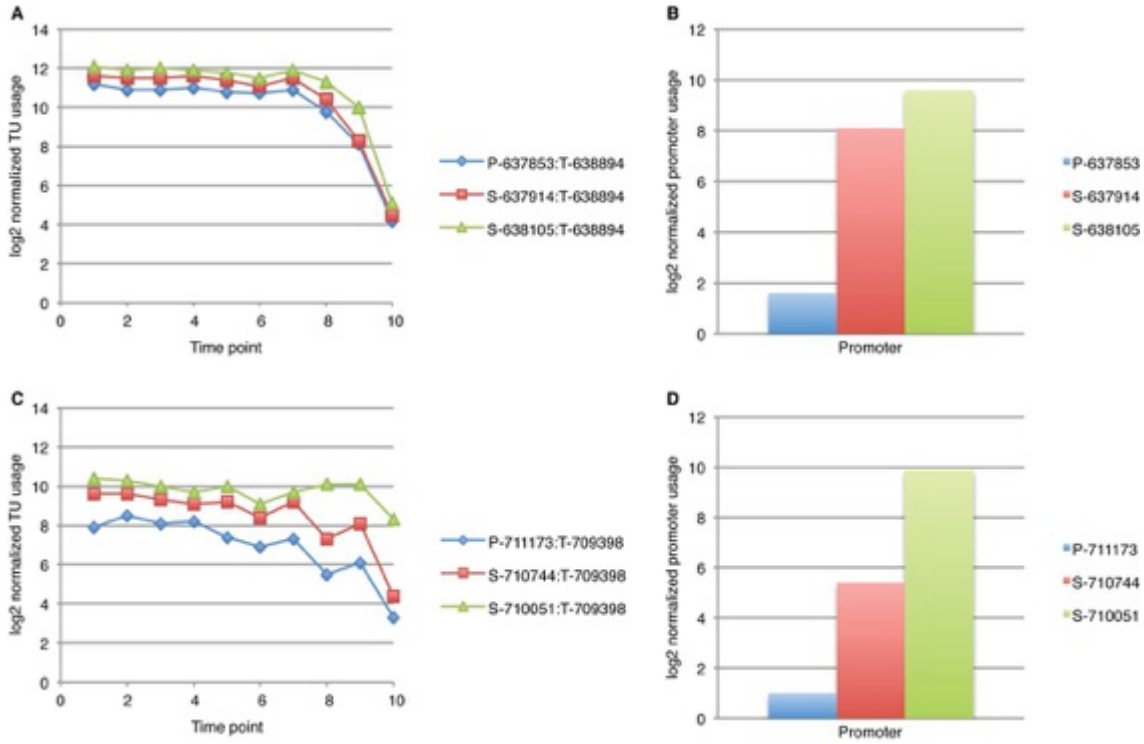


Figure 2-5: Three promoters contribute to expression levels of genes within the ahpCF and the ybfE-fldA-uof-fur operons. (A) WT time series of TU base counts of three overlapping TUs within the ahpCF operon; (B) usage of 3 ahpC promoters (10-base average from TSS +1 to +10) during logarithmic phase (time point 4); (C) TU coverage time series of the ybfE-fldA-uof-fur operon; (D) differential usage of three promoters within the ybfE-fldA-uof-fur operon during log phase. Promoter usage and TU coverage calculations are described in the legend to Fig. 1. (Adopted by T. Conway, 2014)

It is also possible to unravel condition-specific terminator usage by our organizational schema, as illustrated for the internal terminator of the *sdh*CDAB-*suc*ABCD operon. The *sdh*CDAB-*suc*ABCD operon encodes for three enzymes of the tricarboxylic acid cycle (Fig. 2-3). The arrangement of this operon explains how intrinsic termination permits one operon to function independently as two operons under appropriate conditions (79), yet behave as a single operon under other conditions. The examples presented here demonstrate how promoter and terminator activity

calculations can be employed to reveal new biological insights, even from well-understood regions of the genome, from the RNA-seq transcriptome analyses.

Cataloguing Operon Architecture

High-resolution transcriptome mapping of well-characterized regions of the *E. coli* genome has provided glimpses of the intricacies of operon arrangements (Fig. 2-2 to 2-5). Analyses of *E. coli* operons at single-nucleotide resolution revealed numerous instances of transcription complexity throughout the genome. Single-gene operons with a single promoter and terminator make up 45% of all operons, while 19% were classified as "traditional" operons possessing multiple genes and a single promoter and terminator (Fig. 2-4). The remaining operons (36%) were more complex: 21% had multiple promoters (maximum observed was eight), 7% had multiple terminators (as maximum observed was four), and 8% had both multiple promoters and multiple terminators. On average a given operon contains 1.98 genes. The most complex operon observed encoded for genes essential for several core cellular functions, and had eight promoters and four terminators covering fourteen genes, and produced twenty-three biologically relevant transcription units (*yjeF-yjeE-amiB-mutL-miaA-hfq-hflX-hflK-hflC-yjeT-purA-nsrR-rnr-rlmB* operon; see Fig. 2-S3).

Differential transcription unit expression within a given operon can result from the relative activity of secondary and internal promoters, internal terminators, and combinations of these regulatory features. An example of this can be seen in Fig. 2-4, which illustrates how it is possible for an internal promoter and internal terminator to function together to increase the expression of the DksA-specific transcription unit in

stationary phase. The *ybdK* operon, also show in Fig. 2-4, illustrates how differential expression is possible at the 5' and 3' ends of the same operon. This is caused by transcription from a secondary promoter and an internal terminator. The arrangement of these features results in a complete inversion in expression of the two transcription units between logarithmic- and stationary- phases. These findings support the hypothesis that operon architecture permits *E. coli* to adjust relative levels of gene expression within the same operon in response to changes in environmental conditions.

In an effort to quantify differential gene expression within *E. coli* operons, I compared the base counts of transcription units within the same operon under the same growth condition and tabulated the complexity that arises from internal promoters and terminators (data can be seen in T. Conway, 2014). Of the 548 complex operons containing multiple transcription units due to having multiple promoters or terminators (Fig. 2-4), 327 displayed more than 2-fold change in differential expression of one transcription unit compared to other transcription units within the same operon. For the 633 operons that contained more than one gene, a 2-fold or greater change in differential gene expression was observed for 315 of these operons (*e.g.*, see Fig. 2-4). In the instances where polycistronic operons possessed only a single promoter and terminator, it appears that differential decay of the processed transcripts was responsible for the observed variation in gene expression. In total, 43% (642 of 1,510) of all *E. coli* operons displayed a complex gene expression regulatory pattern. Clearly, differential expression of transcription units and genes within the same operon is common in *E. coli*, and worthy of inclusion in a modern model of the bacterial operon.

My analyses of the transcriptome of *E. coli* provided the opportunity to map potential antisense transcription events across the entirety of the genome. In many cases, antisense transcripts completely overlap and are complementary to sense strand transcripts that encode proteins; however, these antisense transcripts do not appear to encode proteins, due to the lack of an open reading frame. For example, the long antisense RNA that is complementary to the *sulA* gene does not appear to be translated, because it has no properly positioned ribosome binding site nearby a start codon, and therefore most likely is an emerging class of regulatory RNA called long noncoding RNA (lncRNA). We found eighteen transcripts either for annotated protein-coding genes or small RNAs that completely overlap operons transcribed in the opposite direction. As a result of this arrangement, the eighteen corresponding operons contain lncRNA transcripts that overlap the coding sequences on the opposite strand.

Since genome annotation relies heavily on identification of coding sequences, it was predicted that the transcriptome analysis described here would reveal a number of unannotated genes. Indeed, 96 novel transcripts that do not correspond to genes on the reference genome and were previously unannotated in *E. coli* K-12 were identified. These 96 novel transcripts include 89 antisense transcripts that have an average length of 397 bases, with the longest being 1,168 bases. The remaining seven transcripts are completely intergenic and do not overlap annotated genes. None of the 96 transcripts appear to code for protein because they all have multiple stop codons within all three reading frames. Of the 89 antisense transcripts, 21 are convergent with known operons that contain genes that encode proteins, seven are divergent with mapped operons, and 40 completely overlap annotated operons. The remaining 21 antisense transcripts

overlap known genes that could not be annotated into operons by the RNA-seq data presented here. The genomic regions corresponding to 72% of these long-noncoding RNAs are highly conserved in >50 *E. coli* and *Shigella* genomes. It was proposed previously that bacterial long-noncoding RNAs might be functional (34, 38), yet this is still questioned by others (30). Similar long-noncoding RNAs have also been found in eukaryotes, and although they are not well understood, they are thought to play a role in regulating gene expression (80). To date, the entirety of what is knowledge concerning long-noncoding RNAs is exclusive to the domain Eukarya. In eukaryotes long-noncoding RNAs are thought to be responsible for shaping the structure of the DNA within the nucleus and regulating its dynamic movement. Until recently it was presumed that prokaryotes lacked this level of gene regulation, once more prescribing to the concept that prokaryotic organisms are too simplistic to have such a sophisticated regulatory mechanism.

A recent study of terminator efficiency showed that only 3% of *E. coli* terminators are "strong" (81). In the context of this paper "strong" was defined as maximally efficient and completely turning off transcription. Inefficient termination however is very common, and would explain how convergent operons sometime result in the production of overlapping transcription (24, 25). I therefore hypothesized that these partial termination events between convergent operons would generate complementary 3' transcript ends, and because of the close proximity and complementary nature of the resulting transcripts they would anneal, thereby adding further complexity to the *E. coli* transcriptome. Figure 1 depicts an intrinsic terminator located between convergent operons, which terminates transcription by 4-fold.

However, read-through transcription of 329 bases of complementary antisense RNA for the 3' end of the convergent operons is still observed. My analyses of 370 instances of convergent operons revealed that 75% demonstrated transcription into an adjacent operon and generated complementary 3' transcript ends that overlapped by an average of 286 bases, with the longest of these being 1,395 bases. In regions of the genome where there were many highly transcribed operons, it was more likely that convergent transcription was observed. Of the genomic regions corresponding to these convergent operons, 74% were highly conserved at the nucleotide sequence level in >50 *E. coli* (and *Shigella*) genomes. It is therefore reasonable to conclude that overlapping transcription of convergent operons is a common feature in bacteria.

Transcription of divergent operons has been shown to result in overlapping transcripts (22, 23). Complementary transcripts generated by divergent promoters have recently been termed "excludons." These excludons are thought to act as negative regulators of genes on the opposite strand (38). The analyses performed here of the 388 instances of divergent operons revealed that 35% have promoters arranged in such a way that their 5' transcript ends overlap by an average of 168 bases, the longest of which is 1,012 bases. The genomic regions corresponding to 81% of these overlapping divergent operons are highly conserved in >50 *E. coli* (and *Shigella*) genomes. The discovery and cataloguing of sequence conservation alone does not begin to explain the function of these features, but the finding that over one-third of divergent operons generate overlapping complementary transcripts supports the idea that excludons may be prevalent in bacteria.

Comparison to Other Data Sets

The data presented here were compared to other high-quality data sets generated by RNA-seq that utilized a similar conservative analytical approach. A concurrent study of the *E. coli* transcriptome by Storz, Sharma, and colleagues focused on AS transcripts (68). Storz, Sharma, and colleagues found that most previously annotated sRNAs are in fact present at high levels, so in an effort to evaluate their observation, we compared our antisense RNA data set to the most highly expressed AS RNAs in their study. Our data were able to corroborate 74 of their 127 most highly expressed antisense RNAs. Furthermore, we corroborated 6 of 14 candidate antisense RNAs tested on Northern blots by the Storz group. However, while their gels verified 6 of the 14, we corroborated only 2 of those 6, indicating that there is substantial variability in these two high-throughput data sets. A recent co-immunoprecipitation study of the double-stranded *E. coli* transcriptome revealed 316 double-stranded RNAs, including partially and fully overlapping transcripts as well as many generated by divergent and convergent operons (37). Our analyses predicted antisense RNAs corresponding to 13 of 21 double-stranded RNAs that were verified in Northern blot analysis (37). It is tempting to speculate that antisense RNAs that are corroborated by RNA-seq studies, verified by Northern blot analysis, and correspond to highly conserved genomic sequences are functional. However, functions have been confirmed for only a limited number of antisense RNAs (82, 83). It therefore is essential that more studies on the function of antisense RNA in bacteria be conducted before a definitive conclusion is made. It remains to be seen how many of the antisense RNAs identified by RNA-seq will prove to be expressed inside the same cell as the sense transcript and display a

phenotype. Such a study would require the use of single-cell transcriptomics, a technique that is on the horizon, but not currently utilized in the study of bacterial transcription.

Bacterial operons compared to eukaryotic genes

It did not escape my attention that the widespread occurrence of bacterial operons with multiple transcription units is in some ways parallel to the events of alternative splicing within eukaryotic organism. Both bacterial operons and eukaryotic genes arise to primary transcripts that are divided into alternative transcripts by either the activity of transcriptional regulatory features, such as internal promoters and terminators in bacteria or RNA splice junctions in eukaryotes. The potential complexity for a given eukaryotic gene is reflected in the number of exons the average gene contains. The number of exons per gene in *Saccharomyces cerevisiae* was estimated to be 1.1 (84), which is fewer than the 1.7 transcription units per operon we observed in *E. coli*. To put the comparison of *E. coli* and *S. cerevisiae* in perspective, higher organisms, such as *C. elegans*, have 4 to 9 introns per gene (85), making them considerably more complex than *E. coli*. It has been proposed that there was a loss of exons that took place in budding yeasts during their evolution from their more primitive eukaryotes ancestors, so this may accentuate their difference from *E. coli* and higher organisms (86). Whatever the cause, it can be concluded that *E. coli* possesses operon complexity comparable to analogous gene structures in budding yeasts.

Concluding statement

This study revealed the power of single- nucleotide resolved RNA-seq data sets for pinpointing transcriptional features and annotating operons across the genome, which I used to evaluate the hypothesis that bacterial operon structure accommodates substantial transcriptional complexity. The level of complexity that was discovered is astounding. A substantial number of overlapping transcripts were identified. In these instances, complementary RNAs were transcribed from both strands, such as those generated by several hundred convergent and divergent operons. More than 100 long antisense transcripts overlapping operons that also are transcribed on the sense strand were also discovered. In total, we found that approximately one in three (519 out of 1,510) operons at least partially overlaps with other operons to generate antisense RNA. These antisense transcripts are highly conserved in *E. coli* and appear to be noncoding, suggesting that they are involved in regulation of gene expression, as has been proposed for the excludon concept in bacteria (38) and the long-noncoding RNA model in eukaryotes (80). It was determined that seven previously unrecognized transcripts that did not correspond with annotated gene(s) were present in expressed operons. The transcriptome complexity we observed in *E. coli* appears to be a general property of the domain bacteria, as the transcriptomes of several other bacteria appear to be similarly intricate (26, 31, 33, 35, 45, 57–59). Whether the same holds true of the Archaea must await high-resolution RNA-seq analysis of representatives of this domain of life (87).

The operon concept presented by Jacob Monod in 1961 articulated a model for the regulation of bacterial gene expression that has stood unchanged for nearly 55 years. While this model still holds true today, it is important to revisit past ideas with modern

methods in an effort to strengthen the scientific process. In its most simple form the operon is analogous to a light switch. Under the correct physiological conditions a bacterial operon will be expressed, i.e., turned on, and will continue to transcribe RNA until the stimulating condition changes and expression is turned off. What can be lost in the current version of the operon model are the subtle layers of complexity that fine-tune the expression of individual genes within a single operon. In reality, bacterial gene expression is not an all-or-nothing event. In fact, individual genes within a single operon are often expressed at different levels. These changes in expression can be explained when a detailed analysis of bacterial promoter and terminator locations are annotated on a high-resolution transcriptome. In addition, gene order within an operon has been shown to play an important role in the abundance of corresponding mRNAs. As RNA polymerase transcribes the genes within an operon it is inevitable that some fraction of the total RNA polymerase population will falter and fall off the template DNA. When these events take place, downstream genes are not transcribed and the abundance of mRNA reflects the occurrence of this phenomena. Therefore transcription of the later genes in a polycistronic operon is less abundant then the genes closer to the promoter.

After years of studying and debating the operon concept in the context of high-resolution RNA-seq data, I am of the opinion that the operon is analogous to an electrical circuit. In an electrical circuit the amount of current can be increased or decreased with capacitors or resistors, respectively. So too can the genes within an operon be fine-tuned to achieve an ideal balance. Hypothetically, if the second gene in an operon is needed in higher quantity, then a secondary or internal promoter can

increase its transcription. Alternatively, a decline in transcript abundance can be achieved by internal terminators, differential RNA decay via antisense transcription, or through the natural loss of RNA polymerase from the DNA template. While it is true that the end result of both models is the same, i.e., a light is turned on, what the circuit model accommodates is the ability to turn on three different light bulbs with three different intensities all with the same initial input.

Acknowledgements

This work was funded primarily by U.S. Public Health Service NIH RC1GM09207 to B.L.W. and T.C. from 2009 to 2011. B.L.W. is currently supported by NSF award 106394. Additional support was from NIH GM095370 to T.C., Grants-in-Aid for Scientific Research 21710198 to T.S. and 17076016, 8310133, and 21241047 to A.I. from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Nano- Biology Project fund from Micro-Nanotechnology Research Center of Hosei University to A.I., Grant-in-Aid for Scientific Research 22241050 and 25250028, Japan Society for the Promotion of Science (JSPS), Granting-Aid for Scientific Research on Innovative Areas 25108716, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and Grant-in-Aid for Scientific Research on Priority Areas to H.M.

References

1. **Jacob F, Monod J.** 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3:318 –356.

2. **Lederberg J, Tatum EL.** 1946. Gene recombination in *Escherichia coli*. Nature 158:558..

3. **Lehman IR, Bessman MJ, Simms ES, Kornberg A.** 1958. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. J. Biol. Chem. 233:163–170.

4. **Lengyel P, Speyer JF, Ochoa S.** 1961. Synthetic polynucleotides and the amino acid code. Proc. Natl. Acad. Sci. U. S. A. 47:1936–1942.

5. **Luria SE, Delbrück M.** 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491–511.

6. **Arber W, Dussoix D.** 1962. Host specificity of DNA produced by *Escherichia coli*. I. Host controlled modification of bacteriophage lambda. J. Mol. Biol. 5:18–36.

7. **Mulligan RC, Berg P.** 1980. Expression of a bacterial gene in mammalian cells. Science 209:1422–1427.

8. **Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR.** 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31:147–157.

9. **Boyer PD, Cross RL, Momsen W.** 1973. A new concept for energy coupling in oxidative phosphorylation based on a molecular explanation of the oxygen exchange reactions. Proc. Natl. Acad. Sci. U. S. A. 70:2837–2839.

10. **Chang CN, Model P, Blobel G.** 1979. Membrane biogenesis: cotranslational integration of the bacteriophage f1 coat protein into an Escherichia coli membrane fraction. Proc. Natl. Acad. Sci. U. S. A. 76:1251–1255.

11. **Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1462.

12. **Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G III, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL.** 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res. 34:1–9.

13. **Balázsi G, Barabási AL, Oltvai ZN.** 2005. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 102:7841–7846.

14. **Price MN, Arkin AP, Alm EJ.** 2006. The life-cycle of operons. PLoS Genet. 2:e96.

15. **Zhang H, Yin Y, Olman V, Xu Y.** 2012. Genomic arrangement of regulons in bacterial genomes. PLoS One 7:e29496.

16. **Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM.** 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS Genet 9:e1003495.

17. **Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN.** 2014. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. BMC Genomics 15:684.

18. **Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC.** 2012. The transcriptional landscape and small RNAs of *Salmonella enterica serovar Typhimurium*. Proc. Natl. Acad. Sci. U. S. A. 109:E1277–E1286. http://dx.doi.org/10.1073/pnas.1201061109.

19. **Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP.** 2014. Conservation of transcription start sites within genes across a bacterial genus. MBio 5:e01398-01314.

20. **Soutourina OA, Monot M, Boudry P, Saujet L, Pichon C, Sismeiro O, Semenova E, Severinov K, Le Bouguenec C, Coppee JY, Dupuy B, Martin-Verstraete I.** 2013. Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. PLoS Genet 9:e1003493.

21. **Taylor K, Hradecna Z, Szybalski W.** 1967. Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA. Proc. Natl. Acad. Sci. U. S. A. 57:1618 –1625.

22. **Piette J, Cunin R, Boyen A, Charlier D, Crabeel M, Van Vliet F, Glansdorff N, Squires C, Squires CL.** 1982. The regulatory region of the divergent argECBH operon in *Escherichia coli* K-12. Nucleic Acids Res.10:8031–8048.

23. **Wek RC, Hatfield GW.** 1986. Nucleotide sequence and in vivo expression of the ilvY and ilvC genes in *Escherichia coli* K-12. Transcription from divergent overlapping promoters. J. Biol. Chem. 261:2441–2450.

24. **Nomura T, Aiba H, Ishihama A.** 1985. Transcriptional organization of the convergent overlapping dnaQ-rnh genes of *Escherichia coli.* J. Biol. Chem. 260:7122–7125.

25. **Sameshima JH, Wek RC, Hatfield GW.** 1989. Overlapping transcription and termination of the convergent ilvA and ilvY genes of *Escherichia coli.* J. Biol. Chem. 264:1224–1231.

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

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

28. **Dornenburg JE, Devita AM, Palumbo MJ, Wade JT.** 2010. Widespread antisense transcription in *Escherichia coli*. mBio 1(1):e00024-10.

29. **Wade JT, Dornenburg JE, Devita AM, Palumbo MJ.** 2010. Reply to "Concerns about recently identified widespread antisense transcription in *Escherichia coli*." mBio 1(2):e00119-10.

30. **Raghavan R, Sloan DB, Ochman H.** 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. mBio 3(4):e00156-12.

31. **Behrens S, Widder S, Mannala GK, Qing X, Madhugiri R, Kefer N, Mraheil MA, Rattei T, Hain T.** 2014. Ultra deep sequencing of *Listeria monocytogenes* sRNA transcriptome revealed new antisense RNAs. PLoS One 9:e83979.

32. **Chatterjee A, Johnson CM, Shu CC, Kaznessis YN, Ramkrishna D, Dunny GM, Hu WS.** 2011. Convergent transcription confers a bistable switch in *Enterococcus faecalis* conjugation. Proc. Natl. Acad. Sci. U. S. A. 108:9721–9726.

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

34. **Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penadés JR, Valle J, Solano C, Gingeras TR.** 2011. Genome-wide antisense transcription drives mRNA processing in bacteria. Proc. Natl. Acad. Sci. U. S. A. 108:20172–20177.

35. **Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH.** 2012. Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. PLoS One 7:e43350.

36. **Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P, Sorek R.** 2012. Comparative transcriptomics of pathogenic and nonpathogenic Listeria species. Mol. Syst. Biol. 8:583.

37. **Lybecker M, Zimmermann B, Bilusic I, Tukhtubaeva N, Schroeder R.** 2014. The double-stranded transcriptome of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 111:3134 –3139.

38. **Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P.** 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol. 11:75– 82.

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

40. **Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BØ.** 2012. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. PLoS Genet. 8:e1002867.

41. **Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E.** 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One 4:e7526.

42. **Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J.** 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 41: D203–D213.

43. **Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J.** 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13:734.

44. **Li S, Dong X, Su Z.** 2013. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K-12 through accurate full-length transcripts assembling. BMC Genomics 14:520.

45. **Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC.** 2013. An infection-relevant transcriptomic compendium for *Salmonella enterica serovar Typhimurium*. Cell Host Microbe 14:683–695.

46. **Neidhardt FC, Bloch PL, Smith DF.** 1974. Culture medium for *enterobacteria*. J. Bacteriol. 119:736–747.

47. **Fabich AJ, Jones SA, Chowdhury FZ, Cernosek A, Anderson A, Smalley D, McHargue JW, Hightower GA, Smith JT, Autieri SM, Leatham MP, Lins JJ, Allen RL, Laux DC, Cohen PS, Conway T.** 2008. Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. Infect. Immun. 76:1143–1152.

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

49. **Wilmes-Riesenberg MR, Wanner BL.** 1992. TnphoA and TnphoA' elements for making and switching fusions for study of transcription, translation, and cell surface localization. J. Bacteriol. 174:4558–4575.

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

51. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078 –2079.

52. **Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F.** 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14:671–683.

53. **Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, Grant GR, Hogenesch JB.** 2014. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol 15:R86.

54. **Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH.** 2009. JBrowse: a next-generation genome browser. Genome Res. 19:- 1630–1638.

55. **Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR.** 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. 27: 3821–3835.

56. **Tao H, Bausch C, Richmond C, Blattner FR, Conway T.** 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol. 181:6425–6440.

57. **Lin YF, A DR, Guan S, Mamanova L, McDowall KJ.** 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation

and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease. BMC Genomics 14:620.

58. **Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H**. 2013. RNA-Seq of *Bacillus licheniformis*: active regulatory RNA features expressed within a productive fermentation. BMC Genomics 14:667.

59. **Balasubramanian D, Kumari H, Jaric M, Fernandez M, Turner KH, Dove SL, Narasimhan G, Lory S, Mathee K.** 2014. Deep sequencing analyses expands the *Pseudomonas aeruginosa* AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response. Nucleic Acids Res. 42:979–998.

60. **Bohannon DE, Connell N, Keener J, Tormo A, Espinosa-Urgel M, Zambrano MM, Kolter R.** 1991. Stationary-phase-inducible "gearbox" promoters: differential effects of katF mutations and role of sigma 70. J. Bacteriol. 173:4482–4492.

61. **Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy T, Taylor J, Nekrutenko A.** 2014. Dissemination of scientific software with Galaxy ToolShed. Genome Biol 15:403.

62. **Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D.** 2010. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26:2204-2207.

63. **Goecks J, Nekrutenko A, Taylor J, Galaxy T.** 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86.

64. **Egan SE, Fliege R, Tong S, Shibata A, Wolf RE, Jr, Conway T.** 1992. Molecular characterization of the Entner-Doudoroff pathway in *Escherichia coli*: sequence analysis and localization of promoters for the edd-eda operon. J. Bacteriol. 174:4638–4646.

65. **Deana A, Celesnik H, Belasco JG.** 2008. The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. Nature 451:355–358.

66. **Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z.** 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. 23:137–144.

67. **Ma Q, Liu B, Zhou C, Yin Y, Li G, Xu Y.** 2013. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. Bioinformatics 29:2261-2268.

68. **Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G.** 2014. Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in *Escherichia coli*. J Bacteriol.

69. **Grant CE, Bailey TL, Noble WS.** 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:-1017–1018.

70. **Robison K, McGuire AM, Church GM.** 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284:241–254.

71. **Mitchell JE, Zheng D, Busby SJ, Minchin SD.** 2003. Identification and analysis of "extended _10" promoters in *Escherichia coli*. Nucleic Acids Res. 31:4689–4695.

72. **Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN.** 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U. S. A. 99:9697–9702.

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

74. **Potrykus K, Murphy H, Chen X, Epstein JA, Cashel M.** 2010. Imprecise transcription termination within *Escherichia coli* greA leader gives rise to an array of short transcripts, GraL. Nucleic Acids Res. 38:1636 –1651.

75. **Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014**. GenBank. Nucleic Acids Res. 42:D32–D37.

76. **Cole ST, Honoré N.** 1989. Transcription of the sulA-ompA region of *Escherichia coli* during the SOS response and the role of an antisense RNA molecule. Mol. Microbiol. 3:715–722.

77. **Michán C, Manchado M, Dorado G, Pueyo C.** 1999. In vivo transcription of the *Escherichia coli* oxyR regulon as a function of growth phase and in response to oxidative stress. J. Bacteriol. 181:2759–2764.

78. **Vecerek B, Moll I, Bläsi U.** 2007. Control of fur synthesis by the noncoding RNA RyhB and iron-responsive decoding. EMBO J. 26:965–975.

79. **Cunningham L, Guest JR.** 1998. Transcription and transcript processing in the sdhCDAB-sucABCD operon of *Escherichia coli*. Microbiology 144(Part 8):2113–2123.

80. **Ponting CP, Oliver PL, Reik W.** 2009. Evolution and functions of long noncoding RNAs. Cell 136:629 – 641.

81. **Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA**. 2013. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat. Methods 10:659 – 664.

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

83. **Georg J, Hess WR.** 2011. Cis-antisense RNA, another level of gene regulation in bacteria. Microbiol. Mol. Biol. Rev. 75:286 –300.

84. **Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW.** 2006. Introns regulate RNA and protein abundance in yeast. Genetics 174: 511–518.

85. **Koralewski TE, Krutovsky KV.** 2011. Evolution of exon-intron structure and alternative splicing. PLoS One 6:e18055.

86. **Carmel L, Rogozin IB, Wolf YI, Koonin EV.** 2007. Patterns of intron gain and conservation in eukaryotic genes. BMC Evol. Biol. 7:192.

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

Supplemental Material

Bacterial Culture Conditions

To annotate operons and characterize their response to carbon starvation, we obtained time series of RNA samples from replicated wild type *E. coli* K-12 (strain BW38038) cultures grown to stationary phase on morpholinopropanesulfonate (MOPS) glucose minimal medium (fig. 2-S1). These conditions are intrinsic to the physiology that allows *E. coli* to colonize the mammalian intestine yet survive in a nutrient-depleted environment until encountering a new host and in the case of *E. coli* pathogens, cause disease (1).

The wild type strain of *E. coli* K-12 used in these studies was *E. coli* BW38038. *E. coli* BW39452 (Δ*rpoS*) was constructed from *E. coli* BW38028 by allelic replacement as described by Datsenko and Wanner (2). *E. coli* BW38028 and BW39452 were grown on lysogeny broth (LB) agar plates overnight from viable frozen stock cultures. Colonies from LB plates were used to inoculate 5 ml cultures of MOPS minimal medium (3) containing 0.05% glucose and grown overnight (16 h) at 37°C with shaking at 250 rpm. To ensure growth through 10 generations prior to taking the first sample, the overnight cultures were diluted 1:10,000 into a 2L B. Braun Biostat® B fermenter with working volume of 1 L MOPS minimal medium with 0.2% glucose, at 37°C, pH was kept constant at 7.4 by the addition of 1 M NaOH, and dissolved oxygen was maintained above 40% of saturation by adjusting the agitation speeds in the range of 270–500 rpm with fixed 1.5 liter/min air flow. Culture samples were harvested by using a homemade sampling device seven times during logarithmic growth and three

times following entry into stationary phase for the WT and two times during logarithmic phase and three times during stationary phase for *E. coli* BW39452 (Δ*rpoS*). OD600 measurements were made on a Beckman Coulter DU 800 spectrophotometer. Samples were harvested directly into ice-cold RNAlater at a 1:1

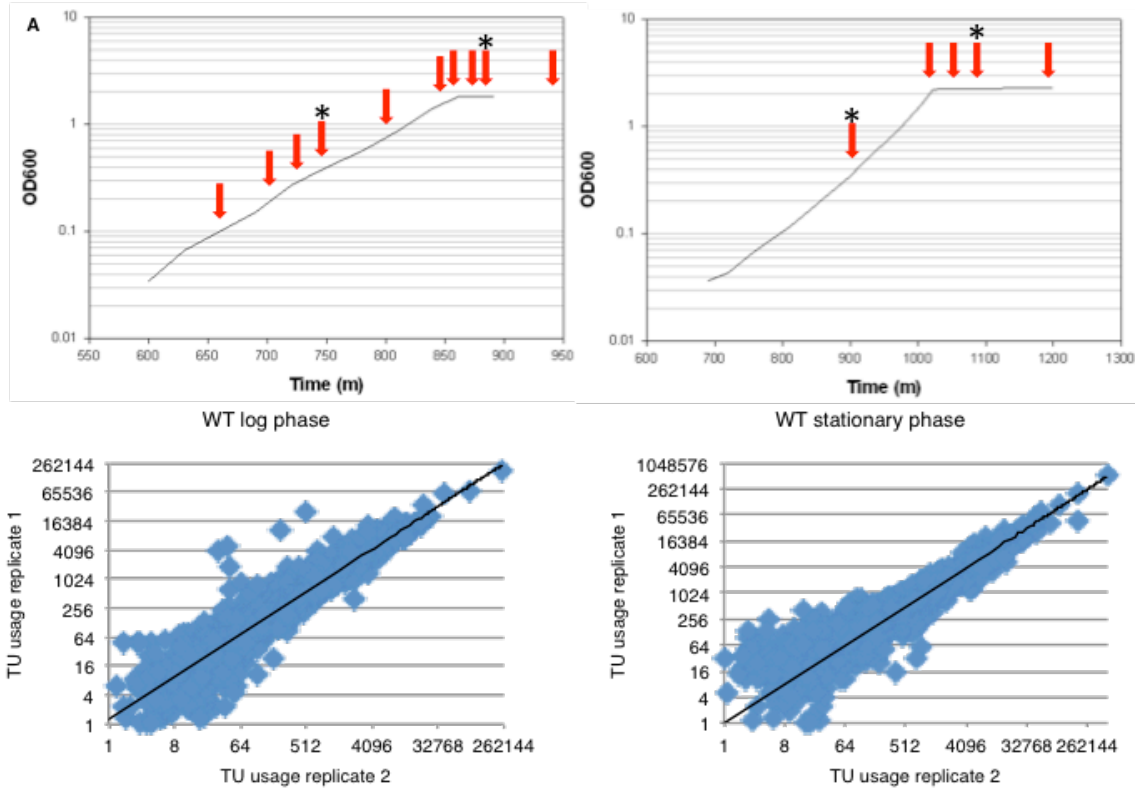


Figure 2-S1: Growth conditions for total RNA sampling and base count data replicates. (A) Wild-type *E. coli* BW38028 was grown on MOPS glucose minimal medium in a 2-liter Biostat B fermenter (Braun Biotech) with a 1-liter working volume at 37°C, pH was kept constant at 7.4 by the addition of 1 M NaOH, and dissolved oxygen was maintained above 40% of saturation by adjusting the agitation speeds in the range of 270 to 500 rpm with fixed 1.5 liters/min airflow. Total RNA was prepared from culture samples taken at 10 time points, indicated by red arrows. Replicate samples from duplicate cultures were taken at times indicated by asterisks. (B) *E. coli* BW39452 (Δ*rpoS*::*cat*) grown as described for panel A; (C) normalized transcription unit (TU) usage values from replicate 1 plotted against values from replicate 2 for logarithmic-phase samples; (D) normalized TU usage values from replicate 1 plotted against values from replicate 2 for stationary-phase samples. All annotated TUs (see Table 2-S2) are plotted. The trend line is shown as a solid black line. The correlations are R = 0.97 for stationary-phase samples and R = 0.96 for logarithmic-phase samples. (Adopted from T. Conway, 2014)

dilution to protect RNA from degradation and cells were then pelleted by centrifugation at 8000rpm for 10 minutes. Cell pellets were stored no longer than 8 weeks at -80°C in an equal volume of RNAlater prior to RNA extraction.

RNA Extraction and Manipulations for sequencing.

Total RNA was prepared as follows. Cell pellets were thawed on ice, resuspended in 200uL of bacterial lysis buffer containing lysozyme, and RNA was extracted and purified by using RNeasy Mini Kits (Qiagen, USA) according to the manufacturers instructions. DNA was digested by on-column DNase treatment. RNA quality and concentration were estimated by measuring A260 to A280 ratio. Since RNeasy columns do not capture small RNAs, these were excluded from the analysis.

Some RNA samples were ribo-depleted prior to sequencing (Table 2-S1). Ribosomal RNA was removed by using a MICROBExpress kit (Ambion, Austin, TX, USA), according to the manufacturer's recommendations. Sequence comparison of ribo-depleted samples with total RNA samples confirmed that ribo-depletion did not affect subsequent transcriptome analysis of normalized datasets, as has been noted by others (4). Subsequently all RNA samples were ribo-depleted to maximize mRNA-specific reads.

In preparation of adapter ligation, total RNA samples were fragmented with RNase III enzyme (Ambion, AM2290) and 15 µg of RNA was digested in 5 µl 10X RNase III Reaction Buffer, 15 µl RNase III (15U), and Nuclease-free water to a final volume of 50 µl. Following incubation for 1 hour at 37°C the fragmented RNA was purified by using Microcon-30 filter columns (Millipore, #42409) per manufacture’s

recommendations. The resulting RNA fragment size distribution was approximately 200 bases.

Table 2-S1: RNA-Seq Datasets

Experiment	Sample conditions	Sample name	Raw reads	Mapped reads	Base counts
WT growth curve replicate 1	A600=0.1	WT_01_rep1	17,916,163	1,392,530	128,850,241
	A600=0.2	WT_02_rep1	25,416,914	1,902,936	178,100,925
	A600=0.3	WT_03_rep1	24,761,958	1,793,606	163,634,704
	A600=0.4-R	WT_04-R_rep1	76,124,279	20,250,551	1,983,779,178
	A600=0.4	WT_04_rep1	29,454,924	2,306,139	197,698,048
	A600=0.8	WT_08_rep1	22,358,046	1,226,176	122,661,558
	A600=1.4	WT_14_rep1	17,456,057	1,037,010	95,075,084
	A600=1.6-R	WT_16-R_rep1	33,685,608	6,485,556	528,960,979
	A600=1.6	WT_16_rep1	14,438,194	886,571	75,267,685
	Stationary +15	WT_15min_rep1	15,391,998	575,313	55,743,748
	Stationary +30	WT_30min-R_rep1	17,099,756	1,463,700	130,000,353
	Stationary +30 -R	WT_30min_rep1	14,067,538	609,131	63,382,552
	Stationary +180	WT_180min_rep1	16,623,224	674,681	67,699,358
WT growth curve replicate 2	A600=0.4	WT_04_rep2	12,532,109	2,833,630	253,524,273
	A600=0.4, TEX	WT_04_TEX	10,966,516	3,630,591	121,688,880
	Stationary +30	WT_30min_rep2	10,935,145	1,647,956	125,991,970
	Stationary +30, TEX	WT_30min_TEX	9,141,228	2,071,630	72,375,264
rpoS growth curve replicate 1	A600=0.4	rpoS_04_rep1	7,606,636	1,864,606	110,258,777
	A600=1.6	rpoS_16_rep1	14,985,280	1,554,515	89,028,552
	Stationary +15	rpoS_15min_rep1	11,618,938	1,512,492	89,838,925
	Stationary +30	rpoS_30min_rep1	11,545,482	1,842,784	107,446,984
	Stationary +180	rpoS_180min_rep1	21,589,212	3,998,538	226,611,694
rpoS growth curve replicate 2	A600=0.4	rpoS_04_rep2	8,189,242	2,029,588	180,023,224
	A600=0.4, TEX	rpoS_04_TEX	9,623,380	4,277,038	151,956,384
	Stationary +30	rpoS_30min_rep2	11,520,689	1,031,601	78,922,992
	Stationary +30, TEX	rpoS_30min_TEX	11,555,208	3,248,876	116,654,064
Total			**476,603,724**	**72,147,745**	**5,515,176,396**

"-R" indicates ribo-depleted samples from WT growth curve replicate 1. All other samples were ribo-depleted.
"TEX" inidcates terminal exonuclease treated samples.
"A600=" indicates culture density at time of sampling.
"+15, +30, +180" indicates time (min) after entry into stationary phase.

Terminator 5'-phosphate-dependent exonuclease (Epicentre, #TER51020) treatment was used to enrich 5'-triphosphate mRNA fragments that correspond to the true transcription start site (5). Each RNA sample was split and a portion was TEX-enriched while the remainder of the sample was sequenced for coverage. TEX enrichment is a standard approach for determining TSS's because the enzyme degrades RNA containing 5' monophosphate ends, which can arise during RNA processing and decay. True TSS's begin with a 5'-triphosphate, as described below in the promoter mapping section.

Tobacco Acid Pyrophosphatase (Epicentre, #T19050) was used to remove 5' triphosphate ends and/or repair 5' monophosphate ends prior to adaptor ligation. On ice, 25 pmol of RNA was added to each reaction in the presence of 5 µl 10X TAP Reaction Buffer, 25 U of TAP and nuclease-free water to equal 50 µl total volume. The reaction was incubated at 37°C for 2 hours, and stopped by phenol extraction followed by ethanol precipitation.

Prior to ligation of SOLiD adaptors, total RNA quantity and quality were assessed via UV spectrophotometry and the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). RNA concentration was estimated by measuring A260 to A280 ratio. RNA integrity values (RIN values) and concentrations were determined by using the Agilent RNA 6000 Pico Chip Kit as specified by the manufacturer.

cDNA Synthesis for the SOLiD System.

To ensure sequencing of the 5' and 3' ends of mRNA, ligation-based chemistry was used. The SOLiD Total RNA-Seq Kit was used to ligate SOLiD specific adaptors to fragmented, end repaired RNA samples (100 ng). This kit was used to create a single stranded DNA-RNA hybrid molecule consisting of the adaptor ligated to the mRNA fragment. Single-stranded cDNA was prepared by reverse transcription using the provided SOLiD RT primer. The resulting cDNA was purified using the Qiagen MinElute PCR Purification system.

cDNA with an approximate size of 150-250 nt was isolated by gel electrophoresis on Novex gels. Using the 250 bp and 150 bp bands on the DNA ladder as a guide this region of the gel was excised and cut into four vertically equal pieces.

Each of the four pieces generated can be used for cDNA amplification. For this experiment the two center fragments were selected for library creation.

SOLiD Library Creation

In order to obtain an acceptable concentration of amplified cDNA, each reaction was prepared in duplicate. The starting material for the amplification reactions was the cDNA contained within the gel slices produced above. The cDNA was amplified using SOLiD PCR primers and AmpliTaq DNA Polymerase for 15 cycles as specified by the manufacturer. The two 100 µl PCR reactions were combined prior to the final purification step. The resulting sequencing libraries were purified by using the PureLink PCR Micro Kit (Invitrogen). Following sequencing library construction the total DNA quantity and quality were assessed via UV spectrophotometry and the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) using the Agilent DNA 1000 Kit. For each sample the percentage of DNA in the 25 to 200 bp range, the median peak size, and molar concentration were estimated.

SOLiD Sequencing

SOLiD sequencing was performed at Purdue University under the direction of Phillip San Miguel. Total RNA, ribo-depleted RNA and TEX treated RNA were prepared at the University of Oklahoma and shipped on dry ice to the Purdue University Genomics Core Facility, where the sequencing libraries were prepared as described above. The resulting cDNA was clonally amplified to by emulsion PCR. The beads are purified, enriched, and modified by terminal transferase to facilitate attachment to flow

chips. Bead enrichment and quality control was monitored to evaluate effectiveness of the process. Paired end sequences with 50-base forward read-lengths and either 25 or 35-base reverse read-lengths were generated on the SOLiD 4 Genetic Analyzer.

Records of the analysis performed by the Purdue University Genomics Core Facility were disseminated via a web-based notebook. All data collected by the Genomics Core Facility was provided to the University of Oklahoma via the Internet. The data are deposited at GEO under accession number GSE52059.

Raw SOLiD sequence data processing

The raw data output (CSFASTA and QUAL files) from the SOLiD 4 Genetic Analyzer were passed through the ABI Sequence Accuracy Enhancement Tool (SAET), which improves the color calling error rate by approximately five-fold. For alignment of the SAET reads to the *E. coli* MG1655 reference genome (RefSeq NC_000913), the short read alignment tool Bowtie ver. 1.8 (6) was utilized in three consecutive passes for each sample dataset. For the first pass, we use paired end color space mapping with a distance cutoff of 350 bases between read mates. Bowtie parameters were set to include only perfect matches and suppress reads that map to more than one genome location, i.e., uniquely mapped reads were retained. In practice we found the efficiency of paired end mapping was between 3 and 10%. To improve the overall alignment we mapped the orphan 5' and 3' end reads in two additional passes with Bowtie (one for the 5' reads and one for the 3' reads). The output of the three passes through Bowtie was three SAM files for each sample. Overall, we achieved 40-60% mapping efficiency with this three-pass strategy. SAMTOOLS (7) utilities were used to sort and index the SAM files and

convert them to BAM format. The sample alignment (BAM) files were displayed in Integrated Genome Viewer (IGV ver. 2) for primary analysis and quality control.

Base count processing of aligned sequence data

Sequence data were processed by conversion of the sample alignment (BAM) files to base count (WIG) files. Conversion of BAM data to WIG data results in a 100-fold reduction in file size, and a more readily "computable" dataset. To accomplish this an in-house script was written to extract strand-specific base count data from BAM files (outputs are positive and negative strand WIG files). First, our solidBam2wig.pl script reads in the paired-end BAM file and counts the nucleotides spanning inserts between the mated 5' and 3' reads as shown here. Next, the script pulls in the orphan 5' and 3' reads from the respective BAM files and increments the base counts at each base location without duplicating the reads already incremented from the paired ends.

Base count normalization.

Base count data were normalized based on the assumption that reads are randomly distributed across the genome and that if sequencing was sufficiently deep, all expressed transcripts would be represented in the dataset. In practice, SOLiD sequencing did not generate datasets in which the lowest abundance transcripts were fully covered by contiguous reads. In addition, inefficient ribo-depletion can bias the number of reads that map to non-rRNA genes. Our normalization strategy accounts for both of these factors by maximizing TU coverage and removing rRNA reads during data processing. Our in-house script, normWIG.pl, reads in the raw WIG files while

excluding counts from all 22 rRNA genes. A simple global normalization approach was utilized that multiplied the count at each base location by 1 billion and divides that value by the sum of base counts at all base locations in the file. This normalization strategy is analogous to the Total Count approach used for normalizing gene-specific read alignments (8). In this way, the base counts are expressed as parts per billion. For display in JBrowse (9), the normalized WIG data was log2 transformed and converted to web browser tracks using the wig-to-jason.pl script that is part of the JBrowse package (available at jbrowse.org).

Promoter mapping.

To map and annotate promoters, we combined differential RNA-Seq (10) and promoter motif analysis. These strategies are described in detail below. Although the state of the art of promoter annotation based on RNA-seq data is a manual process (11), we sought to automate it to the extent that was possible. Therefore, we wrote a simple algorithm to search for changes in normalized base count values exceeding two-fold in replicate TEX enriched and coverage datasets (n=14, WT and rpoS culture samples from log and stationary phase). Consensus of three or more replicates at the identical base location revealed 11,329 putative TSSs. This number of promoters far exceeds the expected promoter density on a genome containing 4492 genes, exemplifying the need for consensus scoring of promoters, as follows.

We used a bioinformatics approach to search the 50 base pair sequences immediately upstream of the putative TSSs for promoter motifs by using FIMO software (12) and screening against a library of *E. coli* transcription factor binding

motifs available at DPInteract (13). We found it necessary to modify the RpoD promoter library according to the characterization of 554 promoters by Mitchell et al. (14), which demonstrated that RpoD promoters have identical -10 and -35 regions differing by spacing of 14 to 20 bases between these promoter elements. We restricted the search output to promoter motifs correctly positioned within +/-3 bases of the TSS, with p-values <0.02, yielding 5653 putative promoters. As *E. coli* RNAseq datasets accumulate this automated strategy for promoter identification almost certainly will improve.

To identify putative promoters missed by TSS mapping we employed Genomic SELEX screening (described in detail below but not discussed in the main body of the manuscript), which was developed for quick identification of genes under the control of specific transcription factors (15). Confirmation of putative TSS's by RNAP binding was employed previously for promoter mapping of *S. Typhimurium* (11). Since RpoD and RpoS recognize similar promoter sequences under the standard conditions for transcription in vitro, we repeated the assays in high concentrations of potassium glutamate, which was previously shown to inhibit RpoD holoenzyme binding in a dose-dependent manner, whereas that of RpoS holoenzyme is activated (16). Combining all four datasets, sites that bind RpoS and/or RpoD exceeding a conservative threshold of 3.0 signal to background ratio identified an additional 1254 putative promoters.

Thus, the combination of consensus promoter mapping and SELEX guided us to 5653 RpoD and 1254 RpoS putative promoters for a total of 6907 we considered during manual annotation. We used a visual graphic environment (J-Browse (9)) that facilitates an interface to an Oracle database to manually document annotation information. From

the list of putative promoters we created a J-Browse track at the corresponding base locations, each displaying a "clickable" URL call that automatically recorded the base location and manually entered metadata, including the type of promoter, regulatory information supported by differential expression analysis, and comments.

We scored and weighted the results of the four determinations to obtain promoter confidence scores (table S2 from Conway, 2014). This strategy quantifies promoter quality on a 1-10 scale and goes beyond the recently proposed rating system for qualitatively classifying evidence codes, considered to be the "gold standard" for annotating *E. coli* regulatory features (20). This approach maximized our confidence that the mapped TSS's were in fact generated by promoter activity. The promoter dataset (table S2 from Conway, 2014) is dominated by P-promoters (66.3%), with a lower number of S- (19.6%), I- (9.8%), and AS- (4.2%) promoters. All possible arrangements and orientations of these promoter types were observed, and collectively generate substantial complexity in the transcriptome.

Operon mapping.

To annotate operons we found it necessary to annotate terminators at the same time as promoters, which allowed documentation of the transcriptional connections between the primary promoters and terminators that define them (Fig. 2-1). We automated the cataloging of base locations of promoters and terminators that define operons by integrating J-Browse with an Oracle database, which sped up the analysis by approximately 10-fold.

Criteria for operon annotation were: 1) the P-promoter must be followed by sequence read coverage across the entire operon; 2) the mapped TES must extend beyond the stop codon of the last gene in the operon; 3) S- and I-promoters must increase coverage of downstream bases; and 4) internal terminators must decrease coverage of downstream bases without interrupting contiguous coverage by read-through transcripts. To annotate transcription units (TU), the user links promoters to terminators (3' transcript ends) by annotation in the database. Users can add comments with each database record, view the history of related comments, flag the location for future analysis, and save and share screen shots with collaborators. The annotation database is a powerful tool for RNA-seq data analysis because it can be queried to generate lists of base locations and associated base count data for any annotated feature, such as TSSs, terminators, and transcription units. We can query large numbers of WIG files and return values representing relative TU and promoter usage, as well as terminator efficiency. Despite mapping reads to 96% of reference genes, this conservative strategy maps transcripts to only two-thirds of genes. An advantage of criterion 1 means we never find orphan promoters. The tradeoff is we map fewer operons, but the advantage is we have greater confidence in those annotated.

In total, we annotated 6463 regulatory features, including 2122 promoters (table S2 from Conway, 2014), 2566 TUs (table S3 from Conway, 2014), and 1774 terminators (table S4 from Conway, 2014). We analyzed the 264 examples of internal terminators and confirmed that all give rise to separate TUs, which apparently were generated by partial termination and hence allowed transcription read-through (Fig. 2-S3). We evaluated the 1774 TES's by using TransTermHP (17) and confirmed that 623

have sequences characteristic of intrinsic terminators that interact directly with RNA polymerase. It has been predicted that one-half of terminators are intrinsic (18). The remaining TES's were not confirmed by TransTermHP, indicating the possibility these terminators require Rho or another protein effector. Since there is no bioinformatics approach to identify protein-dependent terminators, our data represent the most extensive genome-wide prediction of non-intrinsic terminators.

Table S5 (see Conway, 2014) summarizes 1510 annotated operons. Operon and TU quality was assessed by the fraction of bases that were covered by 3 or more reads. Of 2566 annotated TU's 1256 had 100% coverage in at least one sample and 90% of TU's have greater than 90% base coverage (table S3 from Conway, 2014). This conservative strategy annotated operons covering only two-thirds of the genome, but the higher data quality offers greater analytical power.

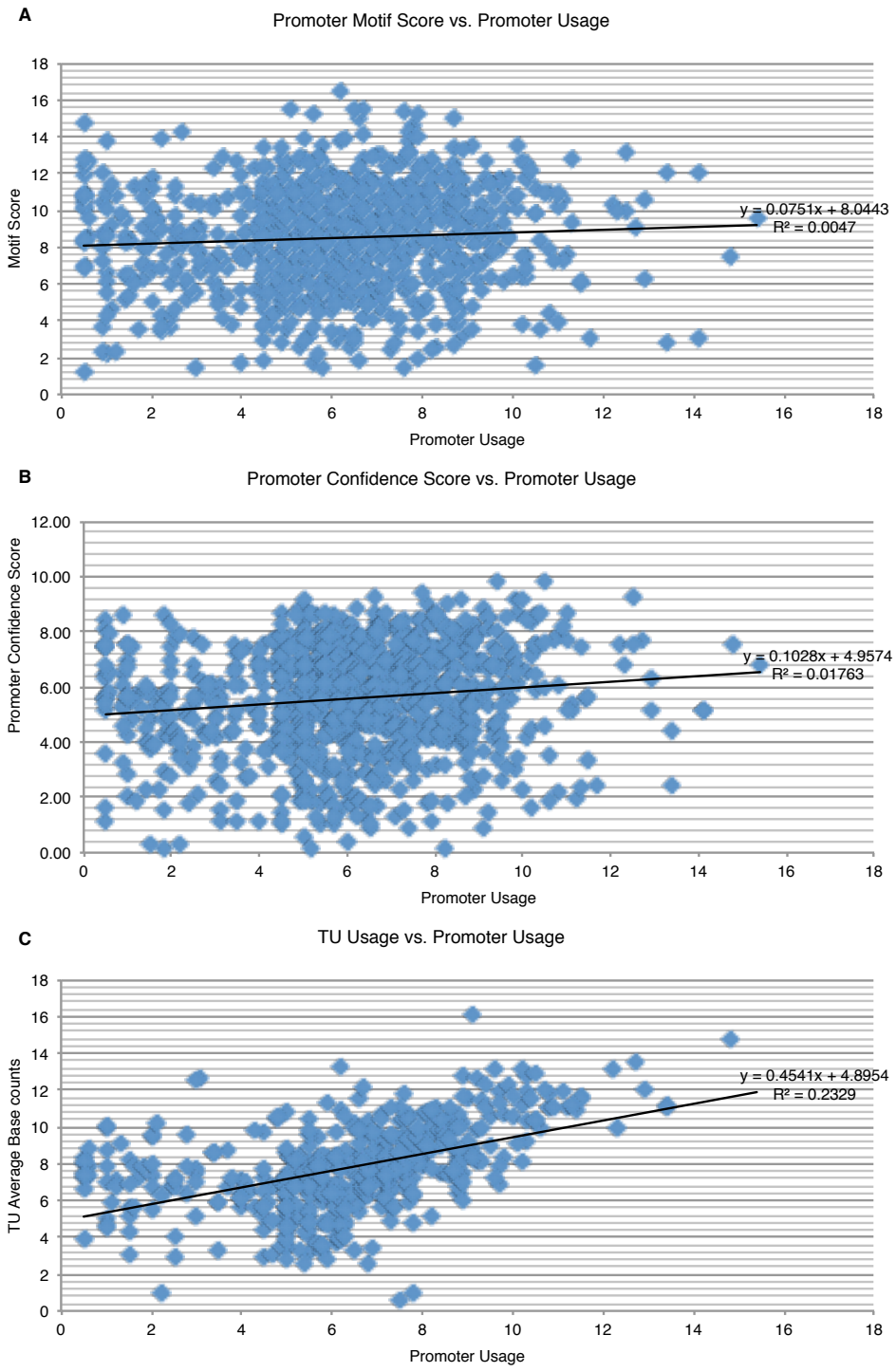


Figure 2-S2: Comparison of promoter usage to promoter metrics and TU usage. (A) Promoter motif score versus promoter usage; (B) promoter confidence score versus promoter usage; (C) TU usage versus promoter usage. Usage values were determined from normalized, log2 base count data, as described in detail in the text. (Adopted from T. Conway, 2014)

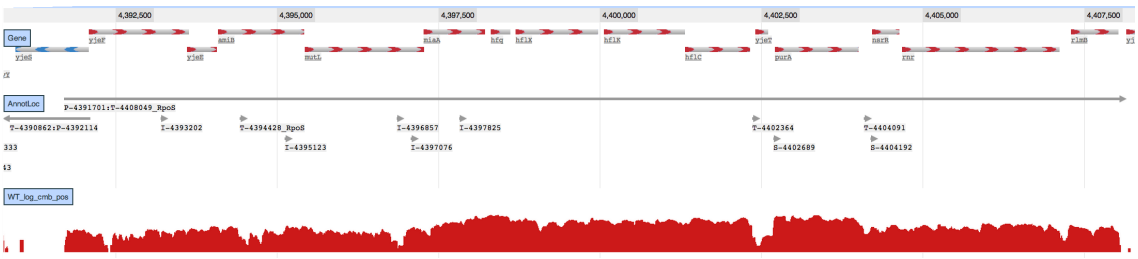


Figure 2-S3: Complex yjeF-yjeE-amiB-mutL-miaA-hfq-hflX-hflK-hflC-yjeT-purA-nsrR-rnr-rlmB operon. This operon has 8 promoters and 4 terminators and contains 23 transcription units created by transcription initiation from S and I promoters, as well as termination and transcriptional read-through at internal terminators. (Adopted from T. Conway, 2014)

Supplemental Material References

1. Fabich AJ, et al. (2008) Comparison of carbon nutrition for pathogenic and commensal Escherichia coli strains in the mouse intestine. Infect Immun 76(3):1143-1152.

2. Datsenko KA & Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A 97(12):6640-6645.

3. Neidhardt FC, Bloch PL, & Smith DF (1974) Culture medium for enterobacteria. J Bacteriol 119(3):736-747.

4. Haas BJ, Chin M, Nusbaum C, Birren BW, & Livny J (2012) How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13:734.

5. Sharma CM & Vogel J (2014) Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin in Microbiol 19:97-105.

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

7. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

8. Dillies MA, et al. (2012) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform.

9. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, & Holmes IH (2009) JBrowse: a next-generation genome browser. Genome research 19(9):1630-1638.

10. Sharma CM, et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464(7286):250-255.

11. Kroger C, et al. (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A 109(20):E1277-1286.

12. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017-1018.

13. Robison K, McGuire AM, & Church GM (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J Mol Biol 284(2):241-254.

14. Mitchell JE, Zheng D, Busby SJ, & Minchin SD (2003) Identification and analysis of 'extended -10' promoters in Escherichia coli. Nucleic Acids Res 31(16):4689-4695.

15. Tuerk C, MacDougal S, & Gold L (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. Proc Natl Acad Sci U S A 89(15):6988-6992.

16. Kusano S, Ding Q, Fujita N, & Ishihama A (1996) Promoter selectivity of Escherichia coli RNA polymerase E sigma 70 and E sigma 38 holoenzymes. Effect of DNA supercoiling. J Biol Chem 271(4):1998-2004.

17. Kingsford CL, Ayanbule K, & Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 8(2):R22.

18. Potrykus K, Murphy H, Chen X, Epstein JA, & Cashel M (2010) Imprecise transcription termination within Escherichia coli greA leader gives rise to an array of short transcripts, GraL. Nucleic Acids Res 38(5):1636-1651.

Chapter 3: Quantitative Bacterial Transcriptomics with RNA-seq

Chapter Summary

The material presented in this chapter was accepted for publication in a special bacterial genomics issue of *Current Opinions in Microbiology* on November 13th 2014. Following the publication of our "*Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing"* article in mBio, Dr. Conway was invited to write a review article detailing our methods for the quantitative analysis of RNA-seq data. Dr. Conway provided me with the opportunity to draft the manuscript, and mentored me through the process of editing, submitting, and revising my first first-authored manuscript.

It was my intent that this article highlights the potential of RNA-seq data for describing transcriptional events in a biologically significant manner. I am of the opinion that this manuscript was a significant contribution to the scientific community, and sheds light on an aspect of bacterial transcriptomics that is often overlooked. That once mapped, the activity of transcriptional features are quantifiable. While technically this manuscript was published as a review article on quantitative bacterial transcriptomics, I would point out that it was written more in the style of a case example. As such, the reader was taken through the analytical process step-by-step with a novel dataset, and the experimental rationale was explained and justified. The resulting manuscript highlights how quantitative transcriptome analysis can reveal biological insights and briefly discusses some of the challenges that face the field of bacterial transcriptomics.

Introduction

Advances in RNA-seq technology have revolutionized the study of bacterial transcriptomes [1,2]. At its core, RNA-seq generates digital information that allows transcriptional features to be located with single-nucleotide precision in a strand specific manner. Since the data are digital, RNA-seq facilitates quantitative computational analysis of any selected region of the transcriptome, but the transcriptome must first be annotated properly. Since bacterial genomes are organized in operons, it is logical that RNA-seq data should be annotated with the operon architecture in mind. In practice, only three transcriptional features need to be defined: 5' transcript ends (promoters), 3' ends (terminators), and RNA sequence read coverage to connect the ends, which together define operons [3-5].

The true power of RNA-seq resides in its potential as an analytical tool for quantifying promoter activity, terminator efficiency, and differential expression of transcripts, including operons, transcription units within operons (e.g. generated by promoters internal to operons), and antisense RNAs. RNA-seq datasets consist of tens of millions of sequence reads and typically the reads are 50 bases in length. The raw sequence reads are aligned to a reference genome and only high quality reads are retained and mapped. Conversion of sequence data into digital format is accomplished by employing freely available computer scripts that count the number of times each transcribed base was sequenced in a read-aligned dataset, thereby converting aligned sequence reads to base count data. Normalization of the base count data is necessary to quantify the differential expression (i.e., relative base counts) of each transcriptional feature within a sample or between different samples. The normalized base count data

can be quantified by averaging the base count across a selected region of the genome. Since the average of the base counts is used, the relative expression of any given transcription feature, regardless of its length, can be expressed in this way. Here we focus on the analysis of an *E. coli* RNA-seq dataset to demonstrate the strategy we developed to quantify the expression of the transcriptional features that define operons in bacteria.

Single-nucleotide resolved RNA-seq dataset

To obtain an RNA-seq dataset suitable for quantitative analysis, we prepared RNA from a culture of *E. coli* K-12 strain BW38028 during logarithmic- and stationary-phase growth on glucose limited minimal medium, as described previously [4]. In addition, we starved *E. coli* BW38028 and its isogenic *rpoS* mutant BW39452 for nitrogen by decreasing by three-fold the amount of ammonium chloride in the growth medium [6]. The RNA was extracted by using the hot-phenol method [7] and DNase I treated to remove contaminating DNA. The RNA samples were not depleted for rRNA prior to sequencing, which tends to eliminate some experimental biases [8]. The RNA samples were shipped on dry ice to vertis Biotechnologie AG (Germany) for library preparation and Illumina HiSeq2000 sequencing, as described by others [7,9]. For library preparation the RNA samples were split and subjected to differential RNA-seq (dRNA-seq) as described [2,10]. Briefly, one portion of the RNA was fragmented by ultrasound and then the fragments were poly(A)-tailed and an RNA adapter was ligated to the 5’ phosphate of the RNA. First strand cDNA synthesis was with a poly(dT) primer and reverse transcriptase. Second strand cDNA synthesis incorporated a

barcoded 3’ Illumina TruSeq adapter. The other portion of the RNA samples were fragmented and treated with 5’-dependent terminator exonuclease (TEX), which enriches for 5’ triphosphate containing transcripts that are generated by transcription initiation at promoters. The TEX treated samples then were tailed and ligated, and cDNA was prepared as described above. The cDNAs were sequenced on an Illumina HiSeq2000 system using 50 bp read length, with each library yielding approximately 20 million reads.

Datasets consisting of 10 million reads per sample are sufficient for transcriptional feature mapping and differential gene expression analysis without ribo-depletion for a transcriptome the size of *E. coli* [9,11]. For quantification the genome-aligned, strand-specific RNA-seq data should be converted from aligned reads to base counts. Our RNA-seq data analysis pipeline involves alignment of the raw data to the reference genome by using Bowtie2 to generate the sequence read alignment file (SAM) [12]. SAMTOOLS [13] were used to convert the SAM file to a binary alignment file (BAM). The BAM file was converted to a BigWig file (base count file), which contains the count of the base at each base location and is the standard for visualization in genome browsers such as J-Browse [14]. Conversion of BAM to BigWig formatted files can be accomplished by using tools available in the Galaxy Toolshed [15] or at UCSC Genome Browser [16].

Alternatively, users can analyze their datasets by using pipelines such as Galaxy [17] or READemption [18], which outputs normalized wiggle files (base count files). A simple and straightforward way to normalize base count data is by using a strategy analogous to the total count approach [19] for normalizing gene-specific read

alignments, which expresses each value as the base count per billion bases counted [4]. Because the BigWig file represents the base count at each nucleotide position, all downstream analysis begins with this file. The advantages of the base count approach are: a) the digital base count data are inherently computable because of their format and smaller size, b) the average base counts of individual transcriptional features can be computed and queried at any desired resolution, from a single nucleotide to an entire operon, to quantify the expression level or activity, c) normalization of base count data makes all samples directly comparable, and d) the use of average base count values eliminates the length bias when comparing transcriptional features of different length [19].

Identification of transcription start sites

Several published RNA-seq studies have focused on transcription start site (TSS) identification [7,9,10,20-28]. The annotation of TSSs is essential for analyzing promoters, 5' UTRs, operon architecture, and for discovering novel transcripts. To assure accuracy, a set of "best practices" for TSS identification has begun to emerge. Enrichment of the 5' RNA ends that are generated by transcription initiation remains critical for accurate TSS identification. The many advantages of dRNA-seq were recently reviewed [2]. The initiating nucleotide in bacteria is a nucleotide triphosphate, which can be distinguished from 5'-monophosphate and 5'-OH containing RNAs that are generated by RNA processing or RppH pyrophosphohydrolase activity [29]. The enrichment strategy preferred by many researchers makes use of 5'-dependent terminator exonuclease (TEX), which degrades RNA with 5'-monophosphate ends to

enrich for primary transcripts that contain 5' triphosphate ends and hence represent the product of transcription initiation [10]. dRNA-seq works by enumerating differences in base counts between TEX-enriched and unenriched sequencing libraries. Experimental replication is critical for accurate TSS identification. Since dRNA-seq is remarkably reproducible, comparison of datasets generated by using the same protocols yet different growth conditions adds confidence to the process and the use of different growth conditions also increases the number of mapped TSSs. RNA samples from many growth conditions can be pooled for dRNA-seq identification of thousands of promoters [9]. For example, a recent dRNA-seq analysis of *Salmonella* using RNA pooled from 22 different growth conditions led to mapping of 96% of the TSSs that could be identified by independently analyzing the 22 samples [9].

When annotating transcriptome data, it is convenient to use widely available computer programs to search dRNA-seq datasets for TSSs [20,30,31]. The advantages of the computational process compared to manual annotation are the speed and precision of recording transcription feature locations. However, like all bioinformatics approaches, some features will be missed and there will be false positives. In the end, human supervision of the results is critical and the state-of-the-art in transcriptome annotation remains a manual process [9]. Manual annotation of TSSs is made more efficient by plotting the count of only the first base at the 5' end of each TEX-enriched read (Fig. 3-1A) [32]. In practice this allows visualization of the 5' triphosphate nucleotide at the TSS.

Subsequent to identification of TSSs by dRNA-seq, bioinformatics and functional analyses can add weight to promoter identification. For example, the DNA

sequences immediately upstream of putative TSSs can be analyzed by using a bioinformatics approach to score sigma factor specific RNA polymerase binding sequence motifs [4,33]. ChIP determination of RNA polymerase binding provides a robust and comprehensive validation of putative promoters [23]. When used in combination, dRNA-seq, consensus amongst experimental replicates, promoter sequence analysis, and RNA polymerase binding assays are a powerful set of tools for the identification of promoters.

Annotation of 3' ends

To obtain the full analytical value of RNA-seq data it is essential to map the 3' transcript ends. Annotating 3' ends is a notably more difficult endeavor than mapping TSSs because there currently is no method of enriching for them. The 3' ends are the primary sites of exonuclease-dependent RNA decay, which may be the reason that RNA base counts decline at the 3' ends of operons, and few reads extend into the stem loop structures of intrinsic terminators (Fig. 3-1C). Further complicating 3' end analysis is that termination is typically inefficient [34], which allows read-through transcription. Currently, the best method for annotating 3' ends is to search for correlation between replicates of the furthermost downstream bases transcribed, keeping in mind that the base counts near the 3' end will be low even for highly expressed transcripts. Comparison of the 3' ends to terminator predictions adds confidence to the analysis. For example, the TransTermHP software package works very well for finding intrinsic terminators [35]. In addition, a ChIP-chip analysis of the distribution of RNA polymerase after treatment with the Rho-specific inhibitor bicyclomycin led to

identification of 200 Rho-dependent terminators [36]. Once both the 5’ and 3’ transcript ends are mapped, it is possible to annotate operons.

Annotation of operons

The transcriptome is a map of the activities of promoters and terminators. These activities are located on both strands of the genome [37] and depending on their arrangement, can give rise to antisense transcription and overlapping, divergent [38,39] and convergent operons [40,41]. To accommodate this naturally occurring complexity it is necessary to annotate the operon architecture. Three transcriptional features are necessary to define operons: 5' ends (promoters), 3' ends (terminators), and sufficient RNA-seq read coverage to connect the ends. If sequence reads cover 90% of the bases, this is a sensible indicator that the operon is real [4,32]. While there are computer algorithms that can find operons [5,42,43], just as for TSS mapping, the state-of-the-art remains a manual process [9]. Once the operons have been mapped, it is a straightforward task to annotate additional promoters and terminators within operons, which add complexity to the transcriptome. Mapping of internal promoters can be done manually or by bioinformatics analysis of mapped promoters that fall within the base locations of annotated operons. The transcriptional feature locations can be formatted as a GenBank feature file by using “promoter”, “terminator” and “operon” as feature keys (see for example, GSE52059 [4]). This format accommodates incremental annotation of condition specific regulatory information and is an accepted standard for disseminating genome annotation data [44]. Once the transcriptional feature locations are annotated, it

is reasonably straightforward to calculate the average base count value for each feature, from each dataset, as described below.

Computing the activities of transcriptional features

Analysis of RNA-seq reads at the base count level permits normalized base counts to be readily averaged across any range of base locations to calculate the relative expression level, activity, or efficiency of individual transcriptional features [4]. We determined empirically that computing the average count of the first 10 transcribed bases accurately represents promoter activity and allows closely spaced promoters to be discriminated [4]. Likewise, the efficiency of transcription termination can be calculated as the relative decline in average base counts in 25-base windows before and after terminators (Fig. 3-1C). The relative transcript levels of operons can be calculated by averaging the base counts from the promoter to the terminator locations. Likewise, the expression levels of alternative transcripts generated by promoter and terminator activities within operons can be calculated. These applications of single-nucleotide-resolution analysis are exemplified in Fig. 3-1, for wild type *E. coli* K-12 during logarithmic growth on glucose minimal medium and during starvation for carbon (stationary phase) or nitrogen, as well as an *rpoS* mutant during nitrogen starvation.

The *cysK-ptsHI-crr* operon contains 4 genes and multiple transcription units (Fig. 3-1A). Conservatively, more than 40% of *E. coli* operons contain multiple transcription units that are differentially expressed, underscoring the need for an annotation system that accommodates operon architecture [4]. In addition to the primary promoter (P-1) and terminator (T-B) that define the operon, there are 8 additional

promoters and one terminator within the operon (Fig. 3-1A). The activities of the promoters range from 12 to more than 10,000 average base counts (calculated from +1 to +10 at each promoter) and their relative activities under the four growth conditions are plotted in Fig. 3-1B.

There are two promoters (P-1 and P-2), separated by 33 base pairs, which drive transcription of *cysK* (Fig. 3-1B). Comparison of the average counts of the first 10 transcribed bases indicates that P-2 is greater than 30-fold more active than P-1. Inefficient termination (approximately 40% of *cysK* transcripts are not terminated, as indicated by the ratio of average base counts) at the internal terminator (T-A) suggests that *cysK* and *ptsHI-crr* are co-transcribed (Fig. 3-1C). Nevertheless, the T-A terminator segments the operon into *cysK* and *ptsHI-crr* specific transcripts, which makes sense because CysK is a cysteine biosynthetic enzyme and the remaining genes encode components of the phosphotransferase system (PTS) involved in sugar uptake [45]. In the current annotation these genes are thought to comprise two operons (*cysK* and *ptsHI-crr*) [46], but the data in Fig. 3-1 show a low but significant number of RNA-seq reads across the terminator T-A, most clearly in the log phase sample. There is also a promoter (P-3) internal to *cysK* that under all four conditions is relatively active compared to the other promoters and could contribute to transcription across the *cysK-ptsH* intergenic region (Fig. 3-1B), yet P-3 activity does not appear to correlate with the base counts in the corresponding unenriched samples and therefore is unlikely to contribute to operon function (Fig. 3-1A). Given its location at the end of a transcript and immediately upstream of an inefficient terminator, this could be an example of a pervasive transcript, which is discussed below.

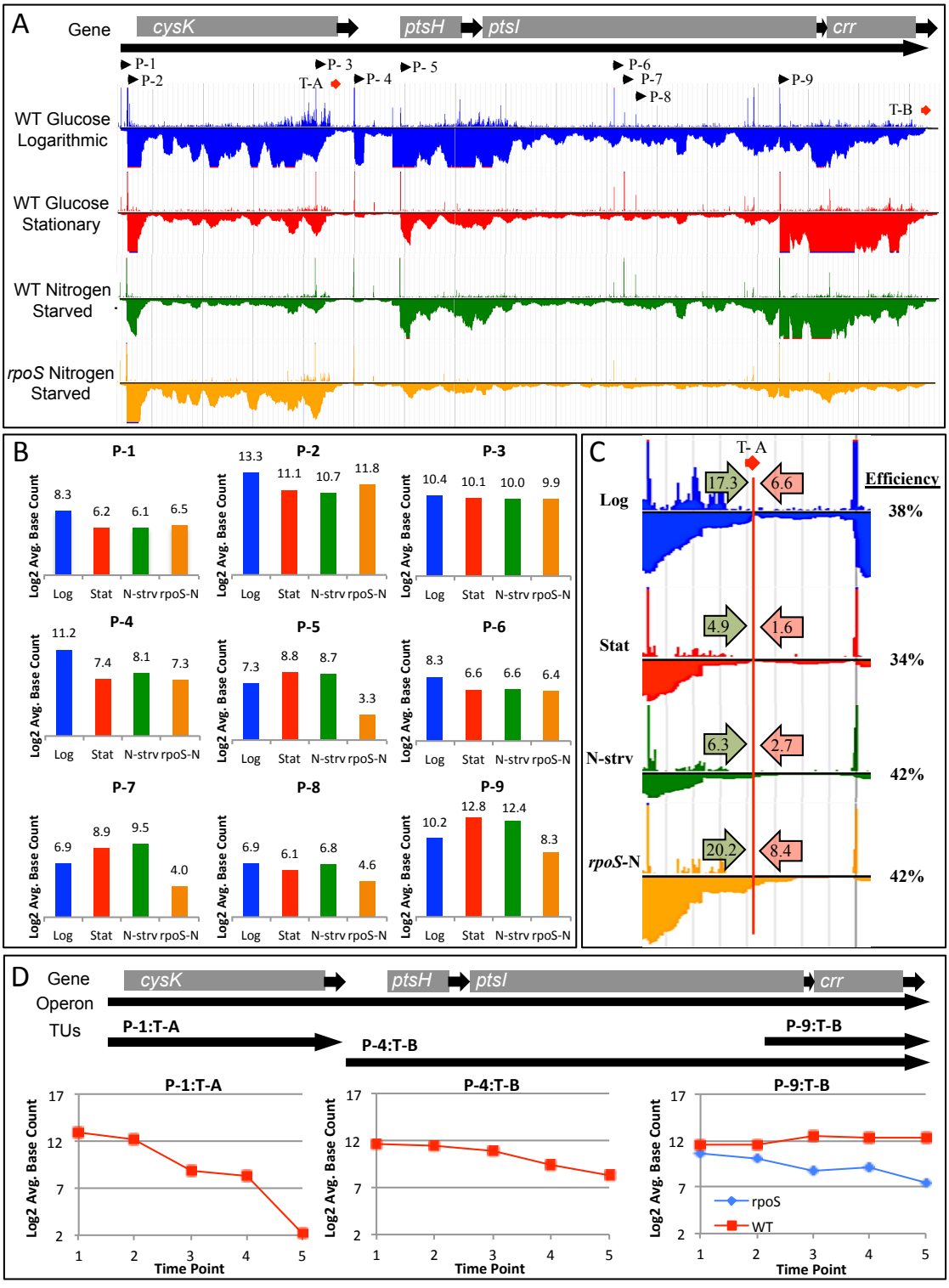


Figure 3-1: Transcriptional feature map and analysis of the *cysK-ptsHI-crr* operon. The dRNA-seq data are available at GEO, GSE58556. (A) The genes and feature locations are drawn to scale and annotated to the positive strand of the *E. coli* MG1655 U00096.3 reference genome. Promoters (P) are indicated by an arrow and are numbered in order from left to right on the positive strand. Terminators (T) are indicated by a diamond. The base count data, consisting of

TEX-treated samples pointing up and unenriched coverage data (fragmented RNA not treated with TEX) pointing down, are visualized in J-Browse [14], as described previously [4]. Only positive strand data are shown. Tracks: wild type (WT), glucose-grown *E. coli* K-12 in logarithmic phase (blue track); WT in stationary phase, 30 min after exhaustion of glucose (red track); WT starved for nitrogen (green track); and an isogenic *rpoS* mutant starved for nitrogen (tan track). The base count scale (on the left) is from 0 to 100, with values exceeding 100 indicated by dark red. (B) The relative activities of the nine promoters is plotted in the graphs as log2 average counts of the first 10 transcribed bases under the four different growth conditions, which are colorized as above. (c) The decrease in average counts of the 25 bases before and after the terminator T-A are shown by light green and pink arrows. (D) Time series analysis of the relative expression levels of three transcripts within the complex *cysK-ptsHI-crr* operon is plotted as the log2 average counts of bases from the indicated promoters to terminators, as described previously [4]. Time point 1 is during middle logarithmic phase, time point 2 is immediately prior to entry into stationary phase, time point 3 is 15 min after entry into stationary phase, time point 4 is 30 min after entry into stationary phase, and time point 5 is 180 min after entry into stationary phase. Additional details of the analysis are described in the text.

Two promoters, P-4 and P-5, which are located within the *cysK-ptsH* intergenic region, drive transcription of *ptsHI-crr*. P-4 is approximately 15 times more active in logarithmic phase than it is under the other three conditions (Fig. 3-1B). On the other hand, P-5 is induced (2.5-fold) in stationary phase and nitrogen-starved conditions by comparison to logarithmic phase and its activity is *rpoS*-dependent, as indicated by a 40-fold reduction in promoter activity by comparison to the wild type under the same conditions (Fig. 3-1B). The transcripts originating from these two promoters apparently are terminated at T-B, downstream of *crr* (Fig. 3-1C). The collective activities of P-4 and P-5 correlate well with the modest decline in average base counts of the P-4:T-B (*ptsHI-crr*) transcript upon entry into stationary phase (Fig. 3-1D). Within the *ptsI* gene are three closely spaced promoters (P-6, P-7, and P-8) that are of relatively low activity compared with the others (Fig. 3-1B). P-6 is expressed approximately equally in the four conditions, P-7 is induced in stationary phase and nitrogen-starved conditions and

is RpoS-dependent, and the least active of the three, P-8, is also dependent RpoS. It does not appear that these three promoters contribute to transcription of the downstream *crr* gene, as indicated by a lack of change in the unenriched base counts visualized in Fig. 3-1A, and so these promoters could also generate pervasive transcripts. On the other hand, P-9 is highly active in stationary phase and nitrogen-starved conditions, is RpoS-dependent, and is located near the 3' end of *ptsI* (Fig. 3-1B), where it apparently drives expression of a *crr* specific transcript (Fig. 3-1A).

Time series analysis shows that the three major transcripts within the operon are differentially expressed during growth and entry into stationary phase (Fig. 3-1D). The *cysK*-specific transcript is expressed at high levels during logarithmic phase and its level declines rapidly during stationary phase. Hence expression of *cysK* reflects the decline in P-1 and P-2 promoter activity in stationary phase and nitrogen-starved conditions. The *ptsHI-crr* transcript level declines little during the first 30 min of stationary phase and then declines modestly 3 hours into stationary phase (Fig. 3-1D), probably because P-4 is less active and P-5 is induced upon entry into stationary phase (Fig. 3-1B). Expression of the *crr* transcript is partially dependent on read-through from promoters within *ptsH* and *ptsI*, and there is no evidence from the base counts to indicate that there is termination within the *ptsI-crr* intergenic region. The *crr*-specific transcript level increases upon entry into stationary phase in the wild type, yet declines in an RpoS-dependent manner in the *rpoS* mutant (Fig. 3-1D). Indeed, P-9 is RpoS dependent, as indicated by 16-fold higher expression in the wild type starved for nitrogen compared to the *rpoS* mutant, and it has a -10 promoter element with the base sequence (CTAnnnTTAA) that is characteristic of RpoS promoters [47].

The primary goal of many RNA-seq experiments is to determine differential gene expression between growth conditions and treatments [9,19,27,32,48-52]. Typically these experiments involve calculating for control and test conditions the number of reads that map to the genome between the start and stop codons of individual genes. Similarly, differential expression of operons can be determined by calculating the average base counts between the promoters and terminators. Since the average operon contains 2 genes, plus intragenic sequences, and 5' and 3' UTRs, there is significantly more information used (more bases) to compute the operon expression level than what is available to represent expression of individual genes. So, the statistical significance of differential expression can be greatly enhanced by using normalized base count data to measure relative operon or transcript expression levels. Differential transcription of operons is readily accomplished by employing algorithms such as DEseq [48] to compute the differential expression and statistics.

Challenges

Massive amounts of RNA-seq data can now be readily obtained. Precise mapping of transcriptional features, logical organization of the annotated data, and meaningful feature quantitation are key to maximizing the value of the resulting transcriptomes. Critical analysis of dRNA-seq data is needed to minimize the number of false positive promoters annotated. Thus it is necessary not only to properly replicate dRNA-seq experiments, but also to augment the analysis with information to corroborate that a predicted TSS is indeed a functional promoter, such as by promoter motif analysis and RNA polymerase binding assays. It would be useful if future

advances in TSS mapping technology include methods to directly label the nucleotides corresponding to TSSs, rather than simply enriching for them. Mapping of 3' transcript ends is an even larger issue and there is a real need for technology that directly labels the 3' ends generated by transcription termination. Perhaps in vitro poly(A) tailing of the 3' ends of RNA prior to fragmentation, followed by sequencing from that end would be helpful. However, it appears from existing RNA-seq data that termination is not a precise biological process and transcripts do not stop at a single nucleotide. For the time being, the state-of-the-art for 3' transcript end mapping remains consensus between replicates.

Lastly, it is important to determine whether "pervasive transcription", defined as TSSs in non-canonical locations [53], is real and if such transcripts have a functional role. Pervasive transcription is seen in yeast, mammals, and fruit flies [54,55] and is frequently observed in viruses and bacteria [32,56,57]. So, there seems to be little doubt that pervasive transcription is real. As to whether pervasive transcripts are functional, that topic was recently reviewed, but it is too early to be sure [53]. The finding that some pervasive transcripts in herpesvirus decreased viral protein production [56] suggests that the functional role of such transcripts should be investigated in bacteria. It is becoming apparent that H-NS and NusG suppress some pervasive transcripts [57,58]. Several potential examples of pervasive transcription can be seen in Fig. 3-1. Using a conservative approach we previously mapped 4 promoters to the *cysK-ptsHI-crr* operon [4]. However, dRNA-seq revealed 9 promoters that map to the operon (Fig. 3-1A), only 4 of which appear to drive transcription of the corresponding genes (P-2, P-4, P-5, and P-9). The other 5 include a weak promoter upstream of the major promoter in front of

cysK and a relatively strong promoter located within the *cysK* coding region and just upstream of the terminator that is intergenic to *cysK-ptsH*. Neither of these promoters appears to contribute to transcript expression levels. The remaining 3 putative pervasive promoters are located within the *ptsI* gene, have relatively low activity levels, and yet all have reasonably well conserved -10 promoter sequence elements, including two that have RpoS promoter motifs and appear to be RpoS-dependent. If these turn out to be real promoters, and there is no reason to think they are not, then the number of promoters on bacterial genomes is being underestimated by perhaps two-fold [9,32].

Acknowledgements

Research in the authors' laboratory was funded by the NIH (GM095370).

References

1. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq**. *Curr Opin Microbiol* 2010, **13**:619-624.

2. **Sharma CM, Vogel J: **Differential RNA-seq: the approach behind and the biological insight gained**. *Curr Opin Microbiol* 2014, **19**:97-105.

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

4. **Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, et al.: **Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing**. *MBio* 2014, **5**:e01442-01414.

5. Li S, Dong X, Su Z: **Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling**. *BMC Genomics* 2013, **14**:520.

6. Neidhardt FC, Bloch PL, Smith DF: **Culture medium for enterobacteria**. *J Bacteriol* 1974, **119**:736-747.

7. *Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G: **Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in *Escherichia coli***. *J Bacteriol* 2014.

8. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, et al.: **IVT-seq reveals extreme bias in RNA sequencing**. *Genome Biol* 2014, **15**:R86.

9. **Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, et al.: **An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium**. *Cell Host Microbe* 2013, **14**:683-695.

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

11. Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J: **How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?** *BMC Genomics* 2012, **13**:734.

12. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357-359.

13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.

14. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome Res* 2009, **19**:1630-1638.

15. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy T, Taylor J, Nekrutenko A: **Dissemination of scientific software with Galaxy ToolShed**. *Genome Biol* 2014, **15**:403.

16. 16. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets**. *Bioinformatics* 2010, **26**:2204-2207.

17. Goecks J, Nekrutenko A, Taylor J, Galaxy T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**:R86.

18. *Forstner KU, Vogel J, Sharma CM: **READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data**. *Bioinformatics* 2014.

19. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al.: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Briefings in bioinformatics* 2012.

20. *Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM: **High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates**. *PLoS Genet* 2013, **9**:e1003495.

21. Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN: **Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis**. *BMC Genomics* 2014, **15**:684.

22. Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BO: **Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling**. *PLoS Genet* 2012, **8**:e1002867.

23. *Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, et al.: **The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium**. *Proc Natl Acad Sci U S A* 2012, **109**:E1277-1286.

24. *Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP: **Conservation of transcription start sites within genes across a bacterial genus**. *MBio* 2014, **5**:e01398-01314.

25. Behrens S, Widder S, Mannala GK, Qing X, Madhugiri R, Kefer N, Mraheil MA, Rattei T, Hain T: **Ultra Deep Sequencing of *Listeria monocytogenes* sRNA Transcriptome Revealed New Antisense RNAs**. *PLoS One* 2014, **9**:e83979.

26. *Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH: **Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis***. *PLoS One* 2012, **7**:e43350.

27. *Soutourina OA, Monot M, Boudry P, Saujet L, Pichon C, Sismeiro O, Semenova E, Severinov K, Le Bouguenec C, Coppee JY, et al.: **Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile***. *PLoS Genet* 2013, **9**:e1003493.

28. *Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H: **RNA-Seq of *Bacillus licheniformis*: active regulatory RNA features expressed within a productive fermentation**. *BMC Genomics* 2013, **14**:667.

29. Deana A, Celesnik H, Belasco JG: **The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal**. *Nature* 2008, **451**:355-358.

30. Bischler T, Kopf M, Voss B: **Transcript mapping based on dRNA-seq data**. *BMC Bioinformatics* 2014, **15**:122.

31. Jorjani H, Zavolan M: **TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data**. *Bioinformatics* 2014, **30**:971-974.

32. **Lin YF, A DR, Guan S, Mamanova L, McDowall KJ: **A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes,* a major contributor to wide-spread human disease**. *BMC Genomics* 2013, **14**:620.

33. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics* 2011, **27**:1017-1018.

34. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA: **Characterization of 582 natural and synthetic terminators and quantification of their design constraints**. *Nat Methods* 2013, **10**:659-664.

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

36. Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R: **Rho directs widespread termination of intragenic and stable RNA transcription**. *Proc Natl Acad Sci U S A* 2009, **106**:15406-15411.

37. Taylor K, Hradecna Z, Szybalski W: **Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA**. *Proc Natl Acad Sci U S A* 1967, **57**:1618-1625.

38. Piette J, Cunin R, Boyen A, Charlier D, Crabeel M, Van Vliet F, Glansdorff N, Squires C, Squires CL: **The regulatory region of the divergent *argECBH* operon in *Escherichia coli* K-12**. *Nucleic Acids Res* 1982, **10**:8031-8048.

39. Wek RC, Hatfield GW: **Nucleotide sequence and in vivo expression of the *ilvY* and *ilvC* genes in *Escherichia coli* K12. Transcription from divergent overlapping promoters**. *J Biol Chem* 1986, **261**:2441-2450.

40. Nomura T, Aiba H, Ishihama A: **Transcriptional organization of the convergent overlapping *dnaQ-rnh* genes of *Escherichia coli***. *J Biol Chem* 1985, **260**:7122-7125.

41. Sameshima JH, Wek RC, Hatfield GW: **Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli***. *J Biol Chem* 1989, **264**:1224-1231.

42. Fortino V, Smolander OP, Auvinen P, Tagliaferri R, Greco D: **Transcriptome dynamics-based operon prediction in prokaryotes**. *BMC Bioinformatics* 2014, **15**:145.

43. *McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B: **Computational analysis of bacterial RNA-Seq data**. *Nucleic Acids Res* 2013, **41**:e140.

44. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2014, **42**:D32-37.

45. De Reuse H, Danchin A: **The *ptsH*, *ptsI*, and *crr* genes of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system: a complex operon with several modes of transcription**. *J Bacteriol* 1988, **170**:3827-3837.

46. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, et al.: **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic Acids Res* 2013, **41**:D203-D213.

47. Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R: **Genome-wide analysis of the general stress response network in *Escherichia coli:* sigmaS-dependent genes, promoters, and sigma factor selectivity**. *J Bacteriol* 2005, **187**:1591-1603.

48. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome biology* 2010, **11**:R106.

49. Balasubramanian D, Kumari H, Jaric M, Fernandez M, Turner KH, Dove SL, Narasimhan G, Lory S, Mathee K: **Deep sequencing analyses expands the *Pseudomonas aeruginosa* AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response**. *Nucleic Acids Res* 2014, **42**:979-998.

50. Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT: **Differential expression analysis of RNA-seq data at single-base resolution**. *Biostatistics* 2014.

51. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nature biotechnology* 2012.

52. Wagner GP, Kin K, Lynch VJ: **A model based criterion for gene expression calls using RNA-seq data**. *Theory Biosci* 2013, **132**:159-164.

53. *Wade JT, Grainger DC: **Pervasive transcription: illuminating the dark matter of bacterial transcriptomes**. *Nat Rev Microbiol* 2014, **12**:647-653.

54. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al.: **Diversity and dynamics of the Drosophila transcriptome**. *Nature* 2014.

55. Jensen TH, Jacquier A, Libri D: **Dealing with pervasive transcription**. *Mol Cell* 2013, **52**:473-484.

56. Canny SP, Reese TA, Johnson LS, Zhang X, Kambal A, Duan E, Liu CY, Virgin HW: **Pervasive transcription of a herpesvirus genome generates functionally important RNAs**. *MBio* 2014, **5**:e01033-01013.

57. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC: **Widespread suppression of intragenic transcription initiation by H-NS**. *Genes Dev* 2014, **28**:214-219.

58. Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R: **Rho and NusG suppress pervasive antisense transcription in Escherichia coli**. *Genes Dev* 2012, **26**:2621-2633.

Chapter 4: RpoS Dependent High-resolution Map of Bacterial Operon Structure Revealed by RNA-seq

Chapter Summary

Escherichia coli is capable of coping with significant changes in environmental conditions, and it adapts to these changes by modulating gene expression through the use of sigma factors. In *E. coli,* gene expression is quickly altered between exponential and stationary phases of growth by RpoD and RpoS sigma factors respectively. Under nearly all growth conditions RpoD regulates the majority of gene expression, however when environmental conditions change, thereby inducing stress, RpoS becomes the prevailing sigma factor and initiates "the general stress response". The RpoS sigma factor plays an important role in the survivability of *E. coli,* and as such elucidating the entirety of the RpoS regulon is of critical importance.

In order to identify and characterize RpoS-dependent operons, genes and promoters under carbon, phosphate and nitrogen starvation, we utilized RNA-seq and dRNA-seq methodologies. RpoS-dependency was identified using DEseq software. Following differential expression analysis, only transcription units, genes and promoters that were statistically significant (p-value ≤ 0.05) and demonstrated a 4-fold or greater change in expression were classified. As a result of our analysis 315 operons, 317 genes, and 278 promoters were classified as RpoS-dependent. These findings are far fewer than were predicted. It was also observed that RpoS-dependency was most impactful when the culture was starved for carbon. Carbon starvation accounted for two-times as many differentially regulated transcription units than nitrogen or

phosphate starvation. Additionally, four new transcripts were identified within the intergenic regions of the genome, and a significant difference in the structure of RpoS-dependent verses independent transcripts was observed. It was observed that most RpoS-dependent operons are monocistronic and approximately half the size of RpoS independent operons. These results and conclusions describe RpoS-dependency at the operon, gene and promoter levels, and elucidate the expansion of the "core" of the RpoS regulon under three different starvation conditions.

Introduction

Escherichia coli, like many bacteria, is adept at exploiting the nutrient resources of a habitat, and as such *E. coli* will experience substantial population growth when environmental conditions are suitable. Alternatively, when resources such as carbon, nitrogen, or phosphate become limited, the growth rate slows and eventually stops. If the resources are limited for a prolonged period of time, death ensues. While it appears that bacteria are more vulnerable to environmental conditions than other species, like mammals, this does not mean that bacteria have not evolved mechanisms to cope with the detrimental effects of starvation or other environmental stressors. Because bacteria are constantly faced with the challenge of coping with changing environmental conditions, they experience a high degree of evolutionary selective pressures. In response to this selection, bacteria have evolved mechanisms of gene expression, described as adaptive modulation, where groups of genes are coordinately regulated in order to respond to environmental stresses and starvation (1). In *E. coli,* gene expression modulation is made possible by one of seven sigma factors that form a complex with

RNA polymerase (RNAP) to orchestrate gene expression needed to adapt to the change in environment (1).

The change in gene expression from rapid growth to stationary phase and back again is regulated by RpoD and RpoS sigma factors, respectively. RpoD is considered the "housekeeping" sigma factor, and is responsible for the majority of gene expression under rapid growth conditions (2). Alternatively, RpoS is the bacterial sigma factor responsible for integrating environmental stress signals and coordinating the change in gene expression termed "the general stress response". RpoS was first discovered as KatF, the regulator of catalase synthesis (KatE) in *E. coli* (3), and was quickly associated with the regulation of a number of other genes. Three years later, Lange and Hengge-Aronis propose that KatF is in fact a sigma factor that "is a central early regulator of the large starvation/stationary phase regulon in *E. coli*" (1).

In *E. coli*, RpoS is under complex regulation at the transcriptional, translational, and post-translational levels (for detailed review see publications by Hengge and Battesti *et. al.*) (4, 5). Transcription of *rpoS* is initiated within the upstream *nlpD* gene (6). This location of the TSS for the *rpoS* transcript results in the formation of a long 5' untranslated region (UTR). RpoS translation is regulated by the formation of a stem-loop within the 5' UTR. When the stem-loop is formed translation cannot occur. This inhibitory secondary structure is overcome when *trans*-encoded a small RNA DsrA anneal to the 5' UTR, linearizes the stem-loop, and exposes the ribosomal-binding site. Alternatively, sRNA OxyS negatively regulates the translation of RpoS, by a mechanism that is not fully understood. Once RpoS is translated, a post-translational form of regulation controls the rate of its degradation. In exponential phase the half-life

of RpoS is approximately 1.5 minutes, in contrast to stationary phase where it is greater than 20 minutes (7, 8). This complex regulation serves to elevate RpoS levels in *E. coli* when it is stressed and helps to ensure survival by altering gene expression until conditions are again favorable for growth.

The elucidation and characterization of the entirety of the RpoS regulon has been elusive, primarily due to the significant overlap between RpoS- and RpoD-dependency. The shared evolutionary history of RpoS and RpoD sigma factors has resulted in considerable structural similarity at the protein level (9). A comparison of the promoter binding consensus motifs for both RpoS and RpoD reveals two subtle differences, 1) RpoS lacks a conserved -35 region and 2) the -10 region of RpoS possesses a cysteine in the "extended -10" region while this is absent in the RpoD motif. These minor differences are not substantial enough to prevent crossover between RpoD- and RpoS- dependent gene expression. Recent studies into RpoS-dependent regulation have revealed two perplexing observations: 1) the existence of a set of genes that are RpoS-dependent and expressed in exponential phase growth (10, 11), and 2) negative regulation of genes associated with the tricarboxylic acid cycle and flagella biosynthesis (12, 13). In addition, there exists a point within the growth curve of *E. coli*, approximately two generations prior to stationary phase, where both RpoS- and RpoD-dependent genes are expressed in parallel. When viewed in totality, this has led many to consider RpoS- and RpoD- dependency as a continuum rather than an absolute, and emphasizes how critical the identification of the RpoS regulon is to fully understanding the function of RpoS.

The RpoS regulon has been investigated for the better part of three decades, and with each new analytical method employed a greater level of detail is revealed. The seminal work by Matin, using two-dimensional gel analysis, identified a set of proteins that responded to carbon, phosphate, and nitrogen starvation, and stimulated further investigation into bacterial stress responses (14). There have been many attempts to elucidate the gene systems under control of RpoS, and by the mid-1990s DNA microarray studies of wild type (WT) and mutant *E. coli* strains led to the identification of several hundred genes that were dependent on RpoS, establishing the RpoS regulon (15). Since then there have been many attempts to elucidate the gene systems under control of RpoS, and over the years the number of genes attributed to the RpoS regulon has increased as a function of the stress conditions studied. It is currently thought that the RpoS regulon consists of more than 500 genes or ~10% of the *E. coli* genome (15). The exact number of genes within the RpoS regulon is a matter of debate, and this uncertainty is compounded by the overlap with other specific stress responses that are actually controlled by alternative sigma factors, like heat (RpoH) and envelope stress (RpoE). At the core of the RpoS regulon are a set 140 genes that are induced in response to all stress conditions tested (15). With the development of RNA-seq analysis, it appears that the elucidation of the fully RpoS regulon is a plausible reality.

For proteobacteria like *E. coli*, the sigma factor RpoS serves to regulate gene expression during the transitions between exponential and stationary phases of growth, and though it is not an essential gene it is clearly important in the colonization of novel habitats (16). The function of the RpoS regulon is expanding with continual research, and with a greater level of understanding comes the realization that RpoS regulated

genes are responsible for biologically impactful changes in phenotypes like biofilm formation (17). In addition, a recent review of pathogenic proteobacteria examined the role that RpoS plays in the infection and colonization processes of pathogens like *E. coli (18)*. While the studies investigating the function of RpoS in pathogenicity of proteobacteria have been inconsistent, the hypothesis persists. Whether the increased pathogenicity of certain proteobacteria is due to RpoS regulation of a virulence factor or because other protective genes within the regulon, like *katE*, are slowing the immune response is still debated. Elucidating the RpoS regulon is essential for understanding, biofilm formation, pathogenicity, and survivability of *E. coli.* Here we explore the RpoS regulon of *E.coli* under carbon, nitrogen, and phosphate starvation conditions and in WT, Δ*rpoS,* Δ*glnG,* and Δ*phoB* mutants. This study significantly increased the number of RpoS-dependent genes within the regulon to now include novel small RNAs and *cis*-encoded antisense RNAs.

Methods and Materials

Bacterial Strains

E. coli BW39452 (Δ*rpoS*:cat), BW39450 (Δ*phoB*::cat), and MG1655 (Δ*glnG*::cat) were constructed from the wild-type (WT) strains BW38028 and MG1655, respectively, using the protocol described by Datsenko and Wanner (19). *E. coli* BW39452 (Δ*rpoS*:cat), BW39450 (Δ*phoB*::cat), MG1655 (Δ*glnG*::cat) and WT BW38028 were grown separately on lysogeny broth (LB) agar plates overnight from viable frozen stock cultures. Single colonies from the LB agar plates were used to inoculate 5mL potassium

morpholinopropanesulfonate (MOPS) minimal medium (20) containing 0.05% glucose and were incubated for 16 h (overnight) at 37°C in a 250 rpm shaker.

Fermenter Grown and Culture Conditions

Overnight *E. coli* BW39452 (Δ*rpoS*:cat), BW39450 (Δ*phoB*::cat), and WT BW38028 cultures were used to inoculate at a 1:10,000 dilution separate 2L Braun Biostat® B Fermenters containing 1 L of MOPS minimal medium with 0.2% glucose. To analyze carbon starvation, 0.2% glucose was sufficient to result in the exhaustion of carbon prior to any other nutrient. To establish phosphate starvation conditions, K_2PO_4 was reduced to 0.2 mM from 1.32 mM in the phosphate replete culture. All other fermenter parameters were kept constant: 37°C, 40% O_2 saturation and a pH of 7.4, which was controlled by the addition of 1M NaOH. Growth of the cultures was monitored via spectrophotometry at 600 nm by using a Beckman Coulter DU800 spectrophotometer. Under carbon limiting growth conditions, representative culture samples were extracted from the fermenter at an OD_{600} of 0.4 (middle-log phase) using a homemade sampling device (21), and once again 30 minutes after entry into stationary phase. For phosphate limiting growth conditions, representative culture samples were collected at an OD_{600} of 0.1 for phosphate replete and 1.0 for phosphate starved samples. The culture samples were withdrawn from the fermenter into an equal volume of ice cold RNAlater to prevent RNA degradation. Cells were pelleted by centrifugation at 8000 rpm for 10 minutes, the RNAlater was decanted, and the cell pellets were stored at -80°C until total RNA was extracted.

Flask Grown and Culture Conditions

For flask cultures the inocula were grown as described above. Overnight cultures were used to inoculate at a 1:10,000 dilution 500mL flasks containing 50mL of MOPS minimal medium and the cultures were incubated at 37°C with constant shaking at 250 rpm. *E. coli* BW39452 (Δ*rpoS*:cat) and WT BW38028 were grown under carbon limitation as described above. *E. coli* BW39450 (Δ*phoB*::cat), BW39452 (Δ*rpoS*:cat), and WT BW38028 were grown under phosphate limitation as described above. To analyze nitrogen starvation, NH_4Cl was reduced to 5nM from 20mM in the nitrogen replete culture, and *E. coli* MG1655 (Δ*glnG*::cat), BW39452 (Δ*rpoS*:cat), and WT BW38028 strains were grown under these conditions. Growth was monitored by spectrophotometry and representative culture samples were collected during log phase and stationary phase. Culture samples were pipetted directly into an equal volume of ice cold RNAlater to prevent RNA degradation and allowed to stand on ice for 5 min before being centrifuged at 8000 rpm for 10 minutes. Then the RNAlater was decanted and the cell pellet was resuspended in 1mL of RNAlater before being transferred to a 1.5mL Eppendorf tube. Cells were pelleted once again by centrifugation at 14,000 x g for 5 min, and residual RNAlater was removed. All cell pellets were stored at -80 until total RNA extraction. Subsequent RNA-seq analysis established that replicate fermenter and flasks cultures yielded nearly identical datasets.

Total RNA Extraction using Quiagen RNeasy Rapid RNA Isolation Kit

Prior to total RNA extraction, bacterial cells were stored at -80°C in an equal volume of RNAlater. Each of the samples were thawed on ice and centrifuged at 5000 x

g for 5 minutes at 4°C. The pellet was resuspended in 200uL of bacterial lysis buffer (30 mM Tris·HCl, pH 8.0, 1 mM EDTA and 15 mg/ml lysozyme (Sigma, St Louis, MO, USA)), and incubated at room temperature for 5 minutes. Following cell lysis, total RNA was extracted and purified from all fermenter growth cultures using the RNeasy rapid RNA isolation kit from Qiagen (Qiagen, USA), following the manufacturer's protocols. Additionally, the optional on-column DNA digestion step was performed using DNase I, without modification from the RNeasy rapid RNA isolation kit protocol. The column based purification process did not retain transcripts less than 50 nucleotides. RNA concentrations were measured on a Beckman Coulter DU800 Spectrophotometer and RNA qualities were assessed using the A260 and A280 ratios. In some samples, rRNA was depleted prior to sequencing to reduce the amount of sequenced rRNA. This was accomplished using the MICROBExpress kit (Ambion, Austin, TX, USA) as specified by the manufacturer's protocol. RNA quality was evaluated prior to sequencing on an Agilent 2100 bioanalyzer with RNA 6000 pico chip. RNA-seq analysis indicated that RNA depletion did not affect the transcriptome analysis (21).

Hot-Phenol method for the Extraction and Purification of total RNA

All flask grown culture samples were extracted and purified using the hot-phenol method described by M. Ares in the protocol "Bacterial RNA Isolation" published in the Cold Springs Harbors molecular technique manual (22). The only modifications were the lengthening of the ethanol precipitation step by 15 minutes and performing the incubation on ice rather than at room temperature. Subsequent to RNA

extraction, all samples were treated with DNAse I to remove DNA contamination prior to sequence library preparation. DNAse I treatment was conducted in accordance with the published protocol described by Kroger *et. al.* (23) . The concentrations of the RNA samples were determined using spectrophotometry on a NanoDrop (ND-1000 spectrophotometer), and RNA quality was determined by using a Shimadzu MultiNA microchip (Shimadzu, Japan). RNA-seq analysis of hot-phenol and RNeasy extracted samples determined, that the relative abundance of small RNA molecules was substantially higher in hot-phenol extracted samples.

cDNA Sequencing Library Preparation for SOLiD 4 Sequencing Platform

Sequencing libraries were prepared at the Purdue University Genomics Core Facility using the SOLiD Total RNA-sequencing kit, as described by Conway *et. al*. (21). In short, total RNA samples were fragmented with RNase III (Ambion, AM2290), resulting in approximately 200 base long RNA fragments. Each RNA sample was divided and one half was treated with terminator 5'-phosphate-dependent exonuclease (TEX) (Epicentre, #TER51020) and sequenced to identify transcription start site (TSSs). The other half of the sample was not treated with TEX. Tobacco Acid Pyrophosphatase (Epicentre, #T19050) was then used to repair 5' monophosphate ends and remove 5' triphosphate ends before the adaptor ligation step. Prior to ligation of SOLiD adaptors, total RNA quantity and quality were assessed using the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA).

In order to maximize the sequencing of the 5' and 3' ends of RNA, the SOLiD Total RNA-Seq Kit, a ligation-based chemistry, was used. Adapters specific to the

SOLiD sequencing platform were directly ligated to the RNA. This resulted in the formation of a DNA-RNA hybrid that was reverse transcribed to produce a cDNA library. Second strand synthesis of the cDNA library was achieved by PCR amplification using SOLiD specific primers. All samples were then purified and quantified using a PureLink Micro Kit (Thermo Fisher Scientific, Grand Island, NY) and the Agilent Bioanalyzer respectively. Emulsion PCR was used to clonally amplify the cDNA libraries and the samples were then purified prior to attachment to the flow chips for SOLiD sequencing. The resulting sequencing libraries were sequencing using a paired-end read protocol at Purdue University Genomics Facility, under the direction of Phillip San Miguel, on a SOLiD 4 Genome analyzer.

cDNA Sequencing Library Preparation for Illumina HiSeq Sequencing Platform

Over the course of this project advances in sequencing technology led to a change in sequencing platforms. Therefore, replicate culture samples were sequenced on the Illumina 2000 HiSeq system. Due to the significant increase in sequencing depth between SOLiD and Illumina sequencing platforms, rRNA depletion was not required prior to sequencing on the Illumina platform. All RNA samples were shipped on dry ice to Vertis Biotechnologie AG in Freisng-Weinenstephan, Germany for library preparation and sequencing, as described previously (24). Prior to sequencing library preparation, all of the RNA samples were divided in half and subjected to differential RNA-seq (dRNA-seq) as described by Sharma *et. al*. (25). In brief, one half of the RNA was fragmented with ultrasound consisting of 4 pulses of 30 seconds at 4°C followed by treatment with Antarctic phosphatase. The RNA fragments were then treated with

poly(A) polymerase in order to poly-adenylate the 3' ends, and an RNA adapter was ligated to the 5' phosphate of the RNA. First strand cDNA synthesis was achieved using a poly(dT) primer and reverse transcriptase. Second strand cDNA synthesis was accomplished using Illumina TruSeq primers and resulted in the incorporation of a barcoded 3' and Illumina TruSeq adapter. The remainder of the RNA samples were fragmented via ultrasound and then treated with TEX to enrich for TSSs. The TEX treated samples were treated with Antarctic phosphatase, poly-adenylated, and ligated

Table 4-1: Metadata for RNA-seq samples

Sample Name	Strain	Growth Condition	Growth Curve Collection	RNA Extraction Method	rRNA Depleted	Sequencing Method
WT_glucose_log	BW38028	0.2% Glucose	Mid-log	Hot-phenol	No	Illumina
WT_glucose_stat	BW38028	0.2% Glucose	Stationary	Hot-phenol	No	Illumina
WT_phos_strv	BW38028	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
WT_N_strv	BW38028	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
glnG_N_strv	MG1655(Δ*glnG*)	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
rpoS_N_strv	BW39452(Δ*rpoS*)	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
rpoS_phos_strv	BW39452(Δ*rpoS*)	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
phoB_phos_strv	BW39450(Δ*phoB*)	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
WT_glucose_log	BW38028	0.2% Glucose	Mid-log	Hot-phenol	No	Illumina
WT_glucose_stat	BW38028	0.2% Glucose	Stationary	Hot-phenol	No	Illumina
WT_phos_strv	BW38028	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
WT_N_strv	BW38028	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
glnG_N_strv	MG1655(Δ*glnG*)	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
rpoS_N_strv	BW39452(Δ*rpoS*)	NH_4Cl 5mM	Stationary	Hot-phenol	No	Illumina
rpoS_phos_strv	BW39452(Δ*rpoS*)	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
phoB_phos_strv	BW39450(Δ*phoB*)	0.2mM K_2PO_4	Stationary	Hot-phenol	No	Illumina
WT_30min_rep1-R WT_30min_rep2	BW38028	0.2% Glucose	Stationary	RNeasy	Yes	SOLiD
WT_04_rep1-R WT_04_rep2	BW38028	0.2% Glucose	Mid-log	RNeasy	Yes	SOLiD
rpoS_30min_rep1 rpoS_30min_rep2	BW39452(Δ*rpoS*)	0.2% Glucose	Stationary	RNeasy	Yes	SOLiD

with a 5' RNA adapter. First and second strand cDNA synthesis was prepared as described above. The sequencing libraries were sequenced from the 5' end on an Illumina HiSeq 2000, generating 50 bp reads, with each library yielding approximately 20 million reads. Metadata for all RNA-seq samples can be located in table 4-1.

Sequence Data Processing and Alignment to Reference Genome

SOLiD and Illumina sequencing platforms are dramatically different technologies, and sequence DNA by two different methods. As such, the raw sequence data generated by these platforms also are different, yet the results were still compatible. SOLiD platforms produce two output files for each run, a CSFASTA and QUAL file, while Illumina generates a single FASTQ file. Raw sequence read files were aligned to the *E. coli* MG1655 genome (NC_000913.3) using Bowtie 2 for Illumina data and Bowtie ver. 1.8 for SOLiD (26). For SOLiD data processing, a three-pass strategy was applied. Pass one consisted of aligning perfectly aligned paired-end reads with maximum distance of 350 bases. The next two passes of Bowtie aligned orphan 5' and 3' end reads (those that could not be aligned as paired reads). This three-pass method increased overall mapping efficiency from 10% to 40-60%. Illumina sequence files, FASTQ format, were aligned in a single pass. Bowtie alignment output files for both Illumina and SOLiD data were SAM files, and all SAM files were converted to binary BAM files using the SAMTOOLS software (27).

Using tools freely available in the Galaxy Toolshed (28) or at UCSC Genome Browser (29), the binary BAM files were converted to BigWig files, which are much smaller and therefore more computable. BigWig files consist of strand specific base

counts at each base location and are readily visualized in the genome browser J-Browse (30). An Oracle database was employed to record all annotation of transcriptional features. Read count data for all BigWig files was normalized using a total count approach (31), which expresses each value as the base count per billion bases counted (21).

Annotation of Transcriptional Features

The annotation of operons across the *E. coli* transcriptome was based on three features: 1) the 5' end, 2) the 3' end, and 3) sufficient read alignment between the 5' and 3' ends to justify connecting them (21). Alternatively, TSSs were annotated as orphans, i.e., they were not contained within mapped operons. Next, TSS locating software TSSpredator (32) and TSSer (33) were used to identify putative promoters. Transcription start site locations were mapped using a clickable J-Browse track that linked to an Oracle database. The clickable locations in this track were generated by using an in-house algorithm that identified two-fold increases in read counts between adjacent bases in the TEX enrichment samples. Manual annotation of TSSs was facilitated by displaying the count of only the first base at the 5' end of each TEX-enriched read in a J-Browse track (34). Putative TSSs were identified and each was added to the database and the promoter type was annotated. Each putative TSS was annotated as one of the following: Primary (P): furthermost upstream in an operon, Secondary (S): located downstream of the primary but not within a coding sequence, Internal (I): located within a coding sequence, Antisense (AS) located in the opposite direction of a transcript, and Orphan (O): not associated with or located in the 3' UTR

of an operon. All primary TSSs and those identified by TSSer or Predator were annotated with the understanding that false TSSs would be included. Following annotation all putative TSSs and the associated promoter region were rigorously analyzed to determine biological significance, see "Promoter Analysis" section below.

Next, the 3' ends of transcripts were annotated. A number of challenges made the annotation of 3' ends more difficult: 1) a 3' end enrichment method does not exist, 2) the number of RNA-seq reads declines at the 3' ends of operons, and 3) stem loop structures associated with intrinsic terminators have varying degrees of efficiency. Therefore, 3' ends of transcripts were annotated by searching for the last aligned base that was consistent between replicates. In addition, the intrinsic terminator prediction software TransTermHP was used to add confidence to the annotation calls (35). Once both the 5' and 3' transcript ends were mapped, the annotation of operons was possible.

To map operons across the *E. coli* transcriptome, the annotated primary promoters were connected to the furthermost downstream terminator by forming a connection in the database. Operons with 90% read coverage between the primary promoter and terminator were considered significant. Once the operon was mapped it was a straightforward process to annotate the additional promoters and terminators within the operon.

Differential Expression Analysis using DEseq

The primary goal of this study was to elucidate and characterize the RpoS regulon at the operon, gene, and promoter levels. To accomplish this objective, RNA-seq data from WT and Δ*rpoS E. coli* cultures under various growth conditions were

analyzed for differences in transcript abundance at the operon, gene and promoter level. Differences in transcript abundance, i.e., differential expression, was determined by using DEseq (36). DEseq is a freely available computational algorithm that normalizes and analyzes RNA-seq data, and then uses a binomial distribution to calculate the difference in expression levels between a "Test" condition and a "Control" condition. DEseq reports the differential expression as log2 fold change and outputs the statistical significance of the difference as a p-value.

The strength of DEseq analysis is dependent on the quality of the transcriptome annotation. Because we use base count datasets, DEseq can be executed on any transcriptional features that can be quantified. Since the transcriptome annotation consists of operons 3' ends, and TSSs, the base counts can be averaged across these features. For example, the average operon contains 2 genes (unpublished data) and UTRs at either transcript end, so DEseq analysis of operons offers greater power than analysis of the genes alone and accounts for the true transcript abundance at the operon level. To elucidate and characterize the RpoS regulon, DEseq was performed at the following levels: 1) promoters (TSS plus 9 bases downstream), 2) transcription units (all promoters within an operon paired to all downstream terminators within that operon, where the largest TU is the operon), and 3) genes (annotated gene locations from reference genome annotation). Table 4-2 summarizes these aspects of DEseq analysis and lists the samples that were analyzed, together with the expected outcome for each comparison.

Following differential expression analysis by DEseq, the log2-fold change and p-value data were evaluated for each "test" and "control" pairing. In an effort to be

conservative, only TUs, genes, and promoters that were statistically significant, i.e., having a p-value of 0.05 or less, and a 4-fold or greater change in expression were classified as RpoS-dependent. All subsequent analysis was performed utilizing only the TUs, genes, and promoters classified as RpoS-dependent by DEseq analysis.

Table 4-2: Differential Expressions Analysis Pairings

Test	Control	Explanation	Sequencing Method
WT_30min_rep1-R WT_30min_rep2	WT_04_rep1-R WT_04_rep2	Stationary Phase Inducible Genes and TUs	SOLiD
WT_glucose_stat	WT_glucose_log	Stationary Phase Inducible Genes and TUs	Illumina
WT_30min_rep1-R WT_30min_rep2	rpoS_30min_rep1 rpoS_30min_rep2	RpoS-dependent Stationary Phase Inducible Genes and TUs	SOLiD
WT_phos_strv	WT_glucose_log	Phosphate Inducible Genes and TUs	Illumina
WT_N_strv	WT_glucose_log	Nitrogen Inducible Genes and TUs	Illumina
WT_glucose_stat	rpoS_phos_strv	RpoS-dependent Phosphate Inducible Genes and TUs	Illumina
WT_glucose_stat	rpoS_N_strv	RpoS-dependent Nitrogen Inducible Genes and TUs	Illumina
WT_30min_rep1-R WT_30min_rep2	WT_04_rep1-R WT_04_rep2	Stationary Phase Inducible Genes and TUs	SOLiD
WT_glucose_stat_TEX	WT_glucose_log_TEX	Stationary Phase Inducible Promoters	Illumina
WT_30min_rep1-R_TEX WT_30min_rep2_TEX	rpoS_30min_rep1_TEX rpoS_30min_rep2_TEX	RpoS-dependent Stationary Phase Inducible Promoters	SOLiD
WT_phos_strv_TEX	WT_glucose_log_TEX	Phosphate Inducible Promoters	Illumina
WT_N_strv_TEX	WT_glucose_log_TEX	Nitrogen Inducible Promoters	Illumina
WT_glucose_stat_TEX	rpoS_phos_strv_TEX	RpoS-dependent Phosphate Inducible Promoters	Illumina
WT_glucose_stat_TEX	rpoS_N_strv_TEX	RpoS-dependent Nitrogen Inducible Promoters	Illumina
WT_30min_rep1-R_TEX WT_30min_rep2_TEX	WT_04_rep1-R_TEX WT_04_rep2_TEX	Stationary Phase Inducible Promoters	SOLiD

Promoter Analysis

Essential to the elucidation of the RpoS regulon is the identification and analysis of the promoters that are driving RpoS-dependent gene expression. Annotation of the *E. coli* transcriptome was a manual process that was aided by TSSpredator and TSSer software. As a result of manual annotation, 11,291 putative promoters were identified.

The quality of each putative promoter was evaluated using three metrics: 1) an increase in coverage reads (non-TEX treated) following a TSS; 2) promoter motif analysis via FIMO software; and 3) promoter activity (the number of sequence reads aligned to the 5'-end of the transcript). The three criteria listed above have been used previously to describe and explain variation in transcript abundance between TUs.

When combined with one another, these metrics provide an effective method for evaluating promoter quality. In order to determine if an increase in coverage was observed at a putative promoter, the 9 base average counts upstream and downstream of the TSS were expressed as a ratio, exemplified by the equation X=(Average(9 bases upstream of TSS))/(Average(9 bases downstream of TSS)). The 50 base pair sequences immediately upstream of each putative TSS were analyzed using the Find Individual Motif Occurrences (FIMO) software associated with the MEME suite. Each 50bp promoter region was screened against a library of consensus motifs for *E. coli* sigma factors. Finally, sequence reads abundance and consensus among replicates was determined. The average of the 9 bases downstream of the TSS was calculated and the resulting values were compared across all samples that originated that a given TSS.

Results

Operon level elucidation of the RpoS Regulon of E. coli

To study regulation by the RpoS sigma factor in *E. coli* on a global scale, the transcript abundance of wild type BW38028 and mutant BW39452 (Δ*rpoS*) *E. coli* were evaluated by dRNA-seq under three starvation conditions. All strains were grown to

stationary phase in MOPS minimal medium with limitation for carbon, nitrogen, or phosphate. Following RNA extraction and RNA sequencing, the sequence data was aligned to the *E. coli* MG1655 reference genome using the methods previously described by Conway *et.al. (21)*. The *E. coli* transcriptome was manually annotated in a Jbrowse environment aided by TSS locating software, TSSpredator and TSSer. Following annotation of the transcriptome, RNA-seq samples were compared one to another using DEseq in order to observe statistically significant changes in transcript levels. The most impactful comparisons for the identification of RpoS-dependent operons were often the Δ*rpoS* mutant compared to the WT strain under the identical growth conditions.

A prevailing observation stemming from the global analysis of the *E. coli* transcriptome is the vast diversity of methods by which *E. coli* modulates gene expression in response to environmental signals. During our analysis, it was observed that regulation of transcription was not exclusively located at the primary promoter of operons. In the majority of occurrences, regulation by RpoS was achieved at the operon level, however differential regulation within operons was observed approximately 30% of the time. Figure 4-1 illustrates RpoS regulation at the operon level, and an investigation of the *osmY-ytjA* operon reveals that it is both stationary phase inducible and RpoS-dependent. As can be seen in figure 4-1, transcript abundance in stationary phase far exceeds (~11.5-fold) that observed in logarithmic phase, indicating that the *osmY-ytjA* operon is stationary phase inducible. Additionally, transcript abundance was significantly larger than 4-fold in the Δ*rpoS* mutant, indicating that transcription at primary promoter, P-4611152, is RpoS-dependent. In this example the primary

promoter (P-4611152) determined transcription of the bicistronic operon *osmY-ytjA*. The remaining three downstream promoters, secondary promoters S-4612063 and S-4612121 and internal promoters I-4612233, in this operon did not contribute to the overall transcript abundance of this portion of the *osmY-ytjA* operon.

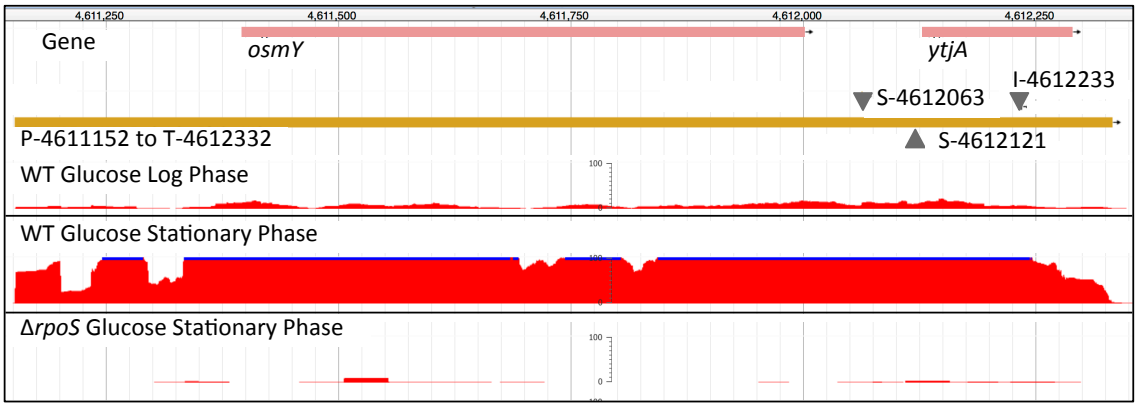


Figure 4-1: Differential expression of RpoS-dependent operon *osmY-ytjA*. This operon has 4 promoters and 1 terminator and contains 4 transcription units created by transcription initiation from secondary and internal promoters. This operon exemplifies the utility of RNA-seq on WT and Δ*rpoS* strains of MG1655 under the same growth conditions. In the absence of the RpoS sigma factor, RpoS-dependent transcripts were not initiated. Blue line across read alignment data indicates greater than 100 reads at that base location.

A global survey of RpoS-dependency revealed that the majority of transcription is regulated at the operon level, but differential expression of genes within an operon also was observed. As seen in figure 4-2A, approximately 70% of RpoS-dependent transcription originates at either the primary or antisense promoter location of operons. In the majority of instances, transcription at either the primary or antisense promoters results in the transcription of all downstream genes within that operon. Alternatively, 30% of transcription occurred at the sub-operon level, either at internal or secondary promoters. An understanding of transcriptional regulation by RpoS can be gleaned from viewing transcription at this level. Within *E. coli*, there exists a set of operons, coding and antisense alike, that are exclusively RpoS-dependent. However, there exists a subset

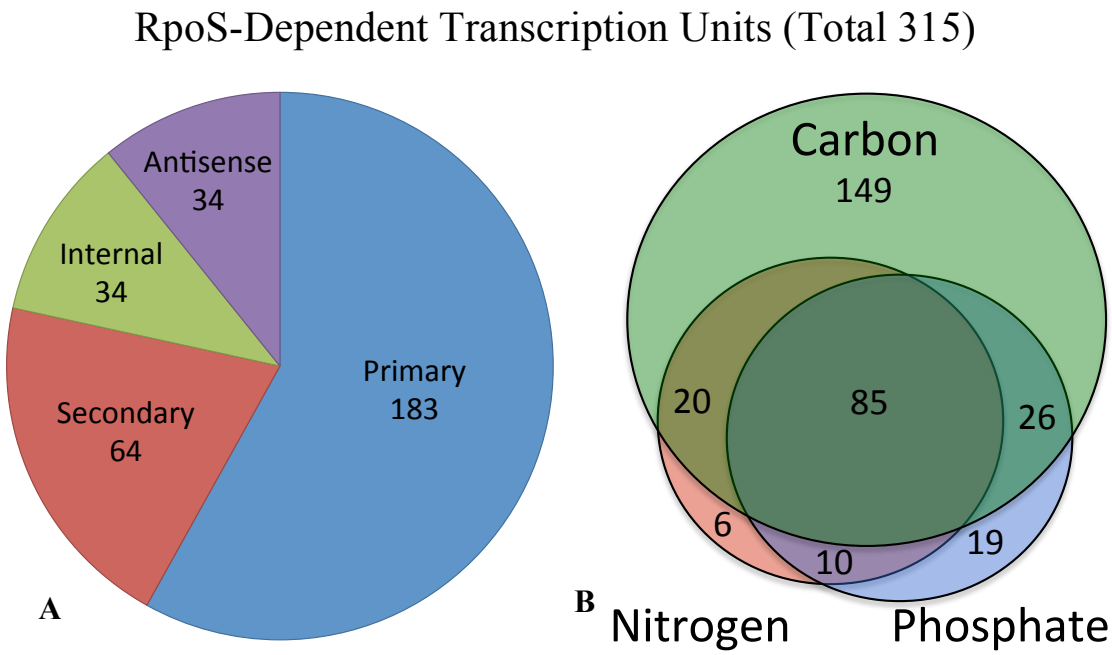


Figure 4-2: RpoS-dependent transcription unit abundance by promoter type and inducible condition. A global survey of RpoS-dependency revealed 315 transcription units that were differentially expressed in WT and Δ*rpoS* strains of MG1655 under the same growth conditions. A) Transcription units originating from primary promoters accounted for 58% of the total dataset. Sub-operon transcription was also observed at secondary (64) and internal (34) promoters. B) Analysis MG1655 (Δ*rpoS*) under nitrogen, carbon, and phosphate starvation reviled variation in the number of statistical significant RpoS-dependent transcription units.

of genes that are regulated by RpoS and RpoD, and in these instances sub-operon level regulation can occur.

Differential RNA-seq analysis of MG1655 (Δ*rpoS*) under nitrogen, carbon, or phosphate starvation revealed the RpoS-dependency for each starvation condition. The Venn diagram in Figure 4-2B illustrates the contribution of all three starvation conditions to the RpoS regulon at the TU level (full list of RpoS-dependent TUs located in Appendix C). It is clear that nitrogen and phosphate starvation are minor contributors to the whole of the RpoS regulon. In fact, only 11% of all TUs were exclusively nitrogen and/or phosphate starvation inducible. Alternatively, of the 35 transcripts

identified as being nitrogen and/or phosphate starvation inducible 4 previously unannotated small RNAs were discovered (average length of 147 bp), demonstrating the value of investigating gene expression under multiple starvation conditions.

Viewing the totality of RpoS-dependent transcription at the operon level, an interesting observation about the RpoS regulon is revealed. Based on our annotation of the *E. coli* transcriptome, a total of 4004 genes are contained within 1796 operons across the entire transcriptome (based on Conway, 2014 and unpublished data). Further more, investigation of the RpoS regulon has identified 368 genes contained within 230 operons. The average number of genes per operon was determined to be 2.2 for all TUs and 1.6 for RpoS-dependent TUs. Additionally, RpoS-dependent TUs were shorter by comparison to all TUs: 1302 bp for RpoS-dependent TUs and 2220 bp for all TUs.

Gene Level Analysis of the RpoS Regulon in E. coli

Differential expression of genes by DEseq was determined by calculating the average transcript abundance between the first and last base of annotated genes. Gene locations were obtained from the *E. coli* MG1655 U00096.3 reference genome available at GenBank. Differential expression at the gene level is the most commonly performed analysis. While it has advantages, there are important drawbacks to not examining transcription abundance at the TU and promoter levels. In this section, I will evaluate search strategies for RpoS-dependent genes, and assess the differences between *de novo* analysis of genes and RpoS-dependent TU directed search strategies.

It is well established that many mRNA transcripts contain 5'- and 3'- UTRs that play an important role in regulating the translation of mRNA into protein. Moreover,

the length and nucleotide sequence of UTRs can aid in understanding the regulatory mechanism. A differential expression search strategy that is limited to evaluating only annotated genes will miss valuable data contained within UTRs, unannotated genes, and many small RNAs. In short, without a properly annotated transcriptome a gene-based strategy cannot evaluate novel genomic features.

The monocistronic operon *pykF,* depicted in figure 4-3, illustrates a challenge that a differentially expressed TU directed search strategy overcomes. It is clear from examination of the data that *pykF* transcription is initiated ~200 bp upstream from the start codon. Further more, it does not appear to be differentially expressed between logarithmic- and stationary phases of growth based on gene annotation alone. A comparison of samples '*WT Glucose Log Phase'* and '*WT Glucose Stationary Phase'* does not display a 4-fold change in transcript abundance between the first and last bases of the *pykF* gene. Alternatively, analysis of all bases between the annotated TSS and

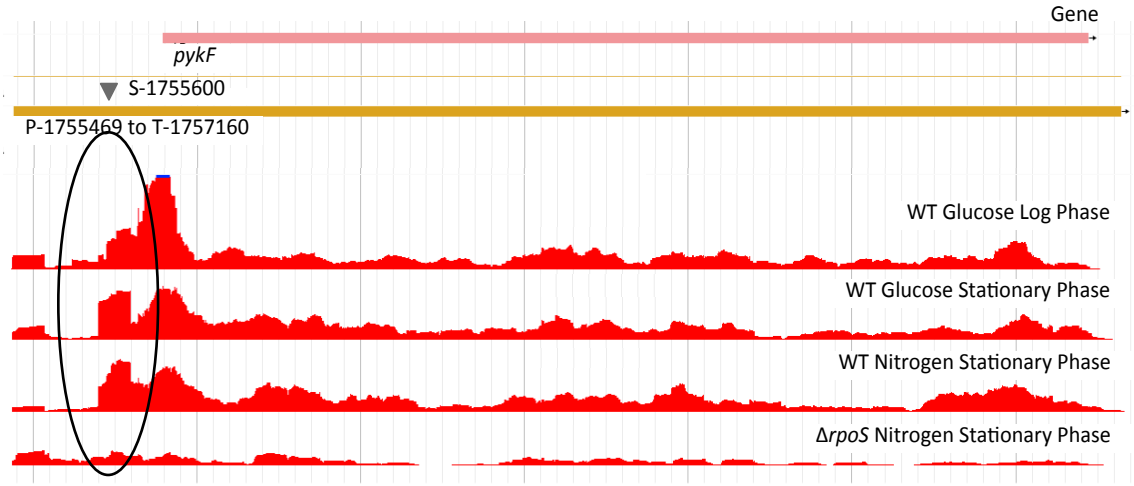


Figure 4-3: Differential expression of RpoS-dependent gene *pykF*. The gene *pykF* is expressed in both logarithmic and stationary growth phase, in addition to being inducible by nitrogen starvation. The *pykF* gene is transcribed by a primary and a secondary promoter. The black circle across all data tracks indicates the decline (greater than 4-fold) in transcript abundance in Δ*rpoS* mutant sample. The secondary promoter at base location 1755600 is RpoS-dependent. A basal level of transcription persists throughout stationary phase due to the activity of the primary promoter at base location 1755469.

terminator for samples '*WT Nitrogen Stationary Phase'* and '*ΔrpoS Nitrogen Stationary Phase'* reveals that the *pykF* operon is RpoS-dependent. Upon closer inspection, it was determined that there are two promoters upstream, a log-phase transcribed primary promoter and a secondary promoter that is both logarithmic- and stationary- phase inducible. Notably, a TU directed search strategy was able to identify 31 additional TUs that display no change in expression between logarithmic- and stationary- phase growth, but are RpoS-dependent.

A comparison of both search strategies for the discovery of differentially expressed RpoS-dependent genes is summarized in figure 4-4. It is evident that neither the annotated gene nor TU directed search strategies are sufficient for identifying all RpoS-dependent genes, and both are required. The annotated gene based search strategy yielded a total of 404 differentially regulated genes (figure 4-4A), the majority of which were identified under carbon and phosphate starvation conditions. Alternatively, a TU

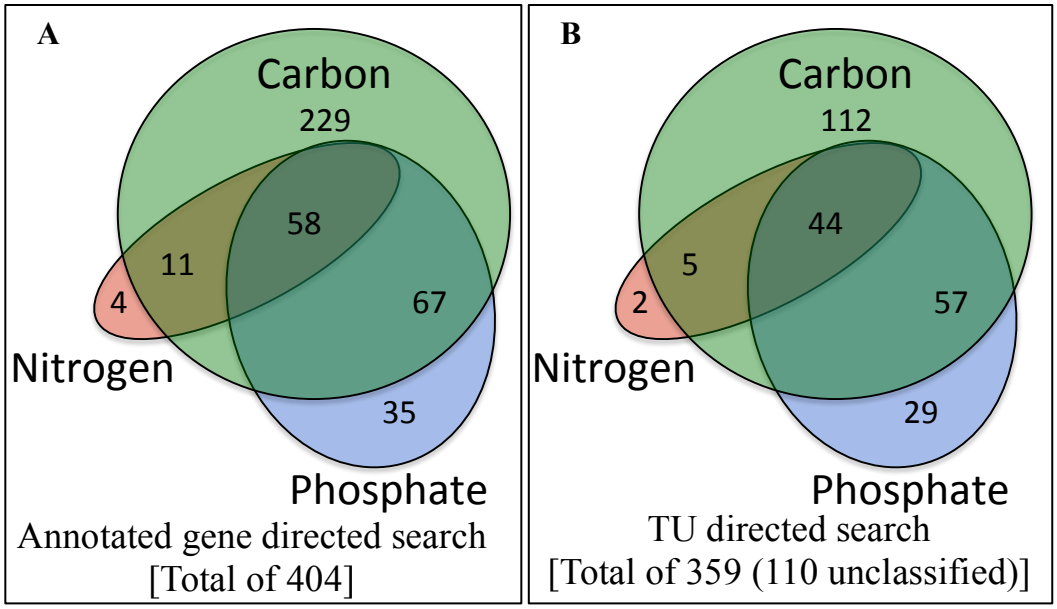


Figure 4-4: RpoS-dependent genes categorized by starvation inducible condition. A) Differential expression analysis identified 404 genes that are RpoS-dependent in one or more growth condition. B) The genes associated with differentially expressed TUs were analyzed further, and classified base on starvation inducible condition.

directed search for differentially expressed RpoS-dependent genes resulted in the identification of 359 genes. However, 110 of these genes contained within differentially expressed TUs did not show a 4-fold change in transcript abundance or were not statistically significant and were therefore not classified. The majority of the 110 unclassified genes found were co-transcribed with another differentially expressed gene that displayed a robust change in transcript abundance between conditions. Manual inspection of all differentally expressed TUs indicated that the analysis of the data from weakly transcribed regions of the genome was difficult to quantify using the gene directed approach (full list of RpoS-dependent genes is located in Appendix C).

Predating the use of dRNA-seq, DNA microarrays were used to investigate the RpoS regulon, and it is only logical that a comparison between the two methods be performed. Based on the microarray studies contained within the *E. coli* Gene Expression Database (GenExpDB), 436 genes were previously characterized as being RpoS-dependent. This set of genes was cross-referenced to the list of RpoS-dependent genes identified by annotated gene- and TU- directed search strategies, figure 4-5. Results indicate that sufficient overlap exists between all three methods, however both search strategies should be utilized to maximize RpoS-dependent gene discovery. Surprisingly, a subset of 97 genes that were identified by microarray were not found by RNA-seq based strategies. It serves to reason, that this is the result of the conservative 4-fold or greater search parameter placed on the RNA-seq data. This is not consistent with the method used for microarray analysis. The list of genes discovered by microarray utilized a 2-fold or greater increase in signal to classify genes as RpoS-dependent. Future RNA-seq studies will be required to resolve these inconsistencies.

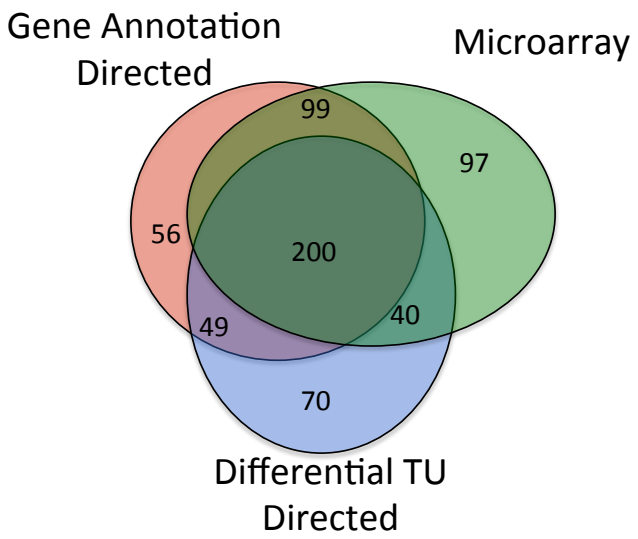


Figure 4-5: Comparison of search strategies for the discovery of RpoS-dependent genes. Venn diagram depicts the union of RpoS-dependent gene discovery by microarray and RNA-seq based approaches.

Promoter Level Analysis of the RpoS Regulon in E. coli

Initiation of transcription is determined by the recruitment of RNA polymerase holoenzyme to the promoter region upstream of the TSS by the binding of sigma factor to conserved nucleotide domains. While the mechanism by which bacterial transcription occurs is well understood, the wide variation in promoter strengths within the same regulon warrants further investigation. Due to the importance of promoters for determining transcription, global analysis of RpoS-dependent promoter activities are vital for understanding the entirety of the RpoS regulon. In order to identify RpoS-dependent promoters, we relied exclusively on the annotated promoter locations obtained from the *E. coli* transcriptome by Conway *et. al.* (21). It was empirically determined that the 9 bases immediately following an annotated TSS were indicative of promoter strength and often were a predictor for transcript abundance across the operon. Therefore, differential expression analysis was performed on the 9 bases following all

annotated TSSs, and once again the results were limited to statistically significant values that exhibited 4-fold or greater change in transcript abundance. In total, 972 promoters were identified in at least one of the starvation conditions tested. However, only 278 of the 972 promoters were determined to be associated with differentially expressed TUs. As can be seen in figure 4-6B, a high number of antisense (AS), internal (I), and orphan (O) promoters were reported as differentially expressed. This large discrepancy in the data, in combination with the elevated incidence of AS, I, and O promoters, highlights the prevalence of pervasive transcription within the *E. coli* transcriptome. In light of the high incidence of pervasive transcription within the dataset, only promoters associated with differentially expressed TUs will be evaluated.

The 278 RpoS-dependent promoters identified were categorized based on their location relative to the annotated operon (primary promoters (P) define the 5'-end of a transcript, secondary (S) are located between genes, internal (I) are within genes,

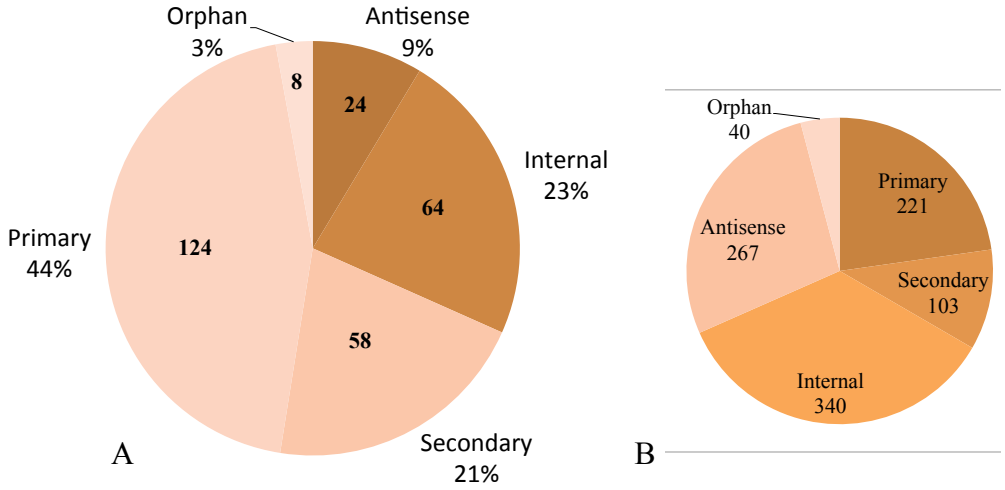


Figure 4-6: Transcription unit directed and *de novo* DEseq analysis for the identification of RpoD-dependent promoters. A) Promoters associated with a differentially expressed TU were classified based on promoter position relative to operon structure. B) Promoters determined by differential expression analysis were classified based on promoter position relative to operon structure.

antisense (AS) are opposite an annotate gene, and orphan (O) which are not associated with robust transcription). In total, 65% of all promoters identified are associated with a differentially expressed TU located directly upstream of a gene (figure 4-6A). A detailed investigation of the transcriptome- annotation and data confirms, that the majority of the RpoS-dependent promoters regulate short operons, consisting of few genes, and often from the primary promoter location.

Consensus analysis was performed on the promoter region for all 278 RpoS-dependent promoters using the bioinformatics software MEME. Figure 4-7A is an illustration of the consensus motif generated by this analysis. As can be seen, the motif

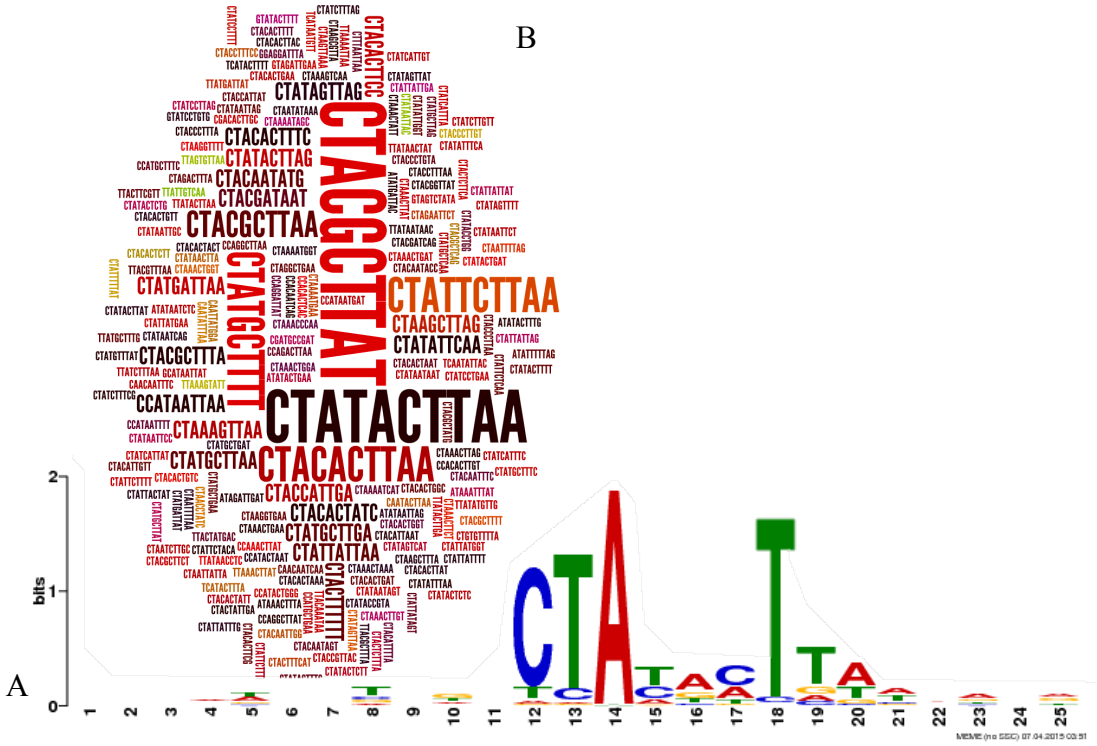


Figure 4-7: Consensus analysis among RpoS-dependent promoters for the -10 region of the sigma factor binding site. A) Consensus motif for all 278 RpoS-dependent promoters identified by differential expression analysis was obtained by employing the bioinformatic software program MEME. B) The relative frequency of occurrence for the -10 region of all 278 RpoS-dependent promoters was determined by use of the online program Wordle, and is intended to be more informational than analytical.

is identical to the ideal RpoS promoter described in the literature (37). Furthermore, comparison of our consensus motif to a database of known motifs, DEInteract, via TOMTOM (a program within the MEME suite of software), a match was returned for RpoS with a p-value of 3.56e-05. Based on this analysis we feel confident that the promoters identified by our differential expression search strategy are RpoS-dependent, and represent the most robust promoters in the RpoS regulon.

Promoter consensus analysis and motif generation are powerful analytical tools for the discovery of conserved domains within promoter regions, however valuable information about the frequency of an individual series of nucleotides is lost. Based on figure 4-7B, it would be assumed that the most common promoter sequences would be CTATACTTAA, however figure 4-7A identifies CTACGCTTAA as the most abundant sequence in the data. While the difference in sequence structure appears minor, the presence of guanine and cytosine at these base locations within the promoter region is sufficient for rendering these promoters exclusively RpoS-dependent. Within the literature there exists a body of work on the concept of a "gearbox" promoter. Gearbox promoters appear as a hybrid between the -10 regions of RpoS and RpoD. An example of a gearbox promoter can be observed in figure 4-7A, CTATACTTAA, and is the second most frequent promoter sequence in the data. Examination of the gearbox promoter sequence reveals the presence of the RpoS and RpoD consensus motif, CTAnnnTTnn and TATnnnTT respectively. Equipped with the knowledge gained by figure 4-7A it becomes apparent why substantial overlap exists between the RpoS- and RpoD- regulons.

Discussion

Differential RNA-seq is a robust and accurate method for the elucidation of the RpoS regulon when coupled with differential expression analysis at the TU, gene, and promoter levels. Our analysis of the RpoS regulon under carbon, nitrogen and phosphate starvation conditions yielded 315 TUs, 359 genes, and 278 promoters that were statistically significant and changed >4-fold in transcript abundance between *E. coli* WT and Δ*rpoS* strains.

At each level of the analysis, insight concerning the events of RpoS-dependent transcription was gained. Promoter level analysis was a valuable indicator of RpoS-dependent transcription initiation, while gene- and TU- level analysis was able to evaluate the abundance of transcripts that were capable of being translated, *i.e.* full length. As such, we were able to observe and subsequently categorize RpoS-dependent gene regulation on a global-scale, to include RpoS-dependent operons that are regulated by multiple promoters in different phases of growth (as seen in figure 4-3), and four newly discovered small RNAs. The enormity of the analysis performed in this study is powerful, but is not without challenges. While global scale investigations, like those performed here, are becoming the norm, the volume of data produced is not readily disseminated within the community. In an effort combat this trend we have converted all annotation calls made on the data to GenBank format using the terms “promoter,” “terminator,” and “operon” as feature keys.

One notable observation that can be drawn from this analysis is the discrepancy in the number of RpoS-dependent TUs, genes and promoters that were identified verses the number predicted. Because of the robust nature of RNA-seq analysis it was our

assumption that more RpoS-dependent transcription features would be observed. Based on the results of this study, we reported 77 fewer TUs and nearly 100 fewer genes than were identified by microarray analysis. Additionally, our comparison to a set of RpoS-dependent genes identified by microarray resulted in 97 genes that were not observed by RNA-seq analysis. It is our opinion that this variation in the data is due to the >4-fold change filter applied to the differential transcription abundance data. A cursory search of the data using a >2-fold change filter, equivalent to what is used in microarray analysis, resulted in the addition of approximately 150 TUs, 100 genes, and 220 promoters. While we are not currently advocating this as a search parameter, the observation of the differences in data sets highlights the need for guidelines concerning the backward compatibility of RNA-seq data to microarray data. Remarkably, analysis of nitrogen and phosphate starvation conditions added little to the number of RpoS-dependent TUs, genes and promoters identified by this study. This supports the concept that RpoS is a "general stress response" and modulates gene expression independently of specific stressors.

An integration of the findings from differential expression at the promoter- and TU- level reveals a set of observations concerning RpoS-dependent gene regulation. It is evident from differential expression analysis at the TU level that the majority of RpoS-dependent operons are either mono- or bicistronic. Combined with the promoter level data, and it becomes clear that there exists a set of operons that are primarily regulated by RpoS exclusively. More over, these operons are shorter in length, have fewer genes contained within them, and are often transcribed from a single promoter.

It is well established that RpoS and RpoD have a shared evolutionary history. It has been determined, based on gene synteny, reciprocal BLAST hit analysis, and insertion/deletion analysis, that the RpoS sigma factor was derived from a duplication of the RpoD gene prior to the divergence of the proteobacteria from its last common ancestor. Following gene duplication, the ancestral RpoS gene underwent a deletion event that resulted in the loss of region 1. This deletion mutation reduced the size of the sigma factor by half. Interestingly, it was determined based on the analysis performed in this study that RpoS-dependent operons are also half the size of their RpoD counterparts, 1302 bp and 2220 bp respectively. More over, the average number of genes per RpoD-dependent operon is ~2.2, while RpoS-dependent operons average 1.6 genes per operon. While these findings could be construed as coincidence, there is inherent value in reflecting upon them further. If it is determined that these observations hold true, then it can be implied that the RpoS regulon is more than the cobbling together of genes from other regulons to deal with environmental stress. Instead, evolutionary selection pressures acted upon *E. coli* to bolster the efficacy of the RpoS regulon through the duplication of genes followed by a reduction in length. This evolutionary model of the RpoS regulon accounts for the size differences observed within our data, and provides an explanation of the evolutionary history for the genes associated with the RpoS regulon.

Acknowledgements

This work was funded primarily by U.S. Public Health Service NIH RC1GM09207 to B.L.W. and T.C. from 2009 to 2011. B.L.W. is currently supported by NSF award 106394. Additional support was from NIH GM095370 to T.C.

References

1. **Lange R, Hengge-Aronis R.** 1991. Identification of a central regulator of stationary-phase gene expression in Escherichia coli. Mol Microbiol **5:**49-59.

2. **Paget MS, Helmann JD.** 2003. The sigma70 family of sigma factors. Genome Biol **4:**203.

3. **Schellhorn HE, Hassan HM.** 1988. Transcriptional regulation of katE in Escherichia coli K-12. J Bacteriol **170:**4286-4292.

4. **Hengge R.** 2009. Proteolysis of sigmaS (RpoS) and the general stress response in Escherichia coli. Res Microbiol **160:**667-676.

5. **Battesti A, Majdalani N, Gottesman S.** 2011. The RpoS-mediated general stress response in Escherichia coli. Annu Rev Microbiol **65:**189-213.

6. **Lange R, Fischer D, Hengge-Aronis R.** 1995. Identification of transcriptional start sites and the role of ppGpp in the expression of rpoS, the structural gene for the sigma S subunit of RNA polymerase in Escherichia coli. J Bacteriol **177:**4676-4680.

7. **Lange R, Hengge-Aronis R.** 1994. The nlpD gene is located in an operon with rpoS on the Escherichia coli chromosome and encodes a novel lipoprotein with a potential function in cell wall formation. Mol Microbiol **13:**733-743.

8. **Lange R, Hengge-Aronis R.** 1994. The cellular concentration of the sigma S subunit of RNA polymerase in Escherichia coli is controlled at the levels of transcription, translation, and protein stability. Genes Dev **8:**1600-1612.

9. **Chiang SM, Schellhorn HE.** 2010. Evolution of the RpoS regulon: origin of RpoS and the conservation of RpoS-dependent regulation in bacteria. J Mol Evol **70:**557-571.

10. **Rahman M, Hasan MR, Oba T, Shimizu K.** 2006. Effect of rpoS gene knockout on the metabolism of Escherichia coli during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations. Biotechnol Bioeng **94:**585-595.

11. **Dong T, Kirchhof MG, Schellhorn HE.** 2008. RpoS regulation of gene expression during exponential growth of Escherichia coli K12. Mol Genet Genomics **279:**267-277.

12. **O'Neal CR, Gabriel WM, Turk AK, Libby SJ, Fang FC, Spector MP.** 1994. RpoS is necessary for both the positive and negative regulation of starvation survival genes during phosphate, carbon, and nitrogen starvation in Salmonella typhimurium. J Bacteriol **176:**4610-4616.

13. **Tsui HC, Feng G, Winkler ME.** 1997. Negative regulation of mutS and mutH repair gene expression by the Hfq and RpoS global regulators of Escherichia coli K-12. J Bacteriol **179:**7476-7487.

14. **Groat RG, Schultz JE, Zychlinsky E, Bockman A, Matin A.** 1986. Starvation proteins in Escherichia coli: kinetics of synthesis and role in starvation survival. J Bacteriol **168:**486-493.

15. **Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R.** 2005. Genome-wide analysis of the general stress response network in Escherichia coli: sigmaS-dependent genes, promoters, and sigma factor selectivity. J Bacteriol **187:**1591-1603.

16. **Stockwell VO, Loper JE.** 2005. The sigma factor RpoS is required for stress tolerance and environmental fitness of Pseudomonas fluorescens Pf-5. Microbiology **151:**3001-3009.

17. **Mika F, Busse S, Possling A, Berkholz J, Tschowri N, Sommerfeldt N, Pruteanu M, Hengge R.** 2012. Targeting of csgD by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in Escherichia coli. Mol Microbiol **84:**51-65.

18. **Dong T, Schellhorn HE.** 2010. Role of RpoS in virulence of pathogens. Infect Immun **78:**887-897.

19. **Datsenko KA, Wanner BL.** 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A **97:**6640-6645.

20. **Neidhardt FC, Bloch PL, Smith DF.** 1974. Culture medium for enterobacteria. J Bacteriol **119:**736-747.

21. **Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL.** 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio **5:**e01442-01414.

22. **Ares M.** 2012. Bacterial RNA isolation. Cold Spring Harb Protoc **2012:**1024-1027.

23. **Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC.** 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A **109:**E1277-1286.

24. **Creecy JP, Conway T.** 2015. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol **23:**133-140.

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

26. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357-359.

27. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.

28. **Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy T, Taylor J, Nekrutenko A.** 2014. Dissemination of scientific software with Galaxy ToolShed. Genome Biol **15:**403.

29. **Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D.** 2010. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics **26:**2204-2207.

30. **Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH.** 2009. JBrowse: a next-generation genome browser. Genome Res **19:**1630-1638.

31. **Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F, French StatOmique C.** 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform **14:**671-683.

32. **Forstner KU, Vogel J, Sharma CM.** 2014. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics **30:**3421-3423.

33. **Jorjani H, Zavolan M.** 2014. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. Bioinformatics **30:**971-974.

34. **Lin YF, A DR, Guan S, Mamanova L, McDowall KJ.** 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in Propionibacterium acnes, a major contributor to wide-spread human disease. BMC Genomics **14:**620.

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

36. **Anders S, Huber W.** 2010. Differential expression analysis for sequence count data. Genome Biol **11:**R106.

37. **Wise A, Brems R, Ramakrishnan V, Villarejo M.** 1996. Sequences in the -35 region of Escherichia coli rpoS-dependent genes promote transcription by E sigma S. J Bacteriol **178:**2785-2793.

Chapter 5: Conclusions and Future Directions

Introduction

The experiments, analytical strategies, and theoretical concepts detailed in this dissertation were performed in an effort to better understand the control of global gene expression in *E. coli*. Novel to these studies has been the use of deep sequencing techniques for the analysis of the total transcriptome of *E. coli.* Utilizing conventional RNA-seq and dRNA-seq methods, I investigated the totality of transcription within many cultures under physiologically relevant conditions and points and phases of growth. The power of this dual approach to total RNA-seq is evident (1). Over the past four years, using the combined weight of more then 25 RNA-seq datasets, Dr. Conway and I have accomplished the following: thousands of new transcriptional features were located, a multitude of transcriptional features previously characterized by others were confirmed, the concept of excludons by complementary convergent and divergent operons was supported, the putative number of excludons was greatly expanded, and the RpoS regulon was characterized. While the work presented here does not fully explain all of the mechanisms of transcriptional regulation, the contribution to the scientific community made by these efforts is notable. Finally, a number of important questions still remain concerning bacterial transcriptomics. In an effort to focus my ongoing investigations in this field of study, I will outline areas of interest that I will pursue in the future.

Chapter 2 Summary

The series of experiments discussed in chapter 2 elucidates the power of high-resolution RNA sequencing data sets for precisely locating transcriptional features and annotating operons across the genome (2). Massive amounts of RNA sequencing data can now be readily obtained. Therefore, precise mapping of transcriptional features, logical organization of the annotated data, and meaningful feature quantitation are key to maximizing the value of the resulting transcriptome data. Single-nucleotide-resolved RNA-seq data offer the best approach to precisely map transcriptional features, and the data presented in chapter 2 were the first to couple RNA-seq techniques with a comprehensive strategy for mapping transcriptional features. this approach revealed a level of transcriptional complexity that was previously uncharacterized in *E. coli*. Our findings allowed for the precise annotation of 2,122 promoters and 1,774 terminators, which defined 1,510 operons with an average of approximately two genes per operon (2). In addition, a large proportion of these operons were complex in nature, possessing internal promoters or terminators that generated multiple transcription units. Differential expression of polycistronic genes within the same operon was also observed, resulting from a host of regulatory mechanisms. In addition, 89 antisense transcripts were also identified. In summation, the transcriptome complexity observed in *E. coli* appears to be a general property of the domain bacteria. However, due to the vital role that *E. coli* plays in the field of molecular biology, a detailed transcriptome map of *E. coli* was vital for the scientific community.

Chapter 3 Summary

Transcriptome analysis by RNA sequencing has emerged as the premier method for evaluating bacterial transcription and transcription regulation. The reduction in the cost of high-throughput sequencing has made transcriptome analysis by RNA-seq a reasonable approach for the majority of research laboratories. However, the truly daunting task remains the analysis of the hundreds of gigabases of data that are obtained from a single study (3, 4).

Vital to the future success of RNA-seq-based transcriptomics studies is the utility of the data for generating biologically insightful conclusions. In chapter 3, I described the methods that were applied to quantitatively analyze RNA-seq data (5). It is my opinion that both the experimental design and analytical procedures must be standardized to insure that only valid studies are published. I acknowledge that the experimental design utilized in our study was not novel, but rather exemplified the use of standardized procedures on a transcriptome previously uncharacterized by RNA-seq, *E. coli* (6). The unique organizational strategy and quantitative methods for analyzing global transcriptome data has however been recognized as novel and highly informative (7-11). Our analytical approach was recently published in a special edition of Current Opinions in Microbiology. It was the editor's intent to use this special issue as a platform to set analytical guidelines for subsequent studies. In addition, in a recent review of regulatory RNA the authors reference our analytical approach as being one of the best resources for sRNA discovery (12). As such, the content described in chapter 3 reflects a contribution to current literature, and describes methods I assisted in

developing to quantifying large transcriptome data sets that were previously considered strictly qualitative in nature.

Chapter 4 Summary

E. coli, by its very nature, is highly capable of coping with dynamic changes in environmental conditions. The adaptability of *E. coli* is the direct result of gene expression modulation, made possible by interchanging sigma factors (13). In *E. coli,* gene expression is quickly altered between exponential and stationary phases of growth by RpoD and RpoS sigma factors, respectively. Under rapid growth conditions RpoD holoenzyme transcribes the majority of genes. When environmental conditions change and begin to induce stress, RpoS becomes the prevailing sigma factor and directs "the general stress response". The RpoS sigma factor plays an important role in the overall success of *E. coli,* and has been implicated in the regulation of genes responsible for biofilm formation and pathogenicity. As such, elucidating the RpoS regulon is of critical importance.

In chapter 4, I utilized RNA-seq and dRNA-seq methodology to investigate RpoS dependency at the operon, gene and promoter levels under carbon, phosphate and nitrogen starvation. RpoS-dependency was identified using DEseq software and a conservative analytical approach. Following differential expression analysis, only transcription units, genes and promoters that were statistically significant (p-value $\leq$ 0.05) and demonstrated a 4-fold or greater change in expression were classified. As a result of my analysis 315 operons, 317 genes, and 278 promoters were classified as RpoS-dependent, far fewer than we were predicting. This is most likely due to the

conservative analytical approach used to identify both genes and promoters. RpoS-dependency was most impactful under carbon starvation conditions accounting for twice as many differentially regulated transcription units than nitrogen or phosphate starvation. Other notable results include the identification of four new transcripts annotated within intergenic regions, a significant difference in the average length of RpoS-dependent (1302 bp) verses independent transcripts (2220 bp), and the observation that RpoS-dependent operons are most often monocistronic. It is my opinion that the results discussed in chapter 4 elucidate the "core" of the RpoS regulon under three different starvation conditions, thereby expanding the number of genes within the "core" of the RpoS regulon.

Future Directions

Transcriptome analysis by RNA sequencing is an ideal method for analyzing global gene expression in bacteria, but what is more important are the multitude of hypotheses that are generated from observing bacterial transcription at this high-level of detail. Moving forward I see the field of bacterial transcriptomics using the insights gained from the global analysis of transcription to focus on investigating poorly understood or previously unknown genetic phenomena, such as pervasive transcription and regulatory RNA discovery. My time studying the *E. coli* transcriptome has left me with more questions than answers. As a conclusion to this dissertation, I will briefly discuss my research plans moving forward.

Biological Significance of Pervasive Transcription in Bacteria

The application of dRNA sequencing methods for the identification of TSSs has been a powerful tool for annotating bacterial transcriptomes. However, an unexpected outcome of dRNA-seq was how robust this method was. An example of the success of this method can be observed in recent literature. Prior to 2009, experimental evidence existed for approximately 800 *E. coli* TSSs (14). By 2009, through the use of a modified 5′ RACE protocol and high-throughput pyrosequencing, more than 1700 TSSs were identified (15). Five years later, the first dRNA-seq analysis of the *E. coli* transcriptome provided support for approximately 2,100 TSSs (2). Finally, in a 2015 study by Thomason *et.al.,* dRNA-seq using an Illumina sequencing platform predicted more than 14,800 candidate TSSs (16). What has become clear from amassing literature is that transcription occurs throughout the *E. coli* genome, and the majority of these transcription events yield RNAs, if they do yield RNAs, with unknown functions.

During the course of analyzing the data presented in chapter 4, I observed the enormity of pervasive transcription using dRNA-seq methods on an Illumina instrument. The volume of transcription observed could be explained two ways: 1) as evidence of a valid biological phenomenon, or 2) as an artifact of Illumina-based sequencing. There is an accumulating body of evidence supporting the hypothesis that pervasive transcription is a valid biological phenomenon (17-19). Within my own work a number of key observations were made that informed my opinion about pervasive transcription. These observations were as follows: pervasive TSSs were consistently observed in replicate samples, viable sigma factor binding sites were identified upstream of pervasive TSSs, pervasive transcript abundance was effected by growth

conditions and mutations, and highly expressed genes appeared to have more upstream pervasive TSSs than weakly expressed genes. As a result of these observations I hypothesized that pervasive transcription does occur and likely provides an advantage to *E. coli* (20). In an effort to better understand the effect pervasive transcription may have on transcript abundance, I propose to insert the green fluorescent protein (GFP) gene directly into the genome of *E. coli* under the control of a wild type RpoD (sigma 70) promoter. Subsequently, I will insert additional RpoD promoter sites, up to five, upstream of the initial promoter and quantify the fluorescence. It is my hypothesis that as the number of pervasive promoters increases so will the abundance of GFP within the cell. Alternatively, a region of the *E. coli* genome containing a number of pervasive promoters can be PCR amplified and ligated directly to the GFP gene sequence. Subsequently, I would mutate the upstream pervasive promoters by replacing the wild type sequence with six consecutive guanines. Whatever the approach the resulting outcome should provide a better understanding of pervasive promoters within the *E. coli* genome.

Bacterial Transcription Regulation by Long-noncoding RNA

The regulatory role of RNA within prokaryotic and eukaryotic cells is more complex than previously depicted. In a recent review, a novel class of regulatory RNA, termed long-noncoding RNA (lncRNA), was added to the ever-growing list of RNA classifications (21). Long-noncoding RNAs are broadly defined as RNA molecules greater than 200 bases in length that do not code for a protein. However, lncRNAs are more accurately described as assisting in the formation of the shape and folding of the

genomic DNA (22, 23). Previously considered to be exclusive to eukaryotic organisms, noncoding transcripts of substantial size (>500 bases) have been identified within the *E. coli* transcriptome (2). It is my hypothesis that lncRNA have a role in regulating the folding of the *E. coli* genome. There are two experimental approaches to be considered: 1) identify a single long-noncoding RNA and investigate its function, or 2) evaluate the function of long-noncoding RNAs globally.

In an effort to better understand the function of a single long-noncoding RNA transcript, I propose to study the transcript *isf*. The *isf* (into sulA function) transcript is an excellent example of lncRNA in the *E. coli* transcriptome. It is a 630 nucleotide long *cis*-encoded RNA that completely overlaps the *sulA* gene, and when expressed is produced at the same level as *sulA* mRNA (24). Considering the features of *isf* and the relationship with the well-studied *sulA* gene, it is logical to assume that *isf* provides a means by which to analyze lncRNA gene regulation in a model system. It has been predicted by others that *isf* down regulates the production of SulA by annealing to the *sulA* mRNA. If the interactions between *isf* and *sulA* RNA do occur, then it can be extrapolated that the production of SulA would decline in response to the formation of double-stranded RNA, and digestion by RNase III. The exact mechanism by which *isf* interacts with *sulA* is uncharacterized and the physiological outcome is unknown. However, the regulation of SulA production has a significant impact on the *E. coli* cell. SulA inhibits cell division by interacting with the FtsZ contractile ring protein. When SulA is bound to FtsZ, formation of the Z-ring is inhibited, which results in the inhibition of cell division. I hypothesize that the *isf* transcript is a member of a novel RNA family, long-noncoding RNA, which controls gene regulation thereby regulating

the production of SulA. I will seek to characterize the expression of *isf* by first identifying *isf* inducible conditions and locating the *isf* promoter region. Subsequent studies will investigate the occurrence and location of putative interactions between *isf* and *sulA* RNAs. Finally, the mechanism by which *isf* regulates the expression of *sulA* will be characterized by artificially over- and under-expressing *isf*.

Alternatively, lncRNAs could be investigated on a global scale. As mentioned above, lncRNA have been observed regulating the shape and folding of genomic DNA. It is logical to hypothesize that some of the 40 lncRNAs annotated in the *E. coli* transcriptome would be essential for maintaining the polarity and folding structure of the *E. coli* genome. Therefore, I propose to systematically mutate the -10 region of the promoter for each of the 40 lncRNAs by replacing the wild type sequence with six consecutive guanines. Following mutation of a single lncRNA promoter, the resulting mutant will be grown in MOPS minimal medium with 0.2% glucose, and the growth rate for each mutant will be evaluated. In addition, the spatial organization of the nucleoid for wild type and mutant *E. coli* will be evaluated for structural variations. Experimental methods and analysis of nucleoid organizational structure are well established, and tools such as MicrobeTracker provide an ideal resource for such an analysis (25).

The era of big data has extended into microbiological research, and the effects are prevalent. As of April 2015, 5339 bacterial genomes have been completely sequenced, transcriptomes and proteomes are published at a remarkable rate, and there is seemingly a metagenome for nearly every environment on Earth. In addition, there is an observable trend within the literature to conduct experiments at a global-scale.

Historically, model organisms, such as *E. coli* and *B. subtillis*, were studied in order to gain understanding about how a particular system functioned, and the insight gained would be extrapolated to other closely related species. With the emergence of RNA-seq and dRNA-seq, the field of bacterial genetics is no longer limited to investigating a single gene or operon in a single model organism. As a result, the number of novel biological insights has abounded. Over the next five years I anticipate that RNA-seq analysis will transform our understanding of microbial genetics, and I look forward to the opportunity that I have been provided to contribute to this field of study.

References

1. **Sharma CM, Vogel J.** 2014. Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin Microbiol **19:**97-105.

2. **Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL.** 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio **5:**e01442-01414.

3. **Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC.** 2013. An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. Cell Host Microbe **14:**683-695.

4. **Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC.** 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A **109:**E1277-1286.

5. **Creecy JP, Conway T.** 2015. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol **23:**133-140.

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

7. **Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW.** 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics **15:**141-161.

8. **Otsuka Y, Muto A, Takeuchi R, Okada C, Ishikawa M, Nakamura K, Yamamoto N, Dose H, Nakahigashi K, Tanishima S, Suharnan S, Nomura W, Nakayashiki T, Aref WG, Bochner BR, Conway T, Gribskov M, Kihara D, Rudd KE, Tohsato Y, Wanner BL, Mori H.** 2015. GenoBase: comprehensive resource database of Escherichia coli K-12. Nucleic Acids Res **43:**D606-617.

9. **Hermes FA, Cronan JE.** 2014. An NAD synthetic reaction bypasses the lipoate requirement for aerobic growth of Escherichia coli strains blocked in succinate catabolism. Mol Microbiol doi:10.1111/mmi.12822.

10. **Alvarez R, Neumann G, Fravega J, Diaz F, Tejias C, Collao B, Fuentes JA, Paredes-Sabja D, Calderon IL, Gil F.** 2015. CysB-dependent upregulation of the Salmonella Typhimurium cysJIH operon in response to antimicrobial compounds that induce oxidative stress. Biochem Biophys Res Commun **458:**46-51.

11. **Romero DA, Hasan AH, Lin YF, Kime L, Ruiz-Larrabeiti O, Urem M, Bucca G, Mamanova L, Laing EE, van Wezel GP, Smith CP, Kaberdin VR, McDowall KJ.** 2014. A comparison of key aspects of gene regulation in Streptomyces coelicolor and Escherichia coli using nucleotide-resolution transcription maps produced in parallel by global and differential RNA sequencing. Mol Microbiol doi:10.1111/mmi.12810.

12. **Miyakoshi M, Chao Y, Vogel J.** 2015. Regulatory small RNAs from the 3' regions of bacterial mRNAs. Curr Opin Microbiol **24:**132-139.

13. **Dong T, Kirchhof MG, Schellhorn HE.** 2008. RpoS regulation of gene expression during exponential growth of Escherichia coli K12. Mol Genet Genomics **279:**267-277.

14. **Huerta AM, Salgado H, Thieffry D, Collado-Vides J.** 1998. RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res **26:**55-59.

15. **Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E.** 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. PLoS One **4:**e7526.

16. **Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G.** 2015. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. J Bacteriol **197:**18-28.

17. **Wade JT, Grainger DC.** 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol **12:**647-653.

18. **Jensen TH, Jacquier A, Libri D.** 2013. Dealing with pervasive transcription. Mol Cell **52:**473-484.

19. **Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R.** 2012. Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev **26:**2621-2633.

20. **Canny SP, Reese TA, Johnson LS, Zhang X, Kambal A, Duan E, Liu CY, Virgin HW.** 2014. Pervasive transcription of a herpesvirus genome generates functionally important RNAs. MBio **5:**e01033-01013.

21. **Ponting CP, Oliver PL, Reik W.** 2009. Evolution and functions of long noncoding RNAs. Cell **136:**629-641.

22. **Nagano T, Fraser P.** 2011. No-nonsense functions for long noncoding RNAs. Cell **145:**178-181.

23. **Bernstein E, Allis CD.** 2005. RNA meets chromatin. Genes Dev **19:**1635-1655.

24. **Cole ST, Honore N.** 1989. Transcription of the sulA-ompA region of Escherichia coli during the SOS response and the role of an antisense RNA molecule. Mol Microbiol **3:**715-722.

25. **Sliusarenko O, Heinritz J, Emonet T, Jacobs-Wagner C.** 2011. High-throughput, subpixel precision analysis of bacterial morphogenesis and intracellular spatio-temporal dynamics. Mol Microbiol **80:**612-627.

Appendix A: Published Articles

RESEARCH ARTICLE

Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing

Tyrrell Conway,[a] James P. Creecy,[a] Scott M. Maddox,[a] Joe E. Grissom,[a] Trevor L. Conkle,[a] Tyler M. Shadid,[a] Jun Teramoto,[b] Phillip San Miguel,[c] Tomohiro Shimada,[d,e] Akira Ishihama,[d] Hirotada Mori,[f] Barry L. Wanner[b]

Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma, USA[a]; Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA[b]; Purdue Genomics Facility, Purdue University, West Lafayette, Indiana, USA[c]; Department of Frontier Bioscience and Micro-Nanotechnology Research Center, Hosei University, Koganei, Tokyo, Japan[d]; Chemical Resource Laboratory, Tokyo Institute of Technology, Nagatsuda, Yokohama, Japan[e]; Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara, Japan[f]

ABSTRACT **We analyzed the transcriptome of *Escherichia coli* K-12 by strand-specific RNA sequencing at single-nucleotide resolution during steady-state (logarithmic-phase) growth and upon entry into stationary phase in glucose minimal medium. To generate high-resolution transcriptome maps, we developed an organizational schema which showed that in practice only three features are required to define operon architecture: the promoter, terminator, and deep RNA sequence read coverage. We precisely annotated 2,122 promoters and 1,774 terminators, defining 1,510 operons with an average of 1.98 genes per operon. Our analyses revealed an unprecedented view of *E. coli* operon architecture. A large proportion (36%) of operons are complex with internal promoters or terminators that generate multiple transcription units. For 43% of operons, we observed differential expression of polycistronic genes, despite being in the same operons, indicating that *E. coli* operon architecture allows fine-tuning of gene expression. We found that 276 of 370 convergent operons terminate inefficiently, generating complementary 3′ transcript ends which overlap on average by 286 nucleotides, and 136 of 388 divergent operons have promoters arranged such that their 5′ ends overlap on average by 168 nucleotides. We found 89 antisense transcripts of 397-nucleotide average length, 7 unannotated transcripts within intergenic regions, and 18 sense transcripts that completely overlap operons on the opposite strand. Of 519 overlapping transcripts, 75% correspond to sequences that are highly conserved in *E. coli* (>50 genomes). Our data extend recent studies showing unexpected transcriptome complexity in several bacteria and suggest that antisense RNA regulation is widespread.**

IMPORTANCE **We precisely mapped the 5′ and 3′ ends of RNA transcripts across the *E. coli* K-12 genome by using a single-nucleotide analytical approach. Our resulting high-resolution transcriptome maps show that ca. one-third of *E. coli* operons are complex, with internal promoters and terminators generating multiple transcription units and allowing differential gene expression within these operons. We discovered extensive antisense transcription that results from more than 500 operons, which fully overlap or extensively overlap adjacent divergent or convergent operons. The genomic regions corresponding to these antisense transcripts are highly conserved in *E. coli* (including *Shigella* species), although it remains to be proven whether or not they are functional. Our observations of features unearthed by single-nucleotide transcriptome mapping suggest that deeper layers of transcriptional regulation in bacteria are likely to be revealed in the future.**

Received 5 June 2014 **Accepted** 16 June 2014 **Published** 8 July 2014

Citation Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL. 2014. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio 5(4):e01442-14. doi:10.1128/mBio.01442-14.

Editor Sankar Adhya, National Cancer Institute, NIH

Address correspondence to Tyrrell Conway, tconway@ou.edu, or Barry L. Wanner, blwanner@purdue.edu.

This article is a direct contribution from a Fellow of the American Academy of Microbiology.

Escherichia coli burst into the realm of model organisms with the discovery of conjugation by Joshua Lederberg in 1946 (1). Just 15 years later, Francois Jacob and Jacque Monod proposed the operon model in *E. coli* (2). For two-thirds of a century, *E. coli* has been an important vehicle for scientific investigation, playing a role in research resulting in no fewer than 10 Nobel prizes (1–10). The *E. coli* K-12 genome was among the early ones sequenced (11) and *E. coli* is unique among model organisms, possessing biochemical or genetic evidence for functions for ca. 75% of its known genes, making it arguably the best understood organism (12). Examination of its genome sequence confirmed what had long been surmised, that genes of related function are frequently arranged in operons (13–15).

Soon after the discovery of the *lac* operon, it became clear that not all operons are transcribed as discrete units of information neatly arranged end to end on the genome. First, it was recognized that regions of phage lambda are transcribed on complementary strands (16). Over the next 40 years, operons were studied, one or two at a time, in line with the technology of the day, revealing occasional glimpses of transcriptional complexity arising from overlapping, divergent (17, 18) and convergent operons (19, 20). Second, the perception of transcriptome complexity was forever

changed when it was found that at least one antisense (AS) transcription start site is associated with nearly one-half (46%) of *Helicobacter pylori* genes (21). There is also a substantial amount of AS transcription in *E. coli* (22–24). While some researchers suggested that extensive AS transcription is a "by-product" of the transcription machinery, largely because AS transcripts did not appear to be conserved in enteric bacteria (25), others concluded the opposite, that AS RNA has an important role in transcriptional regulation (26–32). The recent identification and sequencing of 316 potentially functional double-stranded RNAs in *E. coli* is a step toward laying the argument to rest (33). The "excludon" concept of AS RNA control of divergent operons ascribes an important function to overlapping, complementary transcripts (34). A recent study of *Staphylococcus aureus* suggests that AS transcripts drive RNase III-mediated RNA processing, although a comparison of the AS RNA content of selected bacteria led the authors to infer that the mechanism is prevalent in Gram positives but absent in Gram negatives (30). Amid the mounting evidence for transcriptional complexity in bacteria and the finding that AS transcripts are prevalent in bacteria, we undertook a comprehensive transcriptome analysis of *E. coli*.

RNA sequencing (RNA-Seq) offers tremendous power for high-resolution transcriptome analysis. However, the fullness of its power has yet to be realized for *E. coli*, because all previous studies of the *E. coli* transcriptome failed to annotate both the 5′ and 3′ transcript ends and hence operons were not precisely mapped. We therefore developed an organizational schema described herein to precisely map RNA-Seq reads across entire operons, including both the 5′ and 3′ transcript ends, and to annotate these data in the context of the operon arrangement on the transcriptome. Though others used tiling microarray technology to address bacterial transcriptome organization (28, 35), tiling arrays did not have the resolving power to define transcript ends precisely or to elucidate operons with multiple promoters. More recent transcriptome mapping studies of *E. coli* relied on 5′ end mapping to identify transcription start sites (TSSs) (36, 37). However, our own critical examination of these data sets revealed extensive discrepancies that call into question many candidate TSSs, reinforcing the need for alternative promoter-mapping strategies (38). Recent RNA-Seq analyses of *E. coli* were also unfortunately not designed to map transcript ends accurately because they relied on randomly primed cDNA synthesis (39) or they had a resolution of only ca. 50 nucleotides due to low sequence read coverage (40). The recent development of differential RNA-Seq technology allowed mapping TSSs in *Helicobacter pylori* (21) and *Salmonella enterica* (29, 41); however, operon architecture was not determined because the 3′ ends were not mapped. In evaluating these approaches, we recognized that identification of both 5′ and 3′ transcript ends is essential for precise mapping of transcriptional regulatory features.

Considering the foundational role of *E. coli* in the life sciences, high-resolution RNA-Seq will stimulate progress by unambiguous mapping of the features that control transcription. To annotate operons and characterize their response to carbon starvation, we obtained a time series of RNA samples from wild-type *E. coli* K-12 BW38028 cultures grown to stationary phase on chemically defined, carbon source-limited (glucose) minimal medium. We chose these conditions because they are intrinsic to the physiology that allows *E. coli* to colonize the mammalian intestine yet survive in the environment until encountering a new host and, in the case of *E. coli* pathogens, cause disease (42). We analyzed these RNA samples by deep sequencing with a strand-specific RNA ligation approach (43) that ensures full read coverage and precise mapping of both the 5′ and 3′ transcript ends. In practice, only three transcriptional features are needed to define operon architecture, regardless of complexity. These are the 5′ ends (promoters), the 3′ ends (terminators), and sufficient RNA-Seq read coverage to connect the ends, which together define operons (Fig. 1). Our analyses revealed an unprecedented high-resolution view of *E. coli* operon architecture. Our analytical approach allowed us to test the hypothesis that bacterial operon structure accommodates substantial transcriptional complexity. We offer our annotated *E. coli* K-12 operon map as a community resource upon which others can participate in annotating additional transcriptional regulatory features (GEO accession no. GSE52059).

RESULTS AND DISCUSSION

Single-nucleotide resolved RNA-Seq data sets. *E. coli* K-12 has served as an important model organism for more than a half century and was the first bacterium analyzed by DNA microarray technology (44, 45). While several other bacteria have now been analyzed by RNA-Seq (21, 26, 28, 31, 41, 47–49), the limited RNA-Seq studies of *E. coli* have not provided single-nucleotide resolution (39, 40). Herein, we used a strand-specific RNA ligation-based RNA-Seq strategy, which when coupled with a robust analytical approach, allowed us to define transcriptional features across the whole *E. coli* genome at single-nucleotide resolution. We acquired time series of RNA samples from duplicate cultures of *E. coli* K-12 BW38028 and its isogenic *rpoS* mutant BW39452 during logarithmic- and stationary-phase growth on glucose-limited minimal medium (see Fig. S1 in the supplemental material). In total, we sequenced 26 RNA samples to generate a data set of 72.1 million uniquely mapped sequence reads corresponding to 5.5 gigabases of RNA-Seq data (see Table S1). Appropriate temporal expression of *bolA*, a known glucose starvation-inducible gene (50), confirmed that our RNA-Seq data correctly represented the growth conditions (Fig. 1). Our ongoing analyses of the RpoS regulon will be reported elsewhere. The correlation between replicate cultures was >0.96 (see Fig. S1); this level of biological replication provides a reliable view of the *E. coli* K-12 transcriptome (Fig. 2). The data are available at GEO (accession no. GSE52059).

We developed an in-house computational tool to convert the binary read alignment (BAM) files to base count (WIG) files to facilitate single-nucleotide resolution analyses. We normalized our base count data by using a strategy analogous to the total count approach (51) for normalizing gene-specific read alignments. Accordingly, the resulting WIG files contain only the base location and the number of times each base is read (sequenced) and are >100-fold smaller than the sample read alignment (SAM) files. Advantages of this simple base count approach are several-fold: first, the data are inherently more computable; second, normalization of base count data makes all samples directly comparable and eliminates transcription unit (TU) length bias; third, the base counts of individual features can be computed and queried at any desired resolution from single nucleotide to an entire operon.

Since we analyzed RNA-Seq reads at the base count level, the normalized base counts can be readily averaged across any range of bases to calculate the relative usage of transcriptional features, including promoters, terminators, TUs, and operons (Fig. 1). We empirically determined the number of bases used to calculate pro-

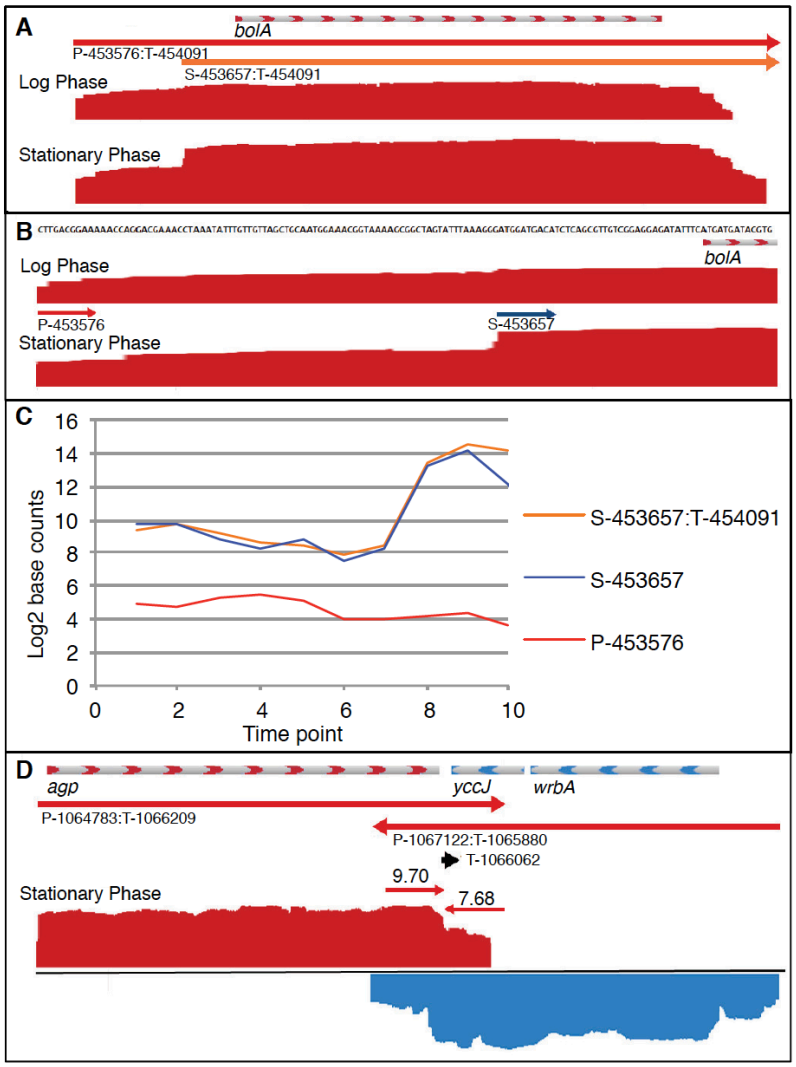


FIG 1 Single-nucleotide resolution of promoters and terminators in example complex operons. (A) The *bolA* operon contains transcription units (TUs) P-453657:T-454091 (red arrow) and S-453688:T-454091 (orange arrow). RNA-Seq data are shown in a JBrowse visualization of positive-strand (red) transcription in logarithmic- and stationary-phase samples (average from three replicates). The base count data were normalized and $\log_2$ transformed such that track heights in JBrowse are directly comparable. (B) *bolA* promoter region showing primary promoter P-453576 and secondary promoter S-453658 at single-nucleotide resolution (drawn to scale). (C) Plot of promoter usage (average count of 10 bases beginning at TSS) and TU usage (average count of bases within TU) for 10 growth curve time points showing *bolA* induction upon entry into stationary phase (see Fig. S1 for growth curve). (D) Terminator usage (average counts of 10 bases preceding and following terminator) is shown for T-1066062, which is shared by converging operons *agp* on positive strand (red) and *wrbA-yccJ* on negative strand (blue).

moter usage by comparing the single base count value at the TSS to 3-, 5-, 10-, and 20-base averages, each beginning at the TSS. In practice, the shorter base count lengths were highly variable, presumably because of staggered starts that are occasionally observed in primer extension experiments (52) and were frequently observed in the RNA-Seq data sets. However, a 20-base-count length was too long to allow discrimination of closely spaced promoters. We therefore used 10-base average counts for quantifying promoter usage (Fig. 1). The same 10-base average worked well for calculating terminator efficiency by comparing the 10-base average counts before and after the termination site (Fig. 1 and 3). We used these base count values to calculate the usage of individual

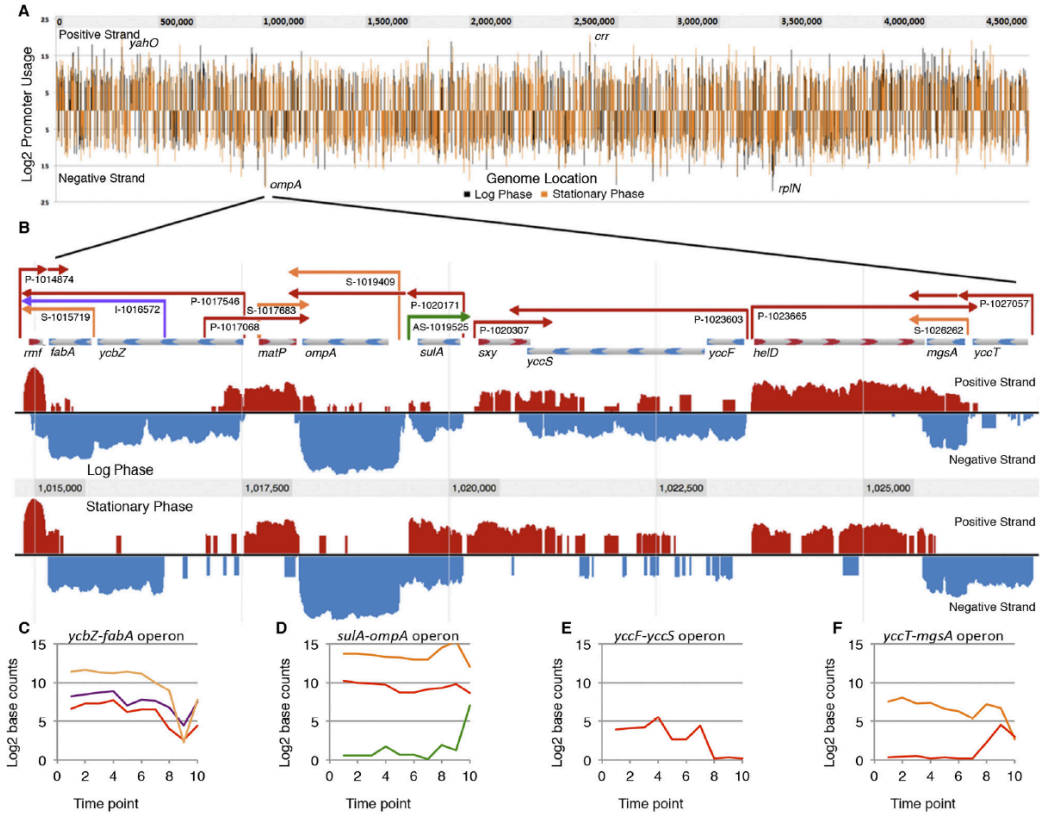


FIG 2 Genome-wide promoter locations and annotated transcriptome map of a selected region. (A) Promoters aligned by genome location. Line heights correspond to normalized, TEX-enriched promoter usage values (see text for details), shown for logarithmic phase (black) and stationary phase (orange). (B) Annotated regulatory features of a selected region of the genome. Positive-strand RNA-Seq data (red) and negative-strand data (blue) were normalized for comparison between logarithmic- and stationary-phase samples. Primary promoters and corresponding TUs (red) are indicated by arrows extending from promoter to terminator, as are secondary promoters (orange), internal promoters (purple), and AS promoters (green). Beginning on the left, *rmf* is transcribed from a primary promoter and depending on growth conditions terminates either before or within the *ycbZ-fabA* operon, which has a primary promoter upstream of *ycbZ*, an internal promoter within *ycbZ*, and a secondary promoter upstream of *fabA*. *matP* is transcribed from primary and secondary promoters. *ompA* is transcribed from a secondary promoter in logarithmic phase and is cotranscribed from the primary promoter of the *sulA-ompA* operon during stationary phase. An AS TU that overlaps the *sulA* sense transcript is turned on in stationary phase. The *sxy* and *yccF-yccS* operons converge. Finally, *mgsA* is transcribed as an independent TU from a secondary promoter in logarithmic phase and also is expressed in the *yccT-mgsA* operon from a promoter that is active only in stationary phase. (C) Plot of TU base counts for *ycbZ-fabA* operon, colorized according to color scheme in panel B; (D) TU plot of *sulA-ompA* operon; (E) TU plot of *yccFS* operon; (F) TU plot of *yccT-mgsA* operon.

transcription features as well as the impact of operon structure on relative TU and gene expression.

Promoter mapping. Our search for promoters was driven by mapping of putative TSSs on the basis of (i) enrichment with terminator exonuclease (TEX), which degrades RNA molecules with 5′-monophosphate ends and consequently enriches for 5′-triphosphate ends corresponding to the nucleotide initiated *de novo* by RNA polymerase (18); (ii) promoter motif analysis; (iii) consensus among replicate data sets; and (iv) sigma factor-specific RNA polymerase binding (SELEX). None of these approaches alone is comprehensive, because each gives rise to false-positive results and fails to find all TSSs (20). For example, TEX treatment does not enrich for some TSSs because RppH phosphatase activity removes the 5′-triphosphates (53). Additionally, not all promoters have consensus motifs that can be identified by computer algorithms (54), nor do all promoters bind RNA polymerase *in vitro* (55).

To facilitate promoter mapping, we wrote a simple algorithm to search for changes in base count values exceeding 2-fold in replicate TEX-enriched and coverage data sets ($n = 14$, wild-type [WT] and *rpoS* culture samples from log and stationary phase). The TSSs of highly expressed genes were apparent in all 14 replicates. However, since the 14 samples represented both logarithmic- and stationary-phase samples, expression of some promoters was condition specific. In order to generate transcriptome maps that are condition independent for annotating the

response to many conditions in the future, we chose a consensus in which three replicates of either logarithmic- or stationary-phase samples have TSSs at the identical base locations as a starting point for promoter mapping. This conservative strategy revealed 11,329 putative TSSs, a value that is similar to the number of promoters found by Thomason and Storz (submitted for publication), and includes known promoters of weakly expressed genes. However, this number far exceeds the expected promoter density on the *E. coli* genome, thus exemplifying the need to use a multifaceted approach to confirm promoters. To identify candidate promoters missed by TSS mapping of regions that had few RNA-Seq reads, we employed genomic SELEX screening, which was developed for quick identification of genes under the control of specific transcription factors (57). Confirmation of tentative TSS's by RNAP binding was previously employed for promoter mapping of *Salmonella enterica* serovar Typhimurium (29). Sites that bound RpoD *in vitro*, exceeding a conservative signal-to-background ratio threshold of 3.0, and corresponded to RNA-Seq reads expressed *in vivo* identified 1,254 additional candidate promoters (see Fig. S2 in the supplemental material).

Next, we used a bioinformatics approach to search the 50-bp sequences immediately upstream of the 12,583 putative TSSs for promoter motifs by using FIMO software (58) to screen against a library of *E. coli* promoter motifs available at DPInteract (59). We found it was necessary to modify the RpoD promoter library according to the characterization of 554 promoters by Mitchell et al. (60), which demonstrated that the RpoD consensus promoter has −10 and −35 regions with spacing of 14 to 20 bases between promoter elements. The search output was restricted to promoter sequences correctly positioned within ±3 bases of the TSS, with *E* values corresponding to *P* values of <0.02. This multifaceted approach yielded 5,653 putative RpoD-dependent promoters, which we evaluated further by manual annotation, which involved direct visual observation.

A visual graphic environment (JBrowse [61]) interfaced to an Oracle database facilitated manual annotation documentation. From the list of candidate promoters, we created a JBrowse track at the corresponding base locations, each displaying a "clickable" URL call to the database that automatically recorded the base location and allowed manual entry of metadata, including the type of promoter, regulatory information supported by differential expression analysis, and comments. We annotated only promoters that could be experimentally associated with operons, by using RNA-Seq data as described in the next section. This comprehensive strategy yielded 2,122 vegetative promoters (Fig. 2), which more than doubled the 811 individually characterized *E. coli* promoters annotated in RegulonDB and calls into question several thousand candidate promoters that were identified by less reliable high-throughput strategies (35, 38). The promoter data set (see Table S2) is dominated by primary promoters (P), defined as the furthest upstream promoter in an operon (66.3%), with a lower number of secondary promoters (S) that are intergenic and downstream of P promoters (19.6%), internal promoters (I) that are intragenic (9.8%), and AS (4.2%) promoters. All possible arrangements and orientations of these promoter types were observed and collectively generated substantial complexity in the transcriptome (Fig. 2).

It is well known that promoter strength, i.e., quality, varies greatly (60) and that variability is reflected in our data set. To quantify promoter quality, we scored the four criteria (metrics) used to map candidate promoters (see Table S2). The promoter quality score was calculated by applying a weighted matrix on the basis of 10 points, where TEX enrichment carries a weight of 4, the promoter motif score carries a weight of 3, the TSS consensus (between replicates) score carries a weight of 2, and the SELEX score carries a weight of 1. The resulting analyses yielded promoters scored on a scale of 0 to 10. The TEX enrichment metric reflects the number of instances among four TEX replicates in which the ratio of TEX-treated versus non-TEX-treated base counts (10-base-count average beginning at the TSS) for a sample exceeded 2-fold. The promoter motif score was calculated in quartiles of *E* values for RpoD-dependent promoter motifs as determined by using FIMO. The TSS consensus score was calculated as the number of occurrences of a TSS at a precise base location divided by the total number of samples evaluated ($n = 14$). The final metric was the presence or absence of SELEX-determined RpoD binding, which was scored as a 1 or 0. The 2,122 promoters ranged in score from 10 to 0.14, with the top 10% of promoters scoring above 7.8, the bottom 10% scoring below 2.9, and the average promoter scoring 5.5.

We found no strong correlation between promoter usage (average count of first 10 transcribed bases) and promoter confidence scores or promoter motif scores (see Fig. S3 in the supplemental material), which is in agreement with an earlier report (60). However, we did find a weak correlation between promoter usage and TU usage (average count of bases from promoter to terminator) (see Fig. S3). We confirmed that TU usage and RNA half-life (62) (measured under similar conditions) do not correspond, as noted previously. Nevertheless, promoter and TU usage values do reflect the physiologically relevant transcript level, as the RNA concentration in the cell is determined both by the frequency of transcription initiation and the rate of RNA decay, which vary substantially for different transcripts (62).

Operon mapping. To annotate operons, it was also necessary to map the 3′ transcript ends, which allowed documenting the connections between promoters and the corresponding downstream terminators (Fig. 1). Our criteria for operon annotation were (i) the P promoter must be followed by sequence read coverage across the entire operon, (ii) the mapped 3′ ends must extend beyond the stop codon of the last gene in the operon, (iii) the S and I promoters must increase read coverage of the downstream bases, and (iv) internal terminators must decrease coverage of downstream bases without interrupting contiguous coverage by readthrough transcripts. Our search for 3′ ends that can be associated with annotated promoter(s) by deep sequence read coverage throughout the operon led to mapping 1,774 candidate terminators (see Table S3 in the supplemental material), 264 of which lie within operons and apparently permit partial readthrough transcription of downstream genes, as demonstrated for the *sdhCDAB-sucABCD* operon (Fig. 3). We evaluated the 1,774 3′ ends by using TransTermHP (63) and confirmed that 623 have sequences characteristic of intrinsic terminators, which extends the number of annotated *E. coli* terminators previously annotated (227 [38]) by nearly 8-fold. It has been predicted that roughly one-half of terminators are intrinsic (64). The remaining 1,151 terminators that were not confirmed by TransTermHP are candidates for ones requiring Rho or another protein factor for termination. Since there is no computational approach to identify factor-dependent terminators, the data in Table S3 represent the

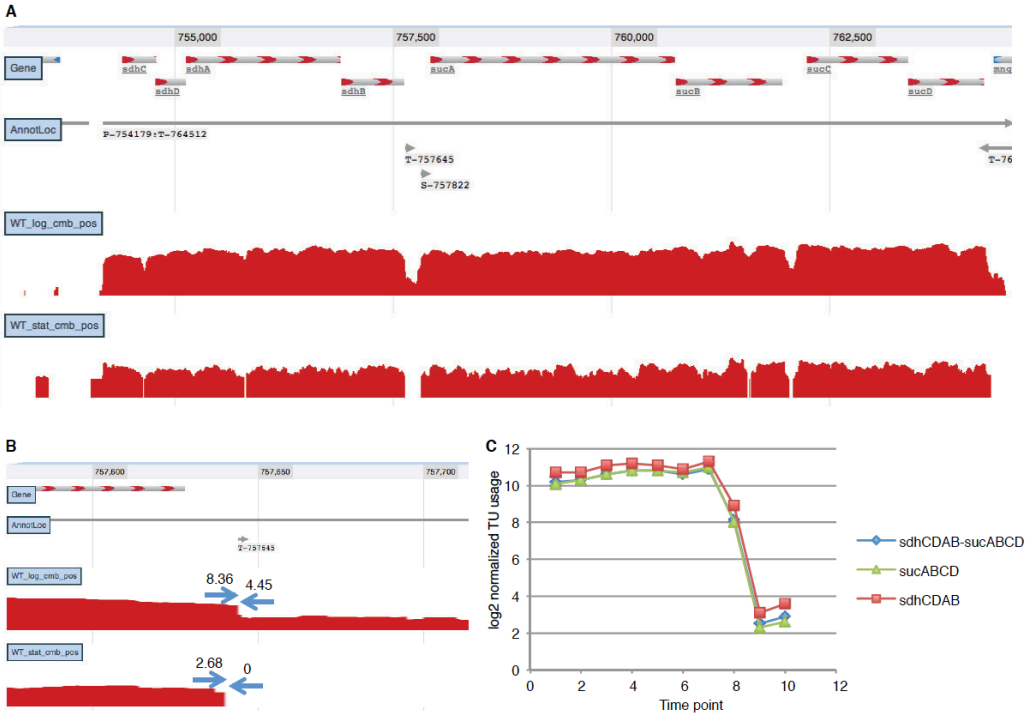


FIG 3 Balanced transcript coverage of the *sdhCDAB-sucABCD* operon achieved by complex interaction of internal terminator and secondary promoter. (A) JBrowse instance showing coverage data; (B) terminator usage in logarithmic (WT_log_cmb_pos) and stationary (WT_stat_cmb_pos) phase; (C) TU coverage time series.

most extensive genome-wide prediction of nonintrinsic terminators.

The preceding analyses of only logarithmic- and stationary-phase samples revealed a total of 6,463 regulatory features, including 2,122 promoters (see Table S2), 1,774 terminators (see Table S3), and 2,566 transcription units (TUs) corresponding to 1,510 operons (see Table S4). The sequence reads covered more than 90% of bases within 90% of the annotated operons. The 1,510 operons cover 2,985 of 4,457 genes (approximately two-thirds) annotated on the reference genome. As more data sets and growth conditions are analyzed, our simple organizational schema should make it straightforward to add newly identified regulatory features to the *E. coli* K-12 transcriptome map. For ready distribution, we converted our data sets to GenBank format by using "promoter," "terminator," and "operon" as feature keys (65). This data format allows annotation of any number of experimental parameters that affect the usage of these features. Our *E. coli* K-12 transcriptome annotation GenBank feature table is available from GEO (accession no. GSE52059).

Operon organization examples. The data in Fig. 2 unequivocally confirm that the *E. coli* genome is organized in operons. In its original conception, the operon has a regulatory region with a single promoter that initiates transcription of a polycistronic mRNA covering the *lac* operon genes and ends with a single terminator. Indeed, many *E. coli* operons fit this model or are even simpler if they contain a single gene (monocistronic). However, the whole *E. coli* transcriptome reveals densely packed regulatory features that cannot be discerned from the genome sequence alone (Fig. 2). Complex operons result from transcripts initiated by S and I promoters, as well as internal terminators. For example, *sulA* and *ompA* are independently transcribed during logarithmic phase, with each gene having its own promoter and terminator. However, during stationary phase, the *sulA* TU reads through a nonintrinsic *sulA* terminator to form the *sulA-ompA* transcript, driven by an S promoter that increases expression of the *ompA* TU (Fig. 2). An AS transcript that fully overlaps the *sulA* coding sequence is also switched on in stationary phase. This arrangement of the *sulA-ompA* operon and AS transcript was postulated as a means for posttranscriptional control of the synthesis of the cell division inhibitor SulA (66), which is further supported by our results. Our organizational schema makes the previously unannotated *sulA* AS transcript (12) and similar regulatory features readily apparent on the *sulA-ompA* transcriptome map (Fig. 2). Such differential expression of TUs within operons can provide bacteria with the ability to modulate gene expression to cope with physiological complexity (28, 30, 34, 41).

It is especially notable that Fig. 2 reveals the *E. coli* transcriptome for only two growth conditions, log phase and stationary

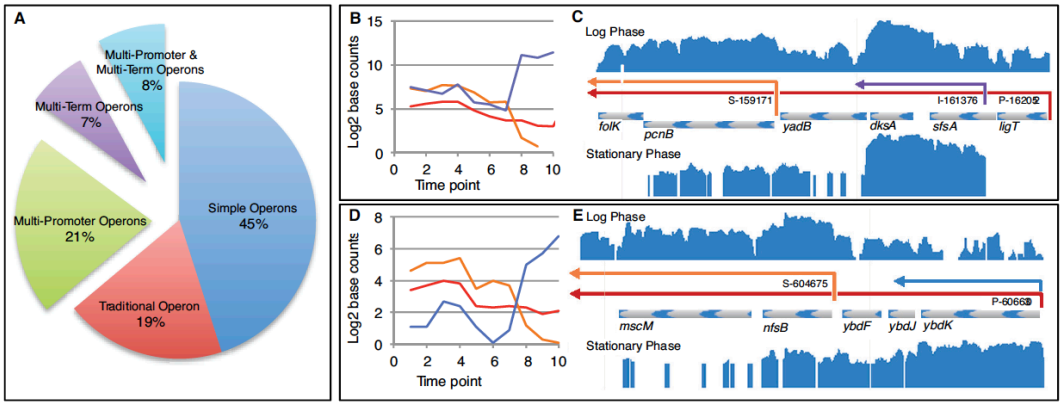


FIG 4 Computational analysis of single-nucleotide resolution data reveals complex operon architecture. (A) Operons organized by increasing complexity; (B) TU usage plot of *ligT-sfsA-dksA-yadB-pcnB-floK* operon. The primary TU corresponding to the entire operon is shown in red. The differentially expressed *dksA*-specific TU driven by promoter I-161376 is shown in purple. The *pcnB-folK* TU driven by S-159171 is shown in orange. Note that transcript levels of *dksA* increase upon entry into stationary phase, whereas *pcnB-folK* decreases. (C) JBrowse instance showing *ligT-sfsA-dksA-yadB-pcnB-floK* operon; (D) TU usage plot of *ybdK-ybdJ-ybdF-nrsB-mbcM* operon. Note the primary TU corresponding to the entire operon (red) decreases only slightly during transition from logarithmic phase into stationary phase, because it is comprised of two differentially expressed TUs, one of which increases and the other decreases during growth: the *nfsB-mbcM*-specific transcript (orange) essentially disappears in stationary phase, whereas the *ybdK*-specific transcript (blue) is induced in stationary phase. (E) JBrowse instance of *ybdK-ybdJ-ybdF-nrsB-mbcM* operon.

phase, due to carbon source limitation. Our analyses show that 29% of operons have more than one promoter and that 15% of operons have more than one terminator under these conditions (Fig. 4). Further, many operons are subject to multiple regulatory inputs (38) that have not been examined here. Differential mRNA decay can also provide an additional layer of control within operons (62). No doubt, future RNA-Seq analysis of the myriad responses to numerous regulatory signals is likely to reveal substantially more variation in operon architecture, as seen for *Salmonella* (41).

Intricacy is readily apparent for operons with internal promoters and terminators. For example, three promoters upstream of the *ahpCF* operon contribute to its expression in an additive fashion (Fig. 5). Such an arrangement allows differential control of alkylhydroperoxidase production in response to stationary phase, osmotic stress, and oxidative stress (67). Likewise, three promoters contribute to *ybfE-fldA-uof-fur* operon expression during logarithmic phase, allowing for continuation of *uof-fur* TU expression, decline of *fldA* expression, and turnoff of *ybfE* expression in the stationary phase (Fig. 5). Although cotranscription of the complex *ybfE-fldA-uof-fur* operon was not previously recognized (68), it makes sense for *uof-fur* to be transcribed independently of *ybfE-fldA* under certain conditions, because *fur* encodes a negative regulator of genes for iron uptake. Furthermore, *uof* expression is controlled indirectly by the *trans*-acting noncoding RNA RhyB, which is itself Fur regulated, thus forming a negative feedback loop responsive to iron limitation (68). The ability to unravel condition-specific terminator usage by our organizational schema is illustrated for the internal terminator of the *sdhCDAB-sucABCD* operon, which encodes three enzymes of the tricarboxylic acid (TCA) cycle (Fig. 3). This arrangement explains how intrinsic termination can allow one operon to function independently as two operons under appropriate conditions (69). These examples demonstrate how our promoter and terminator usage calculations can reveal new biological insights from the RNA-Seq transcriptome analyses.

Catalogue of operon architecture. High-resolution mapping of well-characterized regions of the genome provided glimpses of intricate operon arrangements (Fig. 2 to 5). Our analyses of *E. coli* operons at single-nucleotide resolution further revealed numerous instances of complexity genome-wide. Single-gene operons with a single promoter and terminator make up 47% of all operons, while 17% are "traditional" operons with more than one gene and a single promoter and terminator (Fig. 4). The remaining operons (36%) are more complex: 21% have multiple (as many as 8) promoters, 7% have multiple (as many as 4) terminators, and 8% have both multiple promoters and multiple terminators. The average operon has 1.98 genes (see Table S4 in the supplemental material). In our data set, the most complex operon, which encodes several core cellular functions, has 14 genes, 8 promoters, 4 terminators, and 23 TUs (*yjeF-yjeE-amiB-mutL-miaA-hfq-hflX-hflK-hflC-yjeT-purA-nsrR-rnr-rlmB* operon; see Fig. S4).

Differential TU expression within operons can result from the activity of S and I promoters, internal terminators, and combinations of these regulatory features. For example, Fig. 4 illustrates how an I promoter and internal terminator can function together to increase expression of the DksA-specific TU in stationary phases. For the *ybdK* operon, Fig. 4 shows differential expression of the 5′ and 3′ TUs of the operon caused by transcription from an S promoter and an internal terminator. This arrangement of features results in a complete inversion in expression of the 2 TUs between logarithmic and stationary phases. These findings suggest that operon architecture permits *E. coli* to adjust relative levels of gene expression within the same operon in response to environmental conditions.

To quantify differential gene expression within *E. coli* operons, we compared the base counts of TUs within the same operon under the same growth condition and tabulated the complexity

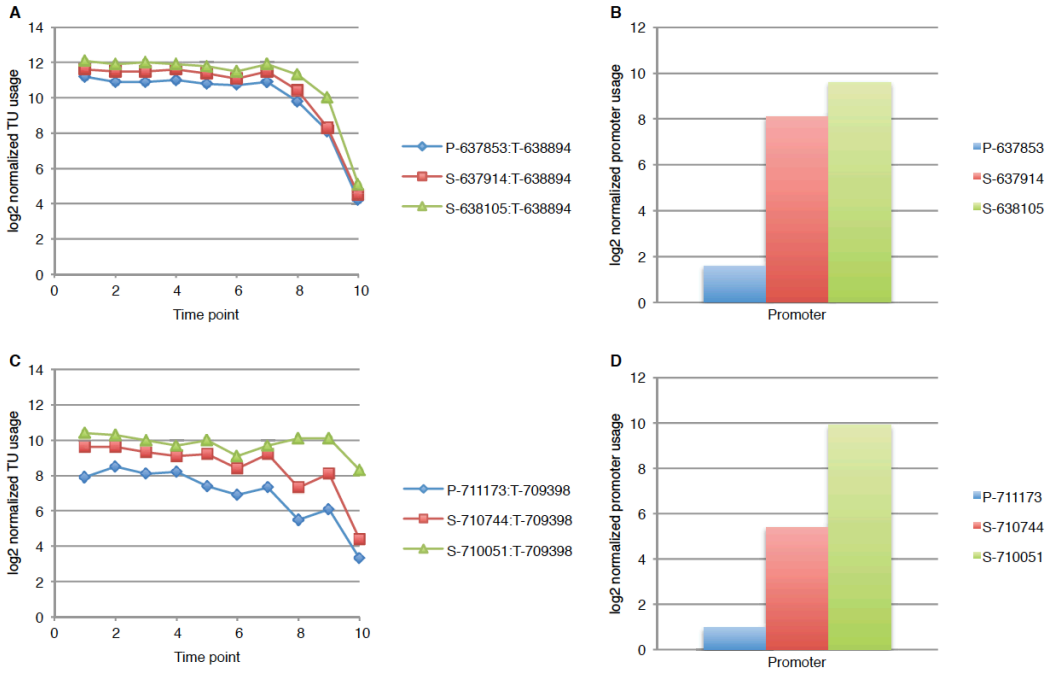


FIG 5 Three promoters contribute to expression levels of genes within the *ahpCF* and the *ybfE-fldA-uof-fur* operons. (A) WT time series of TU base counts of three overlapping TUs within the *ahpCF* operon; (B) usage of 3 *ahpC* promoters (10-base average from TSS +1 to +10) during logarithmic phase (time point 4); (C) TU coverage time series of the *ybfE-fldA-uof-fur* operon; (D) differential usage of three promoters within the *ybfE-fldA-uof-fur* operon during log phase. Promoter usage and TU coverage calculations are described in the legend to Fig. 1.

that arises from internal promoters and terminators (see Table S6 in the supplemental material). Of 548 complex operons displaying multiple TUs due to having multiple promoters or terminators (Fig. 4), 327 showed more than 2-fold differential expression of 1 TU compared to other TUs within the same operon (see Table S6). Of 633 operons containing more than one gene, we observed 2-fold or greater differential gene expression within 315 operons (e.g., see Fig. 4). In the case of polycistronic operons that have only a single promoter and terminator, it appears that differential decay of the processed transcripts is responsible. In total, 43% (642 of 1,510) of all *E. coli* operons show a complex gene expression regulatory pattern (see Table S6). Clearly, differential expression of TUs and genes within the same operon is common in *E. coli*.

Our analyses provided the opportunity to map potential AS transcription across the transcriptome. In many cases, AS transcripts completely overlap and are complementary to sense strand transcripts that encode proteins; however, these AS transcripts do not appear to encode proteins. For example, the long AS RNA that is complementary to *sulA* does not appear to be translated, because it has no properly positioned ribosome binding site nearby a start codon and thus is likely to be a long noncoding RNA (lncRNA). We found 18 transcripts for annotated protein-coding genes and small RNAs that completely overlap operons transcribed in the opposite direction (see Table S5). As a result of this arrangement, the 18 corresponding operons contain long noncoding AS transcripts that overlap the coding sequences on the opposite strand.

Since genome annotation relies heavily on identification of coding sequences, we predicted that our transcriptome analysis would reveal unannotated genes. Indeed, we identified 96 transcripts that do not correspond to genes on the reference genome and were previously unannotated in *E. coli* K-12 (see Table S5). These include 89 AS transcripts that have an average length of 397 bases, the longest of which is 1,168 bases. The remaining 7 transcripts are completely intergenic and do not overlap annotated genes. None of the 96 transcripts appear to code for protein because they all have multiple stop codons in all three reading frames. Of the 89 AS transcripts, 21 are convergent with known operons that code for proteins, 7 are divergent with mapped operons, and 40 completely overlap annotated operons. The remaining 21 AS transcripts overlap known genes that could not be annotated into operons by RNA-Seq. The genomic regions corresponding to 72% of these lncRNAs are highly conserved in >50 *E. coli* and *Shigella* genomes. It was proposed previously that bacterial lncRNAs may be functional (30, 34), yet this was questioned by others (25). Similar lncRNAs have also been found in eukaryotes, and although they are not well understood, they are thought to play a role in regulating gene expression (70).

A recent study of terminator efficiency showed that only 3% of

E. coli terminators are "strong" (71). Inefficient termination would explain how convergent operons sometimes have overlapping transcription (19, 20). Therefore, we tested the hypothesis that partial termination between convergent operons would generate complementary 3′ transcript ends and add further complexity to the transcriptome. Figure 1 shows an intrinsic terminator located between convergent operons, which terminates transcription by 4-fold yet allows readthrough transcription of 329 bases of AS RNA for the 3′ end of the convergent operon transcript. Our analyses of 370 instances of convergent operons revealed that 75% show transcription into an adjacent operon to generate complementary 3′ transcript ends that overlap by an average of 286 bases, the longest of which is 1,395 bases (see Table S5). In genome regions where there are many highly transcribed operons, it is more likely to observe convergent transcription. Of the genome regions corresponding to these convergent operons, 74% are highly conserved at the nucleotide sequence level in >50 *E. coli* (and *Shigella*) complete genomes. It is thus reasonable to suggest that overlapping transcription of convergent operons is a general property in bacteria.

Transcription of divergent operons can result in overlapping transcripts (17, 18). Complementary transcripts generated by divergent promoters recently have been called "excludons," which are thought to act as negative regulators of genes on the opposite strand (34). Our analyses of 388 instances of divergent operons revealed that 35% have promoters arranged such that their 5′ transcript ends overlap by an average of 168 bases, the longest of which is 1,012 bases (see Table S5). The genome regions corresponding to 81% of these overlapping divergent operons are highly conserved in >50 *E. coli* (and *Shigella*) genomes. The finding of sequence conservation says nothing about functionality, but our finding that over one-third of divergent operons generate overlapping complementary transcripts does suggest that excludons may be prevalent in bacteria.

Comparison to other data sets. We compared our AS transcript annotations to other high-quality data sets using a conservative approach. We compared our data sets to highly expressed and experimentally verified AS transcripts from those studies. A contemporaneous single-nucleotide analysis of the *E. coli* transcriptome by Storz, Sharma, and colleagues (submitted for publication) focused on AS transcripts. They found that most previously annotated sRNAs are present at high levels, so we compared our AS RNA data set to the most highly expressed AS RNAs in their study. Our data corroborate 74 of their 127 highest-expressed AS RNAs. Furthermore, we corroborated 6 of 14 candidate AS RNAs tested on Northern blots by the Storz group. However, while their gels verified 6 of the 14, we corroborated only 2 of those 6, indicating that there is substantial variability in these high-throughput data sets. A recent coimmunoprecipitation study of the double-stranded *E. coli* transcriptome revealed 316 double-stranded RNAs, including partially and fully overlapping transcripts as well as many generated by divergent and convergent operons (33). Our analyses predicted AS RNAs corresponding to 13 of 21 double-stranded RNAs that were verified in Northern blot analysis (33). It is tempting to speculate that AS RNAs that are corroborated by RNA-Seq studies, are verified by Northern blot analysis, and correspond to highly conserved genome sequences are functional. However, functions have been confirmed for only a limited number of AS RNAs (56, 72). It remains to be seen how many of the AS RNAs identified by RNA-Seq will prove to be expressed in the same cell as the sense transcript and display a yet unknown phenotype.

Bacterial operons compared to eukaryotic genes. It did not escape our attention that the widespread occurrence of bacterial operons with multiple TUs in some ways resembles alternative splicing of eukaryotic transcripts. From both bacterial operons and eukaryotic genes arise primary transcripts that are divided into alternative transcripts by the activity of transcriptional regulatory features, i.e., internal promoters and terminators in bacteria and RNA splice junctions in eukaryotes. The potential for eukaryotic gene complexity is reflected in the number of exons per gene. The number of exons per gene in *Saccharomyces cerevisiae* is 1.1 (73), which is considerably fewer than the 1.7 TUs per operon in *E. coli*. In contrast, higher organisms have 4 to 9 introns per gene (74), making them more complex than *E. coli*. Perhaps the loss of exons that is proposed to have happened in budding yeasts during evolution from more primitive eukaryotes accentuates their difference from *E. coli* and higher organisms (75). We conclude that *E. coli* possesses operon complexity comparable to analogous gene structures in budding yeasts.

Concluding statement. This study reveals the power of single-nucleotide resolved RNA-Seq data sets for pinpointing transcriptional features across the genome, which we used to annotate operons by precisely mapping their 5′ and 3′ ends. We found an astounding level of overlapping transcription where complementary RNAs are transcribed from both strands, such as those generated by several hundred convergent and divergent operons. We discovered more than 100 long AS transcripts overlapping operons that also were transcribed on the sense strand. In sum, we found that approximately one in three (519 out of 1,510) operons at least partially overlaps with other operons to generate AS RNA. These AS transcripts are highly conserved in *E. coli* and appear to be noncoding RNA, suggesting that they are involved in regulation of gene expression, as has been proposed for excludons in bacteria (34) and lncRNAs in eukaryotes (70). We also found 7 transcripts that did not correspond to an annotated gene and therefore represent previously unrecognized yet potentially functional operons. The transcriptome intricacy we observed in *E. coli* appears to be a general property of the domain bacteria, as the transcriptomes of several other bacteria appear to be similarly intricate (21, 26, 28, 31, 41, 47–49). Whether the same is true of the *Archaea* must await high-resolution RNA-Seq analysis of representatives of this domain of life (83). Since operon arrangements are more highly conserved than gene repertoires (76), it is interesting to speculate that the requirements of primordial life led to the evolution of an operon architecture in bacteria which accommodates substantial variation in gene expression.

MATERIALS AND METHODS

Bacterial strains and growth conditions. To annotate operons and characterize their response to carbon starvation, wild-type *E. coli* BW38028 and *E. coli* BW39452 (Δ*rpoS*::*cat*) were grown in 1 liter of morpholinepropanesulfonic acid (MOPS) minimal medium (77) containing 0.2% glucose in a fermenter at 37°C with constant pH and aeration. MOPS medium solutions were modified as described elsewhere (78), which permits preparation of 40× "M" stock solution, giving the same final medium recipe (77). Cultures were sampled at 10 time points during growth of *E. coli* BW38028 and at five time points for *E. coli* BW39452, as shown in Fig. S1 in the supplemental material. Logarithmic- and stationary-phase samples were duplicated from replicate cultures.

RNA sequencing. RNA was prepared by using an RNeasy kit (Qiagen, USA). Because small RNAs may be preferentially lost during column purification, they are likely underrepresented in our data sets. Replicates of logarithmic- and stationary-phase RNA were treated with Terminator 5′-phosphate-dependent exonuclease (Epicenter, USA) to enrich 5′-triphosphate mRNA fragments for TSS mapping. RNA sequencing libraries (see Table S1) were prepared by using the strand-specific, ligation-based SOLiD Total RNA-Seq kit. Paired-end sequencing was performed on the SOLiD 4 Genetic Analyzer at Purdue University Genomics Facility.

Raw data processing. Sequence reads were aligned to the *E. coli* MG1655 reference genome (U00096.2) with Bowtie version 1.8 (79). For the first pass, we used paired-end color space mapping with a distance cutoff of 350 bases between read mates. Bowtie parameters were set to include only perfect matches and retained only one alignment where a read mapped to more than one genome location. In practice, we found the efficiency of paired-end mapping was between 3 and 10%. To improve the overall alignment, we mapped the orphan 5′- and 3′-end reads in two additional passes with Bowtie (one for the 5′ reads and one for the 3′ reads). The output of the three passes through Bowtie was three SAM files for each sample. Overall, we achieved 40 to 60% mapping efficiency with this three-pass strategy. SAMtools (80) utilities were used to convert SAM files to BAM format and to sort and index them. The binary read alignment (BAM) files were displayed in Integrated Genome Viewer (IGV version 2) for primary analysis and quality control. The BAM files were then converted to base count (WIG) files. We accomplished this by using an in-house script to extract strand-specific base count data from BAM files (outputs are positive- and negative-strand WIG files). First, our solidbam2wig.pl script reads in the paired-end BAM file and counts the nucleotides spanning inserts between the mated 5′ and 3′ reads. Next, the script pulls in the orphan 5′ and 3′ reads from the respective BAM files and increments the base counts at each base location without duplicating the reads already incremented from the paired ends. Base count data were then normalized based on the assumption that reads are randomly distributed across the genome and that if sequencing was sufficiently deep, all expressed transcripts would be represented in the data set (39). In practice, SOLiD sequencing did not generate data sets in which the lowest-abundance transcripts were fully covered by contiguous reads. In addition, inefficient ribo-depletion can bias the number of reads that map to non-rRNA genes. Our normalization strategy accounts for both of these factors by maximizing TU coverage and removing rRNA reads during data processing. Our in-house script, normWIG.pl, reads in the raw WIG files. A simple global normalization approach was utilized that multiplied the count at each base location by 1 billion and divides that value by the sum of base counts at all base locations in the file. This normalization strategy is analogous to the total count approach used for normalizing gene-specific read alignments (51). In this way, the base counts are expressed as parts per billion. For display in JBrowse (61), the normalized WIG files were converted to BIGWIG files by using SAMtools (80). Analysis of the data was conducted in a graphic user interface consisting of JBrowse (61) and an Oracle database.

SELEX. Genomic SELEX was previously described (81). Antibodies against RpoD sigma, RpoS sigma, and core enzyme subunits were produced in rabbits by injecting purified sigma proteins (82).

Nucleotide sequence accession number. RNA sequencing data and curated results were deposited at Gene Expression Omnibus, accession no. GSE52059.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.01442-14/-/DCSupplemental.

Table S1, XLSX file, 0 MB.
Table S2, XLSX file, 0.6 MB.
Table S3, XLSX file, 0.2 MB.
Table S4, XLSX file, 0.3 MB.
Table S5, XLSX file, 0.1 MB.
Table S6, XLSX file, 0.6 MB.
Figure S1, PDF file, 0.1 MB.
Figure S2, PDF file, 0.5 MB.
Figure S3, PDF file, 0.1 MB.
Figure S4, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

This work was funded primarily by U.S. Public Health Service NIH RC1GM09207 to B.L.W. and T.C. from 2009 to 2011. B.L.W. is currently supported by NSF award 106394. Additional support was from NIH GM095370 to T.C., Grants-in-Aid for Scientific Research 21710198 to T.S. and 17076016, 8310133, and 21241047 to A.I. from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Nano-Biology Project fund from Micro-Nanotechnology Research Center of Hosei University to A.I., Grant-in-Aid for Scientific Research 22241050 and 25250028, Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research on Innovative Areas 25108716, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and Grant-in-Aid for Scientific Research on Priority Areas to H.M.

We thank Jay C. D. Hinton for helpful comments during manuscript preparation. We dedicate this work to the memory of Monica Riley (1926 to 2013), a true pioneer of *E. coli* genome annotation.

REFERENCES

1. Lederberg J, Tatum EL. 1946. Gene recombination in *Escherichia coli*. Nature 158:558. http://dx.doi.org/10.1038/158558c0.
2. Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3:318–356. http://dx.doi.org/10.1016/S0022-2836(61)80072-7.
3. Lehman IR, Bessman MJ, Simms ES, Kornberg A. 1958. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. J. Biol. Chem. 233:163–170.
4. Lengyel P, Speyer JF, Ochoa S. 1961. Synthetic polynucleotides and the amino acid code. Proc. Natl. Acad. Sci. U. S. A. 47:1936–1942. http://dx.doi.org/10.1073/pnas.47.12.1936.
5. Luria SE, Delbrück M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491–511.
6. Arber W, Dussoix D. 1962. Host specificity of DNA produced by *Escherichia coli*. I. Host controlled modification of bacteriophage lambda. J. Mol. Biol. 5:18–36. http://dx.doi.org/10.1016/S0022-2836(62)80058-8.
7. Mulligan RC, Berg P. 1980. Expression of a bacterial gene in mammalian cells. Science 209:1422–1427. http://dx.doi.org/10.1126/science.6251549.
8. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31:147–157. http://dx.doi.org/10.1016/0092-8674(82)90414-7.
9. Boyer PD, Cross RL, Momsen W. 1973. A new concept for energy coupling in oxidative phosphorylation based on a molecular explanation of the oxygen exchange reactions. Proc. Natl. Acad. Sci. U. S. A. 70:2837–2839. http://dx.doi.org/10.1073/pnas.70.10.2837.
10. Chang CN, Model P, Blobel G. 1979. Membrane biogenesis: cotranslational integration of the bacteriophage f1 coat protein into an *Escherichia coli* membrane fraction. Proc. Natl. Acad. Sci. U. S. A. 76:1251–1255. http://dx.doi.org/10.1073/pnas.76.3.1251.
11. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462. http://dx.doi.org/10.1126/science.277.5331.1453.
12. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G III, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res. 34:1–9. http://dx.doi.org/10.1093/nar/gnj001.
13. Balázsi G, Barabási AL, Oltvai ZN. 2005. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 102:7841–7846. http://dx.doi.org/10.1073/pnas.0500365102.
14. Price MN, Arkin AP, Alm EJ. 2006. The life-cycle of operons. PLoS Genet. 2:e96. http://dx.doi.org/10.1371/journal.pgen.0020096.

15. Zhang H, Yin Y, Olman V, Xu Y. 2012. Genomic arrangement of regulons in bacterial genomes. PLoS One 7:e29496. http://dx.doi.org/10.1371/journal.pone.0029496.
16. Taylor K, Hradecna Z, Szybalski W. 1967. Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA. Proc. Natl. Acad. Sci. U. S. A. 57:1618–1625. http://dx.doi.org/10.1073/pnas.57.6.1618.
17. Piette J, Cunin R, Boyen A, Charlier D, Crabeel M, Van Vliet F, Glansdorff N, Squires C, Squires CL. 1982. The regulatory region of the divergent *argECBH* operon in *Escherichia coli* K-12. Nucleic Acids Res. 10:8031–8048. http://dx.doi.org/10.1093/nar/10.24.8031.
18. Wek RC, Hatfield GW. 1986. Nucleotide sequence and *in vivo* expression of the *ilvY* and *ilvC* genes in *Escherichia coli* K-12. Transcription from divergent overlapping promoters. J. Biol. Chem. 261:2441–2450.
19. Nomura T, Aiba H, Ishihama A. 1985. Transcriptional organization of the convergent overlapping *dnaQ-rnh* genes of *Escherichia coli*. J. Biol. Chem. 260:7122–7125.
20. Sameshima JH, Wek RC, Hatfield GW. 1989. Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli*. J. Biol. Chem. 264:1224–1231.
21. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464:250–255. http://dx.doi.org/10.1038/nature08756.
22. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM. 2000. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nat. Biotechnol. 18:1262–1268. http://dx.doi.org/10.1038/82367.
23. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT. 2010. Widespread antisense transcription in *Escherichia coli*. mBio 1(1):e00024-10. http://dx.doi.org/10.1128/mBio.00024-10.
24. Wade JT, Dornenburg JE, Devita AM, Palumbo MJ. 2010. Reply to "Concerns about recently identified widespread antisense transcription in *Escherichia coli*." mBio 1(2):e00119-10. http://dx.doi.org/10.1128/mBio.00119-10.
25. Raghavan R, Sloan DB, Ochman H. 2012. Antisense transcription is pervasive but rarely conserved in enteric bacteria. mBio 3(4):e00156-12. http://dx.doi.org/10.1128/mBio.00156-12.
26. Behrens S, Widder S, Mannala GK, Qing X, Madhugiri R, Kefer N, Mraheil MA, Rattei T, Hain T. 2014. Ultra deep sequencing of *Listeria monocytogenes* sRNA transcriptome revealed new antisense RNAs. PLoS One 9:e83979. http://dx.doi.org/10.1371/journal.pone.0083979.
27. Chatterjee A, Johnson CM, Shu CC, Kaznessis YN, Ramkrishna D, Dunny GM, Hu WS. 2011. Convergent transcription confers a bistable switch in *Enterococcus faecalis* conjugation. Proc. Natl. Acad. Sci. U. S. A. 108:9721–9726. http://dx.doi.org/10.1073/pnas.1101569108.
28. Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L. 2009. Transcriptome complexity in a genome-reduced bacterium. Science 326:1268–1271. http://dx.doi.org/10.1126/science.1176951.
29. Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC. 2012. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proc. Natl. Acad. Sci. U. S. A. 109:E1277–E1286. http://dx.doi.org/10.1073/pnas.1201061109.
30. Lasa I, Toledo-Arana A, Dobin A, Villanueva M, de los Mozos IR, Vergara-Irigaray M, Segura V, Fagegaltier D, Penadés JR, Valle J, Solano C, Gingeras TR. 2011. Genome-wide antisense transcription drives mRNA processing in bacteria. Proc. Natl. Acad. Sci. U. S. A. 108:20172–20177. http://dx.doi.org/10.1073/pnas.1113521108.
31. Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH. 2012. Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. PLoS One 7:e43350. http://dx.doi.org/10.1371/journal.pone.0043350.
32. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P, Sorek R. 2012. Comparative transcriptomics of pathogenic and nonpathogenic *Listeria* species. Mol. Syst. Biol. 8:583. http://dx.doi.org/10.1038/msb.2012.11.
33. Lybecker M, Zimmermann B, Bilusic I, Tukhtubaeva N, Schroeder R. 2014. The double-stranded transcriptome of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 111:3134–3139. http://dx.doi.org/10.1073/pnas.1315974111.
34. Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P. 2013. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol. 11:75–82. http://dx.doi.org/10.1038/nrmicro2934.
35. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ. 2009. The transcription unit architecture of the *Escherichia coli* genome. Nat. Biotechnol. 27:1043–1049. http://dx.doi.org/10.1038/nbt.1582.
36. Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BØ. 2012. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. PLoS Genet. 8:e1002867. http://dx.doi.org/10.1371/journal.pgen.1002867.
37. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E. 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One 4:e7526. http://dx.doi.org/10.1371/journal.pone.0007526.
38. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J. 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 41:D203–D213. http://dx.doi.org/10.1093/nar/gkt1054.
39. Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13:734. http://dx.doi.org/10.1186/1471-2164-13-734.
40. Li S, Dong X, Su Z. 2013. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K-12 through accurate full-length transcripts assembling. BMC Genomics 14:520. http://dx.doi.org/10.1186/1471-2164-14-520.
41. Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, Canals R, Grissom JE, Conway T, Hokamp K, Hinton JC. 2013. An infection-relevant transcriptomic compendium for *Salmonella enterica* serovar Typhimurium. Cell Host Microbe 14:683–695. http://dx.doi.org/10.1016/j.chom.2013.11.010.
42. Fabich AJ, Jones SA, Chowdhury FZ, Cernosek A, Anderson A, Smalley D, McHargue JW, Hightower GA, Smith JT, Autieri SM, Leatham MP, Lins JJ, Allen RL, Laux DC, Cohen PS, Conway T. 2008. Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. Infect. Immun. 76:1143–1152. http://dx.doi.org/10.1128/IAI.01386-07.
43. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7:709–715. http://dx.doi.org/10.1038/nmeth.1491.
44. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. 27:3821–3835. http://dx.doi.org/10.1093/nar/27.19.3821.
45. Tao H, Bausch C, Richmond C, Blattner FR, Conway T. 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol. 181:6425–6440.
46. Reference deleted.
47. Lin YF, A DR, Guan S, Mamanova L, McDowall KJ. 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease. BMC Genomics 14:620. http://dx.doi.org/10.1186/1471-2164-14-620.
48. Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H. 2013. RNA-Seq of *Bacillus licheniformis*: active regulatory RNA features expressed within a productive fermentation. BMC Genomics 14:667. http://dx.doi.org/10.1186/1471-2164-14-667.
49. Balasubramanian D, Kumari H, Jaric M, Fernandez M, Turner KH, Dove SL, Narasimhan G, Lory S, Mathee K. 2014. Deep sequencing

analyses expands the *Pseudomonas aeruginosa* AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response. Nucleic Acids Res. 42:979–998. http://dx.doi.org/10.1093/nar/gkt942.
50. Bohannon DE, Connell N, Keener J, Tormo A, Espinosa-Urgel M, Zambrano MM, Kolter R. 1991. Stationary-phase-inducible "gearbox" promoters: differential effects of *katF* mutations and role of sigma 70. J. Bacteriol. 173:4482–4492.
51. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14:671–683. http://dx.doi.org/10.1093/bib/bbs046.
52. Egan SE, Fliege R, Tong S, Shibata A, Wolf RE, Jr, Conway T. 1992. Molecular characterization of the Entner-Doudoroff pathway in *Escherichia coli*: sequence analysis and localization of promoters for the *edd-eda* operon. J. Bacteriol. 174:4638–4646.
53. Deana A, Celesnik H, Belasco JG. 2008. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. Nature 451:355–358. http://dx.doi.org/10.1038/nature06475.
54. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. 23:137–144. http://dx.doi.org/10.1038/nbt1053.
55. Shimada T, Yamamoto K, Ishihama A. 2011. Novel members of the *cra* regulon involved in carbon metabolism in *Escherichia coli*. J. Bacteriol. 193:649–659. http://dx.doi.org/10.1128/JB.01214-10.
56. Thomason MK, Storz G. 2010. Bacterial antisense RNAs: how many are there, and what are they doing? Annu. Rev. Genet. 44:167–188. http://dx.doi.org/10.1146/annurev-genet-102209-163523.
57. Tuerk C, MacDougal S, Gold L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. Proc. Natl. Acad. Sci. U. S. A. 89:6988–6992. http://dx.doi.org/10.1073/pnas.89.15.6988.
58. Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:-1017–1018. http://dx.doi.org/10.1093/bioinformatics/btr064.
59. Robison K, McGuire AM, Church GM. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284:241–254. http://dx.doi.org/10.1006/jmbi.1998.2160.
60. Mitchell JE, Zheng D, Busby SJ, Minchin SD. 2003. Identification and analysis of "extended −10" promoters in *Escherichia coli*. Nucleic Acids Res. 31:4689–4695. http://dx.doi.org/10.1093/nar/gkg694.
61. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. Genome Res. 19:-1630–1638. http://dx.doi.org/10.1101/gr.094607.109.
62. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U. S. A. 99:9697–9702. http://dx.doi.org/10.1073/pnas.112318199.
63. Kingsford CL, Ayanbule K, Salzberg SL. 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. 8:R22. http://dx.doi.org/10.1186/gb-2007-8-2-r22.
64. Potrykus K, Murphy H, Chen X, Epstein JA, Cashel M. 2010. Imprecise transcription termination within *Escherichia coli greA* leader gives rise to an array of short transcripts, GraL. Nucleic Acids Res. 38:1636–1651. http://dx.doi.org/10.1093/nar/gkp1150.
65. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. GenBank. Nucleic Acids Res. 42:D32–D37. http://dx.doi.org/10.1093/nar/gkt1030.
66. Cole ST, Honoré N. 1989. Transcription of the *sulA-ompA* region of *Escherichia coli* during the SOS response and the role of an antisense RNA molecule. Mol. Microbiol. 3:715–722. http://dx.doi.org/10.1111/j.1365-2958.1989.tb00220.x.
67. Michán C, Manchado M, Dorado G, Pueyo C. 1999. *In vivo* transcription of the *Escherichia coli oxyR* regulon as a function of growth phase and in response to oxidative stress. J. Bacteriol. 181:2759–2764.
68. Vecerek B, Moll I, Bläsi U. 2007. Control of fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. EMBO J. 26:965–975. http://dx.doi.org/10.1038/sj.emboj.7601553.
69. Cunningham L, Guest JR. 1998. Transcription and transcript processing in the *sdhCDAB-sucABCD* operon of *Escherichia coli*. Microbiology 144(Part 8):2113–2123. http://dx.doi.org/10.1099/00221287-144-8-2113.
70. Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. Cell 136:629–641. http://dx.doi.org/10.1016/j.cell.2009.02.006.
71. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA. 2013. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat. Methods 10:659–664. http://dx.doi.org/10.1038/nmeth.2515.
72. Georg J, Hess WR. 2011. Cis-antisense RNA, another level of gene regulation in bacteria. Microbiol. Mol. Biol. Rev. 75:286–300. http://dx.doi.org/10.1128/MMBR.00032-10.
73. Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. Genetics 174:511–518. http://dx.doi.org/10.1534/genetics.106.058560.
74. Koralewski TE, Krutovsky KV. 2011. Evolution of exon-intron structure and alternative splicing. PLoS One 6:e18055. http://dx.doi.org/10.1371/journal.pone.0018055.
75. Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Patterns of intron gain and conservation in eukaryotic genes. BMC Evol. Biol. 7:192. http://dx.doi.org/10.1186/1471-2148-7-192.
76. Rocha EP. 2008. The organization of the bacterial genome. Annu. Rev. Genet. 42:211–233. doi:http://dx.doi.org/10.1146/annurev.genet.42.110807.091653.
77. Neidhardt FC, Bloch PL, Smith DF. 1974. Culture medium for enterobacteria. J. Bacteriol. 119:736–747.
78. Wilmes-Riesenberg MR, Wanner BL. 1992. TnphoA and TnphoA' elements for making and switching fusions for study of transcription, translation, and cell surface localization. J. Bacteriol. 174:4558–4575.
79. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25.
80. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
81. Shimada T, Fujita N, Maeda M, Ishihama A. 2005. Systematic search for the cra-binding promoters using genomic SELEX system. Genes Cells 10:907–918. http://dx.doi.org/10.1111/j.1365-2443.2005.00888.x.
82. Jishage M, Ishihama A. 1995. Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. J. Bacteriol. 177:6832–6835.
83. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010. A single-base resolution map of an archaeal transcriptome. Genome Res. 20:133–141.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Quantitative bacterial transcriptomics with RNA-seq

James P Creecy[1,2] and Tyrrell Conway[1]

RNA sequencing has emerged as the premier approach to study bacterial transcriptomes. While the earliest published studies analyzed the data qualitatively, the data are readily digitized and lend themselves to quantitative analysis. High-resolution RNA sequence (RNA-seq) data allows transcriptional features (promoters, terminators, operons, among others) to be pinpointed on any bacterial transcriptome. Once the transcriptome is mapped, the activity of transcriptional features can be quantified. Here we highlight how quantitative transcriptome analysis can reveal biological insights and briefly discuss some of the challenges to be faced by the field of bacterial transcriptomics in the near future.

Addresses
[1] Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, United States
[2] Department of Biology, University of Central Oklahoma, Edmond, OK 73034, United States

Corresponding author: Conway, Tyrrell (tconway@ou.edu)

Current Opinion in Microbiology 2015, **23**:133–140

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay Hinton**

http://dx.doi.org/10.1016/j.mib.2014.11.011

RNA-seq comes of age

Advances in RNA sequencing technology have revolutionized the study of bacterial transcriptomes [1,2••]. At its core, RNA sequence (RNA-seq) generates digital information that allows transcriptional features to be located with single-nucleotide precision in a strand specific manner. Since the data are digital, RNA-seq facilitates quantitative computational analysis of any selected region of the transcriptome, but the transcriptome must first be annotated properly. Since bacterial genomes are organized in operons, it is logical that RNA-seq data should be annotated with the operon architecture in mind. In practice, only three transcriptional features need to be defined: 5′ transcript ends (promoters), 3′ ends (terminators), and RNA-seq read coverage to connect the ends, which together define operons [3•,4••,5].

The true power of RNA-seq resides in its potential as an analytical tool for quantifying promoter activity, terminator efficiency, and differential expression of transcripts, including operons, transcription units within operons (e.g. generated by promoters internal to operons), and antisense RNAs. As described in more detail below, RNA-seq datasets consist of tens of millions of sequence reads and typically the reads are 50 bases in length. The raw sequence reads are aligned to a reference genome and only high quality reads are retained and mapped. Conversion of sequence data into digital format is accomplished by employing freely available computer scripts that count the number of times each transcribed base was sequenced in a read-aligned dataset, thereby converting aligned sequence reads to base count data. Normalization of the base count data is necessary to quantify the differential expression (i.e. relative base counts) of each transcriptional feature within a sample or between different samples. The normalized base count data can be quantified by averaging the base count across a selected region of the genome. Since the average of the base counts is used, the relative expression of any given transcription feature, regardless of its length, can be expressed in this way. Here we focus on the analysis of an *Escherichia coli* RNA-seq dataset to demonstrate the strategy we developed to quantify the expression of the transcriptional features that define operons in bacteria.

Single-nucleotide resolved RNA-seq dataset

To obtain an RNA-seq dataset suitable for quantitative analysis, we prepared RNA from a culture of *E. coli* K-12 strain BW38028 during logarithmic-phase and stationary-phase growth on glucose limited minimal medium, as described previously [4••]. In addition, we starved *E. coli* BW38028 and its isogenic *rpoS* mutant BW39452 for nitrogen by decreasing by three-fold the amount of ammonium chloride in the growth medium [6]. The RNA was extracted by using the hot-phenol method [7•] and DNase I treated to remove contaminating DNA. The RNA samples were not depleted for rRNA prior to sequencing, which tends to eliminate some experimental biases [8]. The RNA samples were shipped on dry ice to vertis Biotechnologie AG (Germany) for library preparation and Illumina HiSeq2000 sequencing, as described by others [7•,9••]. For library preparation the RNA samples were split and subjected to differential RNA-seq (dRNA-seq) as described [2••,10••]. Briefly, one portion of the RNA was fragmented by ultrasound and then the fragments were poly(A)-tailed and an RNA adapter was ligated to the 5′ phosphate of the RNA. First strand cDNA synthesis was with a poly(dT) primer and reverse transcriptase. Second strand cDNA synthesis incorporated a barcoded 3′ TruSeq adapter. The other portion of the RNA samples were fragmented and treated with terminator exonuclease (TEX), which enriches for 5′ triphosphate

containing transcripts that are generated by transcription initiation at promoters. The TEX treated samples then were tailed and ligated, and cDNA was prepared as described above. The cDNAs were sequenced on an Illumina HiSeq2000 system using 50 bp read length, with each library yielding approximately 20 million reads.

Datasets consisting of 10 million reads per sample are sufficient for transcriptional feature mapping and differential gene expression analysis without ribo-depletion for a transcriptome the size of *E. coli* [9••,11]. For quantification the genome-aligned, strand-specific RNA-seq data should be converted from aligned reads to base counts. Our RNA-seq data analysis pipeline involves alignment of the raw data to the reference genome by using Bowtie2 to generate the sequence read alignment file (SAM) [12]. SAMTOOLS [13] were used to convert the SAM file to a binary alignment file (BAM). The BAM file was converted to a BigWig file (base count file), which contains the count of the base at each base location and is the standard for visualization in genome browsers such as J-Browse [14]. Conversion of BAM to BigWig formatted files can be accomplished by using tools available in the Galaxy Toolshed [15] or at UCSC Genome Browser [16].

Alternatively, users can analyze their datasets by using pipelines such as Galaxy [17] or READemption [18•], which outputs normalized wiggle files (base count files). A simple and straightforward way to normalize base count data is by using a strategy analogous to the total count approach [19] for normalizing gene-specific read alignments, which expresses each value as the base count per billion bases counted [4••]. Because the BigWig file represents the base count at each nucleotide position, all downstream analysis begins with this file. The advantages of the base count approach are: first, the digital base count data are inherently computable because of their format and smaller size, second, the average base counts of individual transcriptional features can be computed and queried at any desired resolution, from a single nucleotide to an entire operon, to quantify the expression level or activity, third, normalization of base count data makes all samples directly comparable, and fourth, the use of average base count values eliminates the length bias when comparing transcriptional features of different length [19].

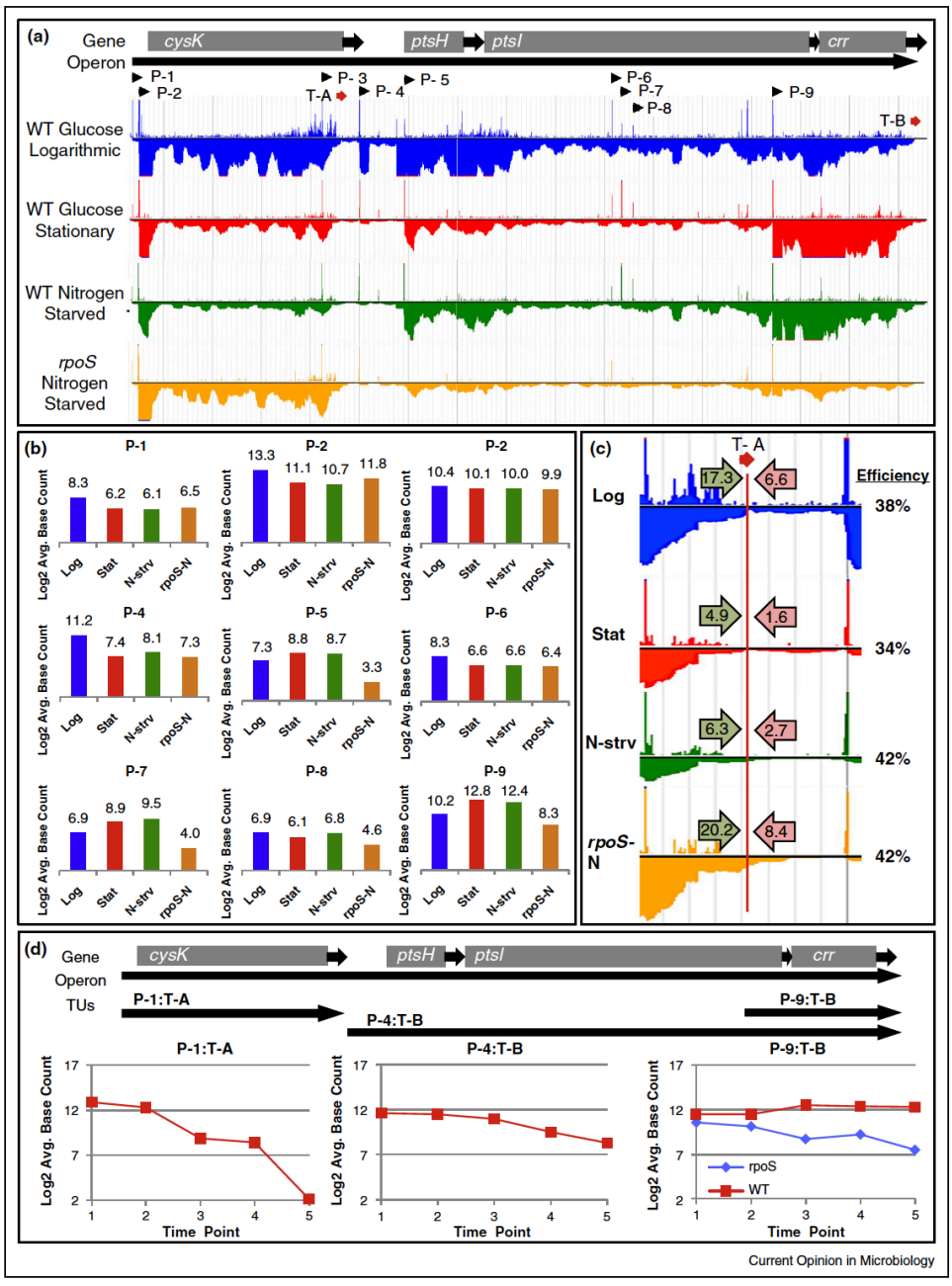
Identification of transcription start sites

Several published RNA-seq studies have focused on transcription start site (TSS) identification [7•,9••,10••,20•,21,22,23•,24•,25,26•,27•,28•]. The annotation of TSSs is essential for analyzing promoters, 5′ UTRs, operon architecture, and for discovering novel transcripts. To assure accuracy, a set of 'best practices' for TSS identification is emerging. Enrichment of the 5′ RNA ends that are generated by transcription initiation is critical for accurate TSS identification. The many advantages of dRNA-seq were recently reviewed [2••]. The initiating nucleotide in bacteria is a nucleotide triphosphate, which can be distinguished from 5′-monophosphate and 5′-OH containing RNAs that are generated by RNA processing or RppH pyrophosphohydrolase activity [29]. The enrichment strategy preferred by many researchers makes use of 5′ monophosphate-dependent terminator exonuclease (TEX), which degrades RNA with 5′ monophosphate ends to enrich for primary transcripts that contain 5′ triphosphate ends and hence represent the product of transcription initiation [10••]. dRNA-seq works by enumerating differences in base counts between TEX-enriched and unenriched sequencing libraries. Experimental replication is critical for accurate TSS identification. Since dRNA-seq is remarkably reproducible, comparison of datasets generated by using the same protocols yet different growth conditions adds confidence to the process and the use of different growth conditions also increases the number of mapped TSSs. RNA samples from many growth conditions can be pooled for dRNA-seq identification of thousands of promoters [9••]. For example, a recent dRNA-seq analysis of Salmonella using RNA pooled from 22 different growth conditions led to mapping of 96% of the TSSs that could be identified by independently analyzing the 22 samples [9••].

When annotating transcriptome data, it is convenient to use widely available computer programs to search dRNA-seq datasets for TSSs [20•,30,31]. The advantages of the computational process compared to manual annotation are the speed and precision of recording transcription feature locations. However, like all bioinformatics approaches, some features will be missed and there will be false positives. In the end, human supervision of the results is critical and the state-of-the-art in transcriptome annotation remains a manual process [9••]. Manual annotation of TSSs is made more efficient by plotting the count of only the first base at the 5′ end of each TEX-enriched read (Figure 1a) [32••]. In practice this allows visualization of the 5′ triphosphate nucleotide at the TSS.

Subsequent to identification of TSSs by dRNA-seq, bioinformatics and functional analyses can add weight to promoter identification. For example, the DNA sequences immediately upstream of putative TSSs can be analyzed by using a bioinformatics approach to score sigma factor specific RNA polymerase binding sequence motifs [4••,33]. ChIP determination of RNA polymerase binding provides a robust and comprehensive validation of putative promoters [23•]. When used in combination, dRNA-seq, consensus amongst experimental replicates, promoter sequence analysis, and RNA polymerase binding assays are a powerful set of tools for the identification of promoters.

Figure 1



Transcriptional feature map and analysis of the *cysK-ptsHI-crr* operon. The dRNA-seq data are available at GEO, GSE58556. **(a)** The genes and feature locations are drawn to scale and annotated to the positive strand of the *E. coli* MG1655 U00096.3 reference genome. Promoters (P) are indicated by an arrow and are numbered in order from left to right on the positive strand. Terminators (T) are indicated by a diamond. The base count data, consisting of TEX-treated samples pointing up and unenriched coverage data (fragmented RNA not treated with TEX) pointing down,

Annotation of 3′ ends

To obtain the full analytical value of RNA-seq data it is essential to map the 3′ transcript ends. Annotating 3′ ends is a notably more difficult endeavor than mapping TSSs because there currently is no method of enriching for them. The 3′ ends are the primary sites of exonuclease-dependent RNA decay, which may be the reason that RNA base counts decline at the 3′ ends of operons, and few reads extend into the stem loop structures of intrinsic terminators (Figure 1c). Further complicating 3′ end analysis is that termination is typically inefficient [34], which allows read-through transcription. Currently, the best method for annotating 3′ ends is to search for correlation between replicates of the furthermost downstream bases transcribed, keeping in mind that the base counts near the 3′ end will be low even for highly expressed transcripts. Comparison of the 3′ ends to terminator predictions adds confidence to the analysis. For example, the TransTermHP software package works very well for finding intrinsic terminators [35]. In addition, a ChIP-chip analysis of the distribution of RNA polymerase after treatment with the Rho-specific inhibitor bicyclomycin led to identification of 200 Rho-dependent terminators [36]. Once both the 5′ and 3′ transcript ends are mapped, it is possible to annotate operons.

Annotation of operons

The transcriptome is a map of the activities of promoters and terminators. These activities are located on both strands of the genome [37] and depending on their arrangement, can give rise to antisense transcription and overlapping, divergent [38,39] and convergent operons [40,41]. To accommodate this naturally occurring complexity it is necessary to annotate the operon architecture. Three transcriptional features are necessary to define operons: 5′ ends (promoters), 3′ ends (terminators), and sufficient RNA-seq read coverage to connect the ends. If sequence reads cover 90% of the bases, this is a sensible indicator that the operon is real [4••,32••]. While there are computer algorithms that can find operons [5,42,43•], just as for TSS mapping, the state-of-the-art remains a manual process [9••]. Once the operons have been mapped, it is a straightforward task to annotate additional promoters and terminators within operons, which add complexity to the transcriptome. Mapping of internal promoters can be done manually or by bioinformatics analysis of mapped promoters that fall within the base locations of annotated operons. The transcriptional feature locations can be formatted as a GenBank feature file by using 'promoter', 'terminator' and 'operon' as feature keys (see for example, GSE52059 [4••]). This format accommodates incremental annotation of condition specific regulatory information and is an accepted standard for disseminating genome annotation data [44]. Once the transcriptional feature locations are annotated, it is reasonably straightforward to calculate the average base count value for each feature, from each dataset, as described below.

Computing the activities of transcriptional features

Analysis of RNA-seq reads at the base count level permits normalized base counts to be readily averaged across any range of base locations to calculate the relative expression level, activity, or efficiency of individual transcriptional features [4••]. We determined empirically that computing the average count of the first 10 transcribed bases accurately represents promoter activity and allows closely spaced promoters to be discriminated [4••]. Likewise, the efficiency of transcription termination can be calculated as the relative decline in average base counts in 25-base windows before and after terminators (Figure 1c). The relative transcript levels of operons can be calculated by averaging the base counts from the promoter to the terminator locations. Likewise, the expression levels of alternative transcripts generated by promoter and terminator activities within operons can be calculated. These applications of single-nucleotide-resolution analysis are exemplified in Figure 1, for wild type *E. coli* K-12 during logarithmic growth on glucose minimal medium and during starvation for carbon (stationary phase) or nitrogen, as well as an *rpoS* mutant during nitrogen starvation.

The *cysK-ptsHI-crr* operon contains four genes and multiple transcription units (Figure 1a). Conservatively, more than 40% of *E. coli* operons contain multiple transcription units that are differentially expressed, underscoring the need for an annotation system that accommodates operon architecture [4••]. In addition to the primary promoter (P-1) and terminator (T-B) that define the operon, there are eight additional promoters

(Figure 1 Lengd Continued) are visualized in J-Browse [14], as described previously [4••]. Only positive strand data are shown. Tracks: wild type (WT), glucose-grown *E. coli* K-12 in logarithmic phase (blue track); WT in stationary phase, 30 min after exhaustion of glucose (red track); WT starved for nitrogen (green track); and an isogenic *rpoS* mutant starved for nitrogen (tan track). The base count scale (on the left) is from 0 to 100, with values exceeding 100 indicated by dark red. **(b)** The relative activities of the nine promoters is plotted in the graphs as log 2 average counts of the first 10 transcribed bases under the four different growth conditions, which are colorized as above. **(c)** The decrease in average counts of the 25 bases before and after the terminator T-A are shown by light green and pink arrows. **(d)** Time series analysis of the relative expression levels of three transcripts within the complex *cysK-ptsHI-crr* operon is plotted as the log 2 average counts of bases from the indicated promoters to terminators, as described previously [4••]. Time point 1 is during middle logarithmic phase, time point 2 is immediately prior to entry into stationary phase, time point 3 is 15 min after entry into stationary phase, time point 4 is 30 min after entry into stationary phase, and time point 5 is 180 min after entry into stationary phase. Additional details of the analysis are described in the text.

Annotation of 3′ ends

To obtain the full analytical value of RNA-seq data it is essential to map the 3′ transcript ends. Annotating 3′ ends is a notably more difficult endeavor than mapping TSSs because there currently is no method of enriching for them. The 3′ ends are the primary sites of exonuclease-dependent RNA decay, which may be the reason that RNA base counts decline at the 3′ ends of operons, and few reads extend into the stem loop structures of intrinsic terminators (Figure 1c). Further complicating 3′ end analysis is that termination is typically inefficient [34], which allows read-through transcription. Currently, the best method for annotating 3′ ends is to search for correlation between replicates of the furthermost downstream bases transcribed, keeping in mind that the base counts near the 3′ end will be low even for highly expressed transcripts. Comparison of the 3′ ends to terminator predictions adds confidence to the analysis. For example, the TransTermHP software package works very well for finding intrinsic terminators [35]. In addition, a ChIP-chip analysis of the distribution of RNA polymerase after treatment with the Rho-specific inhibitor bicyclomycin led to identification of 200 Rho-dependent terminators [36]. Once both the 5′ and 3′ transcript ends are mapped, it is possible to annotate operons.

Annotation of operons

The transcriptome is a map of the activities of promoters and terminators. These activities are located on both strands of the genome [37] and depending on their arrangement, can give rise to antisense transcription and overlapping, divergent [38,39] and convergent operons [40,41]. To accommodate this naturally occurring complexity it is necessary to annotate the operon architecture. Three transcriptional features are necessary to define operons: 5′ ends (promoters), 3′ ends (terminators), and sufficient RNA-seq read coverage to connect the ends. If sequence reads cover 90% of the bases, this is a sensible indicator that the operon is real [4••,32••]. While there are computer algorithms that can find operons [5,42,43•], just as for TSS mapping, the state-of-the-art remains a manual process [9••]. Once the operons have been mapped, it is a straightforward task to annotate additional promoters and terminators within operons, which add complexity to the transcriptome. Mapping of internal promoters can be done manually or by bioinformatics analysis of mapped promoters that fall within the base locations of annotated operons. The transcriptional feature locations can be formatted as a GenBank feature file by using 'promoter', 'terminator' and 'operon' as feature keys (see for example, GSE52059 [4••]). This format accommodates incremental annotation of condition specific regulatory information and is an accepted standard for disseminating genome annotation data [44]. Once the transcriptional feature locations are annotated, it is reasonably straightforward to calculate the average base count value for each feature, from each dataset, as described below.

Computing the activities of transcriptional features

Analysis of RNA-seq reads at the base count level permits normalized base counts to be readily averaged across any range of base locations to calculate the relative expression level, activity, or efficiency of individual transcriptional features [4••]. We determined empirically that computing the average count of the first 10 transcribed bases accurately represents promoter activity and allows closely spaced promoters to be discriminated [4••]. Likewise, the efficiency of transcription termination can be calculated as the relative decline in average base counts in 25-base windows before and after terminators (Figure 1c). The relative transcript levels of operons can be calculated by averaging the base counts from the promoter to the terminator locations. Likewise, the expression levels of alternative transcripts generated by promoter and terminator activities within operons can be calculated. These applications of single-nucleotide-resolution analysis are exemplified in Figure 1, for wild type *E. coli* K-12 during logarithmic growth on glucose minimal medium and during starvation for carbon (stationary phase) or nitrogen, as well as an *rpoS* mutant during nitrogen starvation.

The *cysK-ptsHI-crr* operon contains four genes and multiple transcription units (Figure 1a). Conservatively, more than 40% of *E. coli* operons contain multiple transcription units that are differentially expressed, underscoring the need for an annotation system that accommodates operon architecture [4••]. In addition to the primary promoter (P-1) and terminator (T-B) that define the operon, there are eight additional promoters

(Figure 1 Lengd Continued) are visualized in J-Browse [14], as described previously [4••]. Only positive strand data are shown. Tracks: wild type (WT), glucose-grown *E. coli* K-12 in logarithmic phase (blue track); WT in stationary phase, 30 min after exhaustion of glucose (red track); WT starved for nitrogen (green track); and an isogenic *rpoS* mutant starved for nitrogen (tan track). The base count scale (on the left) is from 0 to 100, with values exceeding 100 indicated by dark red. **(b)** The relative activities of the nine promoters is plotted in the graphs as log 2 average counts of the first 10 transcribed bases under the four different growth conditions, which are colorized as above. **(c)** The decrease in average counts of the 25 bases before and after the terminator T-A are shown by light green and pink arrows. **(d)** Time series analysis of the relative expression levels of three transcripts within the complex *cysK-ptsHI-crr* operon is plotted as the log 2 average counts of bases from the indicated promoters to terminators, as described previously [4••]. Time point 1 is during middle logarithmic phase, time point 2 is immediately prior to entry into stationary phase, time point 3 is 15 min after entry into stationary phase, time point 4 is 30 min after entry into stationary phase, and time point 5 is 180 min after entry into stationary phase. Additional details of the analysis are described in the text.

Challenges

Massive amounts of RNA sequencing data can now be readily obtained. Precise mapping of transcriptional features, logical organization of the annotated data, and meaningful feature quantitation are key to maximizing the value of the resulting transcriptomes. Critical analysis of dRNA-seq data is needed to minimize the number of false positive promoters annotated. Thus it is necessary not only to properly replicate dRNA-seq experiments, but also to augment the analysis with information to corroborate that a predicted TSS is indeed a functional promoter, such as by promoter motif analysis and RNA polymerase binding assays. It would be useful if future advances in TSS mapping technology include methods to directly label the nucleotides corresponding to TSSs, rather than simply enriching for them. Mapping of 3′ transcript ends is an even larger issue and there is a real need for technology that directly labels the 3′ ends generated by transcription termination. Perhaps *in vitro* poly(A) tailing of the 3′ ends of RNA prior to fragmentation, followed by sequencing from that end would be helpful. However, it appears from existing RNA-seq data that termination is not a precise biological process and transcripts do not stop at a single nucleotide. For the time being, the state-of-the-art for 3′ transcript end mapping remains consensus between replicates.

Lastly, it is important to determine whether 'pervasive transcription', defined as TSSs in non-canonical locations [53•], is real and if such transcripts have a functional role. Pervasive transcription is seen in yeast, mammals, and fruit flies [54,55] and is frequently observed in viruses and bacteria [32••,56,57]. So, there seems to be little doubt that pervasive transcription is real. As to whether pervasive transcripts are functional, that topic was recently reviewed, but it is too early to be sure [53•]. The finding that some pervasive transcripts in herpesvirus decreased viral protein production [56] suggests that the functional role of such transcripts should be investigated in bacteria. It is becoming apparent that H-NS and NusG suppress some pervasive transcripts [57,58]. Several potential examples of pervasive transcription can be seen in Figure 1. Using a conservative approach we previously mapped 4 promoters to the *cysK-ptsHI-crr* operon [4••]. However, dRNA-seq revealed nine promoters that map to the operon (Figure 1a), only four of which appear to drive transcription of the corresponding genes (P-2, P-4, P-5, and P-9). The other five include a weak promoter upstream of the major promoter in front of *cysK* and a relatively strong promoter located within the *cysK* coding region and just upstream of the terminator that is intergenic to *cysK-ptsH*. Neither of these promoters appears to contribute to transcript expression levels. The remaining three putative pervasive promoters are located within the *ptsI* gene, have relatively low activity levels, and yet all have reasonably well conserved −10 promoter sequence elements, including two that have RpoS promoter motifs and appear to be RpoS-dependent. If these turn out to be real promoters, and there is no reason to think they are not, then the number of promoters on bacterial genomes is being underestimated by perhaps two-fold [9••,32••].

Acknowledgement

Research in the authors' laboratory was funded by the NIH (GM095370).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq**. *Curr Opin Microbiol* 2010, **13**:619-624.
2. •• Sharma CM, Vogel J: **Differential RNA-seq: the approach behind and the biological insight gained**. *Curr Opin Microbiol* 2014, **19**:97-105.
3. • Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO: **The transcription unit architecture of the *Escherichia coli* genome**. *Nat Biotechnol* 2009, **27**:1043-1049.
4. •• Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A *et al.*: **Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing**. *MBio* 2014, **5** e01442–01414.
5. Li S, Dong X, Su Z: **Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling**. *BMC Genomics* 2013, **14**:520.
6. Neidhardt FC, Bloch PL, Smith DF: **Culture medium for enterobacteria**. *J Bacteriol* 1974, **119**:736-747.
7. • Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G: **Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in *Escherichia coli***. *J Bacteriol* 2014.
8. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS *et al.*: **IVT-seq reveals extreme bias in RNA sequencing**. *Genome Biol* 2014, **15**:R86.
9. •• Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K *et al.*: **An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium**. *Cell Host Microbe* 2013, **14**:683-695.
10. •• Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R *et al.*: **The primary transcriptome of the major human pathogen *Helicobacter pylori***. *Nature* 2010, **464**:250-255.
11. Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J: **How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?** *BMC Genomics* 2012, **13**:734.
12. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357-359.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.
14. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome Res* 2009, **19**:1630-1638.
15. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy T, Taylor J, Nekrutenko A: **Dissemination of scientific software with Galaxy ToolShed**. *Genome Biol* 2014, **15**:403.

16. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets**. *Bioinformatics* 2010, **26**:2204-2207.

17. Goecks J, Nekrutenko A, Taylor J, Galaxy T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**:R86.

18. • Forstner KU, Vogel J, Sharma CM: **READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data**. *Bioinformatics* 2014.

19. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F, French StatOmique C: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Brief Bioinform* 2013, **14**:671-683.

20. • Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM: **High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates**. *PLoS Genet* 2013, **9**:e1003495.

21. Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN: **Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis***. *BMC Genomics* 2014, **15**:684.

22. Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BO: **Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling**. *PLoS Genet* 2012, **8**:e1002867.

23. • Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K *et al.*: **The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium**. *Proc Natl Acad Sci U S A* 2012, **109**:E1277-E1286.

24. • Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP: **Conservation of transcription start sites within genes across a bacterial genus**. *MBio* 2014, **5** e01398–01314.

25. Behrens S, Widder S, Mannala GK, Qing X, Madhugiri R, Kefer N, Mraheil MA, Rattei T, Hain T: **Ultra deep sequencing of *Listeria monocytogenes* sRNA transcriptome revealed new antisense RNAs**. *PLoS ONE* 2014, **9**:e83979.

26. • Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH: **Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis***. *PLoS ONE* 2012, **7**:e43350.

27. • Soutourina OA, Monot M, Boudry P, Saujet L, Pichon C, Sismeiro O, Semenova E, Severinov K, Le Bouguenec C, Coppee JY *et al.*: **Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile***. *PLoS Genet* 2013, **9**:e1003493.

28. • Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H: **RNA-Seq of *Bacillus licheniformis*: active regulatory RNA features expressed within a productive fermentation**. *BMC Genomics* 2013, **14**:667.

29. Deana A, Celesnik H, Belasco JG: **The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal**. *Nature* 2008, **451**:355-358.

30. Bischler T, Kopf M, Voss B: **Transcript mapping based on dRNA-seq data**. *BMC Bioinformatics* 2014, **15**:122.

31. Jorjani H, Zavolan M: **TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data**. *Bioinformatics* 2014, **30**:971-974.

32. •• Lin YFA, Guan DR, Mamanova S, McDowall LKJ: **A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease**. *BMC Genomics* 2013, **14**:620.

33. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics* 2011, **27**:1017-1018.

34. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA: **Characterization of 582 natural and synthetic terminators and quantification of their design constraints**. *Nat Methods* 2013, **10**:659-664.

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

36. Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R: **Rho directs widespread termination of intragenic and stable RNA transcription**. *Proc Natl Acad Sci U S A* 2009, **106**:15406-15411.

37. Taylor K, Hradecna Z, Szybalski W: **Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA**. *Proc Natl Acad Sci U S A* 1967, **57**: 1618-1625.

38. Piette J, Cunin R, Boyen A, Charlier D, Crabeel M, Van Vliet F, Glansdorff N, Squires C, Squires CL: **The regulatory region of the divergent *argECBH* operon in *Escherichia coli* K-12**. *Nucleic Acids Res* 1982, **10**:8031-8048.

39. Wek RC, Hatfield GW: **Nucleotide sequence and in vivo expression of the *ilvY* and *ilvC* genes in *Escherichia coli* K12. Transcription from divergent overlapping promoters**. *J Biol Chem* 1986, **261**:2441-2450.

40. Nomura T, Aiba H, Ishihama A: **Transcriptional organization of the convergent overlapping *dnaQ-rnh* genes of *Escherichia coli***. *J Biol Chem* 1985, **260**:7122-7125.

41. Sameshima JH, Wek RC, Hatfield GW: **Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli***. *J Biol Chem* 1989, **264**:1224-1231.

42. Fortino V, Smolander OP, Auvinen P, Tagliaferri R, Greco D: **Transcriptome dynamics-based operon prediction in prokaryotes**. *BMC Bioinformatics* 2014, **15**:145.

43. • McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B: **Computational analysis of bacterial RNA-Seq data**. *Nucleic Acids Res* 2013, **41**:e140.

44. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2014, **42**:D32-D37.

45. De Reuse H, Danchin A: **The *ptsH*, *ptsI*, and *crr* genes of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system: a complex operon with several modes of transcription**. *J Bacteriol* 1988, **170**:3827-3837.

46. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A *et al.*: **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic Acids Res* 2013, **41**:D203-D213.

47. Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R: **Genome-wide analysis of the general stress response network in *Escherichia coli:* sigmaS-dependent genes, promoters, and sigma factor selectivity**. *J Bacteriol* 2005, **187**:1591-1603.

48. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**:R106.

49. Balasubramanian D, Kumari H, Jaric M, Fernandez M, Turner KH, Dove SL, Narasimhan G, Lory S, Mathee K: **Deep sequencing analyses expands the *Pseudomonas aeruginosa* AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response**. *Nucleic Acids Res* 2014, **42**:979-998.

50. Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT: **Differential expression analysis of RNA-seq data at single-base resolution**. *Biostatistics* 2014, **15**:413-426.

51. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nat Biotechnol* 2013, **31**:46-53.

52. Wagner GP, Kin K, Lynch VJ: **A model based criterion for gene expression calls using RNA-seq data**. *Theory Biosci* 2013, **132**:159-164.

53. • Wade JT, Grainger DC: **Pervasive transcription: illuminating the dark matter of bacterial transcriptomes**. *Nat Rev Microbiol* 2014, **12**:647-653.

54. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, Wan KH, Yu C, Zhang D, Carlson JW, Cherbas L, Eads BD, Miller D, Mockaitis K, Roberts J, Davis CA, Frise E, Hammonds AS, Olson S, Shenker S, Sturgill D, Samsonova AA, Weiszmann R, Robinson G, Hernandez J, Andrews J, Bickel PJ, Carninci P, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Lai EC, Oliver B, Perrimon N, Graveley BR, Celniker SE: **Diversity and dynamics of the *Drosophila* transcriptome**. *Nature* 2014, **512**:393-399.

55. Jensen TH, Jacquier A, Libri D: **Dealing with pervasive transcription**. *Mol Cell* 2013, **52**:473-484.

56. Canny SP, Reese TA, Johnson LS, Zhang X, Kambal A, Duan E, Liu CY, Virgin HW: **Pervasive transcription of a herpesvirus genome generates functionally important RNAs**. *MBio* 2014, **5** e01033–01013.

57. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC: **Widespread suppression of intragenic transcription initiation by H-NS**. *Genes Dev* 2014, **28**:214-219.

58. Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R: **Rho and NusG suppress pervasive antisense transcription in *Escherichia coli***. *Genes Dev* 2012, **26**:2621-2633.

Appendix B: Sequencing Provider's Protocol and Notebook

vertis Biotechnologie AG 10.04.2014

Preparation of 21 cDNA libraries for Illumina sequencing

1 Material supplied

Twelve RNA samples from *E. coli* delivered on dry ice, as indicated in Table 1.

Table 1: Samples delivered

No.	Sample	Date of delivery	Conc. (ng/µl)	Vol (µl)	Conc. (ng/µl)	Total amount (µg)	ratio S23/S16
			customer-specified		own measurement (see Fig.1)		
1	Pool Coverage	10.01.14	1.218,0	107,2	3.528,0	141,1	1,1
2	Pool TEX	10.01.14	1.218,0	107,2	4.074,0	162,9	1,1
3	CFT Mouse	27.01.14	1.434,0	45,0	2.837,6	147,6	NA
4	CFT 0,2% Glucose (Log)	10.01.14	859,2	30,1	2.348,0	93,9	0,9
5	WT 0,2% glucose (log)	10.01.14	1.640,3	57,4	3.685,2	147,4	1,3
6	WT 0,2% glucose (station.)	10.01.14	1.116,9	39,1	3.438,0	137,5	1,2
7	WT phosphate stationary	10.01.14	2.412,8	84,4	7.346,0	293,8	1,0
8	phoB phosphate stationary	10.01.14	1.367,4	47,9	6.083,0	243,3	1,0
9	rpoS phosphate stationary	10.01.14	1.190,5	41,7	2.030,0	81,2	0,9
10	WT nitrogen stationary	10.01.14	883,1	30,9	2.441,0	97,6	0,9
11	rpoS nitrogen stationary	10.01.14	1.209,9	42,3	3.899,0	156,0	1,3
12	glnG nitrogen stationary	10.01.14	772,0	27,0	2.010,0	80,4	0,9

The RNA samples were analyzed by capillary electrophoresis (Figure 1).

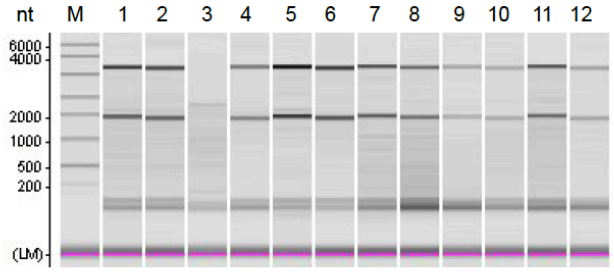


Figure 1: Analysis of the total RNA samples on a Shimadzu MultiNA microchip electrophoresis system. M = RNA marker

2 cDNA synthesis

2.1 Standard procedure

The samples were fragmented with ultrasound (4 pulses of 30 sec at 4°C) followed by a treatment with antarctic phosphatase and re-phosphorylated with polynucleotide kinase (PNK). Afterwards, the RNA fragments were poly(A)-tailed using poly(A) polymerase and a RNA adapter was ligated to the 5'-phosphate of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and M-MLV reverse transcriptase. The resulting cDNA was PCR-amplified to about 20-30 ng/µl using a high fidelity DNA polymerase (cycle numbers and barcode sequences which are part of the 3' TruSeq sequencing adaptor are indicated in Table 2). The cDNA was purified using the Agencourt AMPure XP kit and analyzed by capillary electrophoresis.

2.2 TEX treatments of RNA samples

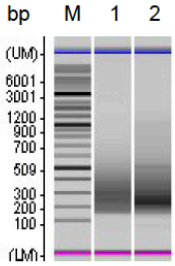
RNA samples Nr. 2, 4 - 12 were first fragmented with ultrasound (4 pulses of 30 sec at 4°C) and then treated with polynucleotide kinase (PNK). This was followed by treatment with Terminator exonuclease (TEX) and then the RNA samples were poly(A)-tailed using poly(A) polymerase. The 5'PPP structures were removed using a 5' Polyphosphatase. Afterwards, a RNA adapter was ligated to the 5'-phosphate of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. The resulting cDNAs were PCR-amplified to about 20-30 ng/µl using a high fidelity DNA polymerase (cycle numbers and barcode sequences which are part of the 3' TruSeq sequencing adaptor are indicated in Table 2). The cDNA was purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics) and analyzed by capillary electrophoresis.

The following tables indicate the Illumina barcodes of the cDNA samples representing the cDNA pools and numbers of PCR cycles, used for cDNA amplification. The cDNA samples were analyzed on a Shimadzu MultiNA microchip electrophoresis system (M = 100 bp ladder).

Table 2: Illumina barcodes for samples representing pools 1 to 3

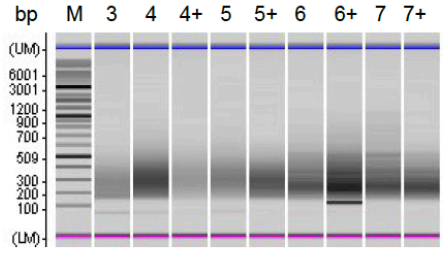
Pool 1

No.	Sample	Barcode	PCR cycles
1	Pool Coverage	CAAAAG	13
2	Pool TEX	ATCACG	21



Pool 2

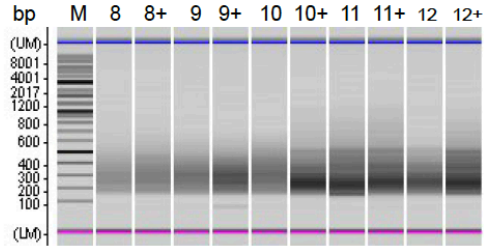
No.	Sample	Barcode	PCR cycles
3	CFT Mouse	AGTTCC	13
4	CFT 0,2% Glucose (Log)	CCGTCC	13
4	CFT 0,2% Glucose (Log)+TEX	TTAGGC	20
5	WT 0,2% glucose (log)	GTTTCG	12
5	WT 0,2% glucose (log)+TEX	TGACCA	21
6	WT 0,2% glucose (station.)	ATGAGC	12
6	WT 0,2% glucose (station.)+TEX	ACAGTG	20
7	WT phosphate stationary	CAACTA	13
7	WT phosphate stationary +TEX	GCCAAT	20



Pool 3

No.	Sample	Barcode	PCR cycles
8	phoB phosphate stationary	GTGAAA	12
8	phoB phosphate stationary+TEX	CAGATC	21
9	rpoS phosphate stationary	GGTAGC	13
9	rpoS phosphate stationary+TEX	ACTTGA	22
10	WT nitrogen stationary	CACCGG	13
10	WT nitrogen stationary+TEX	GATCAG	21

11	rpoS nitrogen stationary	ATGTCA	13
11	rpoS nitrogen stationary+TEX	TAGCTT	20
12	glnG nitrogen stationary	GAGTGG	13
12	glnG nitrogen stationary+TEX	GGCTAC	21



3 Pool generation and size fractionation

For Illumina sequencing, the cDNA samples as indicated in Table 2 were pooled in equimolar amounts. The cDNA pools were fractionated in the size range of 150–550 bp using a differential clean-up with the Agencourt AMPure kit. An aliquot of each size fractionated cDNA pool was analyzed by capillary electrophoresis (Fig. 2).

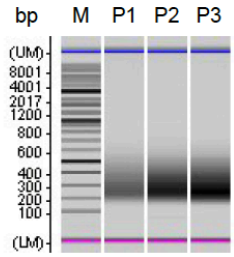


Figure 2: Analysis of the size fractionated cDNA pools 1 - 3 a Shimadzu MultiNA microchip electrophoresis system. M = 100 bp ladder.

4 Sample description

The primers used for PCR amplification were designed for TruSeq sequencing according to the instructions of Illumina.

The following adapter sequences flank the cDNA inserts:

TrueSeq_Sense_primer
5´- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN-3'

TrueSeq_Antisense_ NNNNNN_primer Barcode
5'-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3'

The combined length of the flanking sequences is 148 bases.

The following tables indicate the Illumina barcodes for the cDNA samples representing cDNA pools 1 – 3. Included in the tables are the numbers of PCR cycles used for cDNA amplification. The cDNA samples were analysed on a Shimadzu MultiNA microchip electrophoresis system (M = 100 bp ladder).

5 Ilumina sequencing

The cDNA pools were sequenced on a Illumina HiSeq2000 system using 50 bp read length. The files provided are summarized in Table 3.

Pool	Sample	File archive	Reads
1	Pool Coverage	NG-7271_VB1395_1_CAAAAG_lib39507_2212_7_1.fastq.bz2	70.100.593
	Pool TEX	NG-7271_VB1395_1_ATCACG_lib39512_2212_7_1.fastq.bz2	76.237.119
2	CFT Mouse	NG-7271_VB1395_2_AGTTCC_lib39508_2140_6_1.fastq.bz2	20.335.701
	CFT 0,2% GI (L)	NG-7271_VB1395_2_CCGTCC_lib39513_2140_6_1.fastq.bz2	22.285.927
	CFT 0,2% GI (L) +T	NG-7271_VB1395_2_TTAGGC_lib39517_2140_6_1.fastq.bz2	24.716.248
	WT 0,2% GI (l)	NG-7271_VB1395_2_GTTTCG_lib39514_2140_6_1.fastq.bz2	10.908.259
	WT 0,2% GI (l) +T	NG-7271_VB1395_2_TGACCA_lib39518_2140_6_1.fastq.bz2	24.861.247
	WT 0,2% GI (st)	NG-7271_VB1395_2_ATGAGC_lib39515_2140_6_1.fastq.bz2	20.741.761
	WT 0,2% GI (st) +T	NG-7271_VB1395_2_ACAGTG_lib39519_2140_6_1.fastq.bz2	25.073.211
	WT phos st	NG-7271_VB1395_2_CAACTA_lib39516_2140_6_1.fastq.bz2	23.854.994
	WT phos st +TEX	NG-7271_VB1395_2_GCCAAT_lib39520_2140_6_1.fastq.bz2	24.273.577
3	phoB phos st	NG-7271_VB1395_3_GTGAAA_lib39509_2140_7_1.fastq.bz2	16.300.148
	phoB phos st +T	NG-7271_VB1395_3_CAGATC_lib39525_2140_7_1.fastq.bz2	27.344.682
	rpoS phos st	NG-7271_VB1395_3_GGTAGC_lib39521_2140_7_1.fastq.bz2	19.020.367
	rpoS phos st +T	NG-7271_VB1395_3_ACTTGA_lib39526_2140_7_1.fastq.bz2	24.716.248
	WT nit st	NG-7271_VB1395_3_CACCGG_lib39522_2140_7_1.fastq.bz2	17.086.304
	WT nit st +TEX	NG-7271_VB1395_3_GATCAG_lib39527_2140_7_1.fastq.bz2	22.926.696
	rpoS nit st	NG-7271_VB1395_3_ATGTCA_lib39523_2140_7_1.fastq.bz2	21.172.604
	rpoS nit st +TEX	NG-7271_VB1395_3_TAGCTT_lib39528_2140_7_1.fastq.bz2	22.212.822
	glnG nit st	NG-7271_VB1395_3_GAGTGG_lib39524_2140_7_1.fastq.bz2	18.525.661
	glnG nit st +TEX	NG-7271_VB1395_3_GGCTAC_lib39529_2140_7_1.fastq.bz2	23.316.287

Illumina fastq format

Four lines for each sequence:

1- Unique identifier, with the following format:

@<machine_id>:<lane>:<tile>:<x_coord>:<y_coord>#<index>/<read_#>

Starting with **Casava 1.8** the header information changed to:

@<machine_id>:<run id>:<flowcell id>:<lane>:<tile>:<x_coord>:<y_coord>;<member of a pair: "1" or "2">:<read filtered (Y)es or (N)o>:<control bits on "0" or otherwise it is an even number>: <index>

2- Sequence (A, T, C, G or N (undetermined) only)
3- Orientation (always forward without mapping)
4- Quality value for each base, corresponding to a Phred-like score encoded in ASCII format. With **CASAVA 1.7** and previous pipeline versions, there is an offset of 64 in relation with the ASCII table (e.g. "f" gives a value of 38). With **CASAVA 1.8**, the offset is of 33 (e.g. "J" gives a value of 41) and is in accordance with Sanger FASTQ format.

Example (< Casava 1.8):

@HWI-ST132_403:1:1:17205:2161#AGGACC/1
CAATTAACACTCTAAGAGCTTCTTCTCCCTCGCTCTGGGTCTTTTTCACA
+
fffffefffedffffeccedd affceeeeceeeeceecbeeededdе`c

Example (>= Casava 1.8):

@HWI-ST865:45:D0BWNFBYX:7:1101:17932:2151 2:N:0:ACAGCG
AAGAATGACACTGGACTTCTTTTCAGAAATCACATAAATAATAAGATGAGA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJIJJJIJJJJJJJJJJIJJJJ

The sequence files are compressed as .bz2 archives.

Appendix C: Chapter 4 Supplemental Tables

Supplemental Table S4-1: RpoS-Dependent Transcription Units and Genes

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-53:P-132	-		P-132
P-17311:T-20084	+	nhaA	P-17311
T-33725:AS-34027	-		AS-34027
P-49788:T-50527	+	folA	P-49788
T-50160:I-52034	-	apaG:apaH	I-52034
T-138648:P-141263	-	gcd	
I-145610:T-146571	+		I-145610
T-180636:AS-181048	-		
T-194917:AS-194965	-		AS-194965
P-209658:T-213728	+	ldcC:yaeR:tilS	
I-214621:T-216124	+	arfB:nlpE	I-214621
T-318634:P-320033	-	rclA	P-320033
T-331514:P-331672	-		P-331672
T-339496:O-340091	-		O-340091
P-342818:T-343975	+	yahK:yahL	P-342818
S-344993:T-345117	+		S-344993
P-346432:T-346972	+	yahO	P-346432
T-346618:P-348469	-	prpR	P-348469
T-380062:I-380445	-		I-380445
AS-382902:T-383388	+		
T-395032:P-396363	-	ampH	
S-406954:T-408492	+	yaiA	S-406954
T-436731:S-438284	-	xseB:ispA:dxs:yajO	S-438284
S-454434:T-454867	+	bolA	
S-457336:T-458755	+	clpX	
P-475306:T-475998	+	ybaY	P-475306
T-501139:P-503295	-		
P-511574:T-514655	+	glsA:ybaT:cueR	P-511574
I-533632:T-533865	+		I-533632
I-533632:T-540536	+	gcl:hyi:glxR:ybbW:allB	I-533632
P-576858:T-577278	+		P-576858
P-583582:T-584677	+	appY	
T-637782:S-638633	-	dsbG	
S-674993:T-676044	+	ybeL	S-674993
T-720993:I-722210	-	kdpF:kdpA:kdpB:kdpC:kdpD:kdpE	
P-738980:T-741000	+	ybgA:phr	P-738980
T-784893:P-785462	-		P-785462
S-798555:T-799595	+	pgl	S-798555

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-803326:P-803436	-		
T-807410:S-807929	-	ybhB	S-807929
P-819858:T-820636	+	ybhL	
T-821433:P-824524	-		P-824524
T-837449:P-838482	-	ybiI:ybiJ	P-838482
T-841763:P-842088	-	mcbA	P-842088
T-848341:P-848950	-	dps	P-848950
P-850128:T-851006	+	ompX	P-850128
P-878722:T-880048	+	yliI	P-878722
P-880618:T-881952	+	dacC	P-880618
T-900804:P-903843	-	artP:artI:artQ:artM	P-903843
T-903910:P-904522	-	ybjP	P-904522
P-904517:T-906055	+	ybjQ:amiD	P-904517
T-905670:P-911076	-	ltaE:ybjT:ybjS	P-911076
AS-914828:T-915089	+		AS-914828
T-915220:P-916231	-	aqpZ	P-916231
T-944756:P-945616	-	ycaC	
P-945841:T-947040	+	ycaD	
P-956736:T-957530	+	ycaP	P-956736
P-980888:T-982964	+	ldtD	
P-987402:T-987632	+		
T-1000643:AS-1000992	-		
P-1027748:T-1028826	+	yccU	
P-1029962:T-1030612	+	yccX	
S-1037707:T-1041945	+	cbdA:cbdB:cbdX:appA	S-1037707
T-1062508:S-1063828	-	cbpA:cbpM	
P-1065559:T-1066855	+	agp	
T-1066836:P-1067983	-	wrbA:yccJ	P-1067983
P-1068001:T-1068507	+	ymdF	P-1068001
T-1113538:P-1114212	-	msyB	P-1114212
T-1148961:AS-1149109	-		AS-1149109
I-1157250:T-1159356	+	ptsG	I-1157250
P-1185738:T-1187102	+	pepT	
P-1215752:T-1217110	+	ycgZ:ymgA:ariR:ymgC	
T-1218433:AS-1218876	-		AS-1218876
T-1235674:P-1237285	-		P-1237285
P-1244678:T-1245637	+	ymgE:ycgY	P-1244678
T-1245665:P-1247422	-	treA	P-1247422
P-1258738:T-1259100	+	ychH	
P-1289106:T-1292389	+	rssA:rssB:galU	
T-1301000:AS-1301196	-		AS-1301196

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-1314685:P-1316092	-	yciG:yciF:yciE	P-1316092
I-1323765:T-1324865	+	yciO	
P-1335781:T-1338985	+	acnA	
T-1343015:S-1343368	-	osmB	S-1343368
T-1343576:P-1346842	-		P-1346842
P-1361052:T-1362588	+	puuD:puuR	
I-1362751:T-1367027	+	puuB:puuE	
T-1406494:P-1407869	-	ydaM	P-1407869
P-1409283:T-1410954	+	dbpA	P-1409283
T-1426762:AS-1426973	-		AS-1426973
AS-1433018:T-1433419	+	ttcA	AS-1433018
T-1437213:P-1440845	-	pfo	
T-1441282:S-1441777	-	ldhA:hslJ	S-1441777
I-1448821:T-1449043	+		
S-1495265:T-1496689	+	ydcJ:opgD:ydcH:rimL	
T-1499401:P-1500471	-	ydcK	P-1500471
P-1502429:T-1503408	+	ydcL	P-1502429
P-1511497:T-1517055	+	ydcS:ydcT:ydcU:ydcV:patD	
P-1518999:T-1521073	+	curA:mcbR	P-1518999
P-1526225:T-1526888	+	yncG	P-1526225
T-1532934:P-1533874	-		P-1533874
T-1535903:P-1544162	-	narU:narZ:narY:narW:narV	P-1544162
T-1551928:S-1553898	-		S-1553898
T-1555774:S-1556047	-	sra	S-1556047
T-1555774:P-1556339	-	sra	P-1556339
S-1556600:T-1557362	+	osmC	S-1556600
T-1563304:P-1567294	-	dosC:dosP	P-1567294
T-1568910:P-1572073	-	gadC	P-1572073
S-1607323:T-1608642	+	tam	S-1607323
T-1618172:P-1618940	-		
P-1624557:T-1625450	+	ydeJ	P-1624557
AS-1631394:T-1631789	+		
P-1646146:T-1646778	+	flxA	
P-1657539:T-1657890	+	ynfD	P-1657539
S-1671916:T-1672955	+	ydgD	
I-1680761:T-1681778	+	folM	
P-1689794:T-1691563	+	ydgA	
T-1696447:S-1698122	-	uidR	S-1698122
T-1723981:S-1724679	-	sodC	
T-1723981:S-1725664	-	ydhL:ydhF:sodC	S-1725664
T-1723981:P-1725945	-	ydhF:sodC	

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
P-1747091:T-1748971	+	ydhS	P-1747091
S-1755600:T-1757160	+	pykF	S-1755600
T-1757693:S-1758864	-	sufA:sufB:sufC:sufD:sufS:sufE:ldtE	S-1758864
T-1757693:P-1764500	-	ldtE	P-1764500
P-1768977:T-1770211	+	ydiK	
T-1770417:AS-1770569	-		
T-1792243:S-1794143	-	btuE:btuD:nlpC	S-1794143
T-1795224:I-1795720	-	pheM:pheS:pheT:ihfA	
S-1807375:T-1808700	+	ydiZ:yniA	S-1807375
P-1813813:T-1816333	+	katE	P-1813813
T-1821867:P-1822283	-	osmE	P-1822283
P-1866747:T-1870266	+	yeaG:yeaH	P-1866747
T-1878976:P-1879297	-	yeaQ	P-1879297
S-1894017:T-1894696	+	yoaC	S-1894017
P-1898383:T-1900109	+	yoaD	P-1898383
P-1916143:T-1920182	+	yebS:yebT	
P-1921727:T-1922662	+	yebV	P-1921727
T-1921892:P-1923206	-		P-1923206
T-1929793:S-1930416	-	yebF	S-1930416
T-1979975:P-1982442	-	otsA	P-1982442
T-1989247:S-1989501	-	yecH	
T-1996071:S-2000410	-	dcyD:yecS:yecC:sdiA	
P-2006132:T-2007648	+	amyA	P-2006132
P-2009802:T-2011180	+	yedK:yedL	P-2009802
T-2024353:P-2024841	-	dsrB	P-2024841
P-2024927:T-2026468	+	yodD:yedP	P-2024927
T-2026260:S-2028396	-	yedQ	S-2028396
P-2035631:T-2036790	+	hchA	P-2035631
I-2039749:T-2041246	+	yedZ	I-2039749
T-2062358:S-2063353	-	ldtA	S-2063353
P-2165132:T-2167163	+	yegP:yegQ:cyaR	P-2165132
P-2168685:T-2169666	+	yegS	P-2168685
T-2177331:P-2178630	-	fbaB	P-2178630
T-2192476:P-2192843	-	yehE	
P-2214840:T-2215949	+	mlrA:yohO	P-2214840
T-2218501:P-2219523	-	osmF	P-2219523
T-2225015:P-2225664	-		P-2225664
P-2225777:T-2226627	+	yohD	
T-2226436:P-2227296	-	yohF	P-2227296
P-2228958:T-2229199	+	yohP	P-2228958
P-2313084:T-2313180	+	micF	

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-2380687:S-2381053	-	elaA:elaB	S-2381053
T-2460624:P-2461145	-	yfcZ	
P-2461205:T-2462772	+	fadL	P-2461205
T-2464223:S-2465135	-	mlaA	
S-2465274:T-2466391	+	yfdC:argW	S-2465274
P-2470748:T-2471140	+	tfaS	P-2470748
T-2487526:I-2488293	-		I-2488293
P-2509455:T-2510942	+	yfeO	P-2509455
T-2511075:S-2512853	-	mntH	
I-2525595:T-2525975	+		I-2525595
S-2533730:T-2536391	+	ptsH:ptsI:crr	S-2533730
T-2560938:P-2561057	-		P-2561057
P-2578590:T-2581783	+	talA:tktB	P-2578590
T-2593016:S-2596763	-	tmcA:ypfH	
T-2664305:P-2664356	-		
P-2665411:T-2666881	+	csiE	P-2665411
S-2673321:T-2673802	+	yphA	S-2673321
T-2697213:P-2698592	-	tadA	
AS-2707107:T-2707348	+		
P-2731193:T-2731551	+		
T-2771214:I-2772008	-		I-2772008
T-2778136:I-2781211	-		I-2781211
T-2778136:I-2781601	-		I-2781601
AS-2780661:T-2780823	+		
P-2788927:T-2796496	+	csiD:lhgO:gabD:gabT	P-2788927
T-2796039:S-2796813	-	yqaE:ygaU	S-2796813
T-2796039:P-2797085	-	ygaU	
T-2799516:P-2800092	-	ygaC	P-2800092
P-2800121:T-2800543	+	ygaM	P-2800121
P-2818759:T-2819074	+		P-2818759
T-2818822:S-2819273	-	csrA	S-2819273
T-2824480:P-2825606	-	mltB	
P-2903986:T-2904093	+		P-2903986
T-2904707:P-2905442	-	queE	
T-2906608:I-2910110	-	relA:mazE:mazF:mazG:pyrG:eno	I-2910110
T-2929527:AS-2929575	-		
T-2972622:I-2974560	-	lplT	I-2974560
T-2976084:P-2976189	-	omrA	
P-3015058:T-3016158	+	mocA:ygfK:ssnA:ygfM:xdhD	
P-3033041:T-3033901	+	idi	
P-3051012:T-3051150	+		

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-3067290:S-3068126	-	argO:yggE	
T-3082583:I-3085928	-	speB	
T-3083837:I-3083951	-		
P-3086567:T-3087918	+	metK	P-3086567
AS-3100654:T-3101060	+		AS-3100654
AS-3121327:T-3121617	+		
T-3146781:P-3147766	-	yghX	P-3147766
P-3149636:T-3150811	+	yghA	P-3149636
S-3156572:T-3157662	+	dkgA	S-3156572
T-3158140:AS-3158189	-		AS-3158189
T-3169250:P-3169709	-	ygiW	P-3169709
P-3177866:T-3179627	+	tolC	
T-3191642:S-3192002	-	glgS	S-3192002
T-3214837:P-3215496	-	mug	
P-3219459:T-3220906	+	patA	P-3219459
P-3248888:T-3250954	+	yqjC:yqjD:yqjE:yqjK	P-3248888
P-3250976:T-3252032	+	yqjG	P-3250976
T-3252091:P-3252139	-		P-3252139
P-3252263:T-3252703	+	yhaH	
P-3298908:T-3299620	+	yhbO	P-3298908
P-3303412:T-3304930	+	yhbW	P-3303412
AS-3313034:T-3313213	+		AS-3313034
T-3344982:AS-3345199	-		AS-3345199
T-3367412:P-3367769	-		P-3367769
T-3368968:I-3370087	-	nanR:nanA:nanT:nanE:nanK:yhcH	I-3370087
S-3380653:T-3382143	+	degQ:degS	S-3380653
T-3385810:P-3386169	-	yhcO	P-3386169
T-3390490:S-3392421	-	yhdE:rng:yhdP:tldD	
AS-3397835:T-3397992	+		
AS-3398471:T-3398627	+		AS-3398471
S-3418971:T-3420093	+	yhdW:yhdW:yhdW	
I-3435065:T-3437936	+	rsmB:trkA	
S-3438001:T-3438485	+	mscL	
AS-3441061:T-3441240	+		AS-3441061
T-3465941:S-3466747	-		S-3466747
P-3478501:T-3478619	+		P-3478501
T-3490266:P-3491648	-	yhfG:fic:pabA	P-3491648
I-3528874:T-3530670	+	hslR:hslO	I-3528874
T-3580919:P-3581018	-	ryhB	
T-3584411:P-3584459	-		
I-3584765:T-3585151	+		

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
T-3585071:P-3586854	-	ggt	P-3586854
P-3586920:T-3587598	+	yhhA	
T-3590998:P-3592426	-	ugpB	
P-3611841:T-3613590	+	yhhT:acpT	
T-3618885:P-3619164	-		P-3619164
P-3634742:T-3635932	+	yhiM	P-3634742
P-3637396:T-3639355	+	pitA	P-3637396
T-3638889:P-3639846	-		P-3639846
P-3640838:T-3642357	+	dtpB	P-3640838
P-3653936:T-3654571	+	slp:dctR	P-3653936
T-3655062:P-3656791	-	yhiD	P-3656791
T-3655918:P-3656791	-	hdeB	P-3656791
P-3656960:T-3657612	+	hdeD	P-3656960
P-3657800:T-3663735	+	gadE:mdtE:mdtF	
T-3657980:AS-3658054	-	arrS	AS-3658054
T-3663530:S-3665841	-	gadW	S-3665841
T-3663530:P-3667608	-		P-3667608
P-3669546:T-3671568	+	treF	P-3669546
S-3673332:T-3674463	+	yhjD	S-3673332
T-3676056:S-3678380	-		S-3678380
T-3683612:I-3686203	-	yhjK	I-3686203
T-3695858:S-3696393	-	yhjR	S-3696393
P-3707887:T-3708310	+		
T-3710620:P-3713010	-		P-3713010
AS-3712290:T-3712399	+		AS-3712290
P-3719431:T-3719814	+	yiaG	P-3719431
T-3727878:I-3729964	-	xylB	
AS-3736592:T-3736723	+		
T-3754552:S-3756535	-		S-3756535
T-3765825:AS-3766091	-		AS-3766091
T-3767581:AS-3767679	-		AS-3767679
T-3770070:S-3771811	-	yibH	S-3771811
AS-3801038:T-3801278	+		AS-3801038
P-4012886:T-4015663	+	metE	P-4012886
T-4015323:S-4016202	-	ysgA	S-4016202
T-4020871:AS-4020921	-		AS-4020921
S-4102800:T-4103543	+	yiiM	
I-4107371:T-4108602	+	pfkA	I-4107371
T-4110729:S-4111569	-		
T-4133195:P-4133767	-		
T-4161750:I-4162921	-		

RpoS-Dependent TUs (315)	Strand	RpoS-Dependent Genes (359)	RpoS-Dependent Promoters
P-4214233:T-4215260	+	metA	P-4214233
P-4233667:T-4235588	+	pgi	
P-4259179:T-4259609	+	yjbJ	P-4259179
T-4263218:P-4264304	-	qorA	
T-4277930:P-4278099	-		P-4278099
T-4325219:P-4325766	-		P-4325766
T-4332036:I-4334291	-	basR:basS	
S-4332095:T-4332397	+	pmrR	
S-4351813:T-4352373	+	yjdI:yjdJ:yjdK:yjdO	
T-4362517:S-4363330	-	yjdC:pheU	S-4363330
AS-4366893:T-4367329	+		AS-4366893
P-4375565:T-4376695	+	efp:ecnA:ecnB:sugE	P-4375565
S-4376510:T-4376695	+	ecnB:sugE	S-4376510
T-4377103:P-4377746	-		P-4377746
P-4414249:T-4415960	+	aidB	P-4414249
T-4415971:P-4416869	-	bsmA:yjfN	P-4416869
T-4424474:I-4424536	-	yjfY	
T-4424474:P-4424843	-		
T-4429589:AS-4429731	-		
S-4436564:T-4437515	+	cysQ:ytfI	S-4436564
T-4438638:P-4439310	-	ytfJ	
P-4439134:T-4439831	+	ytfK	P-4439134
P-4449915:T-4454752	+	ytfQ	P-4449915
T-4457069:P-4457969	-	yjgA	
T-4460297:P-4462800	-	nrdG	P-4462800
T-4488423:S-4490083	-	yjgR	
T-4495171:P-4496233	-	ahr	P-4496233
T-4537650:P-4538902	-		P-4538902
I-4608389:T-4611051	+	yjjG:prfC	I-4608389
P-4611152:T-4612332	+	osmY:ytjA	P-4611152
P-4616180:T-4621649	+	deoC:deoA:deoB:deoD	P-4616180

Supplemental Table S4-2: RpoS-Dependent Promoters

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
AS-122	-		Contained within Operon
P-132	-		
P-17311	+	nhaA>	
I-18025	+	nhaR>	Contained within Operon
AS-34027	-		
P-49788	+	folA>	
I-52034	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
I-145610	+	yadD>	
AS-194965	-		
I-214621	+	arfB>	
P-320033	-		
P-331672	-		
O-339813	-		Contained within Operon
O-339906	-		Contained within Operon
O-339999	-		Contained within Operon
O-340091	-		
P-342818	+	yahK>	
S-344993	+	yahM>	
P-346432	+	yahO>	
P-348469	-		
I-380445	-		
AS-383123	+	tauA>	Contained within Operon
S-406954	+	yaiA>	
S-407344	+	aroM>	Contained within Operon
I-437997	-		Contained within Operon
S-438284	-		
P-475306	+	ybaY>	
P-511574	+	glsA>	
I-533632	+	gcl>	
P-576858	+	essD>	
S-674993	+	ybeL>	
I-675045	+	ybeR>	Contained within Operon
P-738980	+	ybgA>	
I-739823	+	ybgI>	Contained within Operon
P-785462	-		
S-798555	+	pgl>	
S-807929	-		
I-822075	-		Contained within Operon
P-824524	-		
P-838482	-		
P-842088	-		
P-848950	-		
P-850128	+	ompX>	
P-878722	+	yliI>	
P-880618	+	dacC>	
S-903791	-		Contained within Operon
P-903843	-		
P-904517	+	ybjQ>	
P-904522	-		
P-911076	-		
AS-914828	+	ybjD>	
P-916231	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
P-956736	+	ycaP>	
I-982010	+	ycbK>	Contained within Operon
S-1037707	+	cbdA>	
S-1067758	-		Contained within Operon
P-1067983	-		
P-1068001	+	ymdF>	
I-1113974	-		Contained within Operon
P-1114212	-		
AS-1149109	-		
I-1157250	+	ptsG>	
I-1216959	+	ycgG>	Contained within Operon
AS-1218876	-		
P-1237285	-		
P-1244678	+	ymgE>	
P-1247422	-		
S-1290177	+	rssB>	Contained within Operon
AS-1301196	-		
P-1316092	-		
S-1343368	-		
I-1344314	-		Contained within Operon
P-1346842	-		
P-1407869	-		
P-1409283	+	dbpA>	
AS-1426973	-		
AS-1433018	+	micC>	
I-1440772	-		Contained within Operon
S-1441777	-		
P-1500471	-		
P-1502429	+	ydcL>	
P-1518999	+	curA>	
I-1519702	+	mcbR>	Contained within Operon
I-1519742	+	mcbR>	Contained within Operon
S-1520134	+	mcbR>	Contained within Operon
P-1526225	+	yncG>	
P-1533874	-		
P-1544162	-		
O-1552037	-		Contained within Operon
O-1552215	-		Contained within Operon
O-1552322	-		Contained within Operon
S-1553898	-		
S-1556047	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
P-1556339	-		
S-1556600	+	osmC>	
S-1567269	-		Contained within Operon
P-1567294	-		
I-1571396	-		Contained within Operon
P-1572073	-		
S-1607323	+	tam>	
P-1624557	+	ydeJ>	
P-1657539	+	ynfD>	
S-1698122	-		
S-1725664	-		
P-1747091	+	ydhS>	
S-1755600	+	pykF>	
S-1758864	-		
P-1764500	-		
S-1794143	-		
S-1807375	+	ydiZ>	
P-1813813	+	katE>	
I-1815997	+	nadE>	Contained within Operon
P-1822283	-		
P-1866747	+	yeaG>	
S-1866816	+	yeaG>	Contained within Operon
P-1879297	-		
S-1894017	+	yoaC>	
P-1898383	+	yoaD>	
P-1921727	+	yebV>	
I-1921985	+	yebW>	Contained within Operon
P-1923206	-		
S-1930416	-		
I-1980197	-		Contained within Operon
I-1981650	-		Contained within Operon
I-1981940	-		Contained within Operon
P-1982442	-		
P-2006132	+	amyA>	
P-2009802	+	yedK>	
P-2024841	-		
P-2024927	+	yodD>	
S-2028396	-		
P-2035631	+	hchA>	
I-2039749	+	yedZ>	
S-2063353	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
P-2165132	+	yegP>	
P-2168685	+	yegS>	
P-2178630	-		
P-2214840	+	mlrA>	
P-2219523	-		
I-2225443	-		Contained within Operon
P-2225664	-		
P-2227296	-		
P-2228958	+	yohP>	
S-2229077	+	yohJ>	Contained within Operon
S-2381053	-		
P-2461205	+	fadL>	
S-2461250	+	fadL>	Contained within Operon
S-2465274	+	yfdC>	
P-2470748	+	tfaS>	
I-2488293	-		
P-2509455	+	yfeO>	
I-2525595	+	yfeH>	
S-2533730	+	ptsH>	
I-2534840	+	crr>	Contained within Operon
P-2561057	-		
P-2578590	+	talA>	
I-2593199	-		Contained within Operon
I-2593969	-		Contained within Operon
I-2594383	-		Contained within Operon
P-2665411	+	csiE>	
S-2673321	+	yphA>	
I-2772008	-		
I-2781211	-		
I-2781601	-		
P-2788927	+	csiD>	
I-2790910	+	gabD>	Contained within Operon
S-2796813	-		
P-2800092	-		
P-2800121	+	ygaM>	
P-2818759	+	srlA>	
S-2819273	-		
P-2903986	+	yqcG>	
I-2904019	+	yqcG>	Contained within Operon
I-2908439	-		Contained within Operon
I-2910110	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
I-2974560	-		
I-3067780	-		Contained within Operon
P-3086567	+	metK>	
S-3086683	+	metK>	Contained within Operon
AS-3100654	+	mutY>	
I-3147682	-		Contained within Operon
P-3147766	-		
P-3149636	+	yghA>	
S-3156572	+	dkgA>	
AS-3158189	-		
P-3169709	-		
S-3192002	-		
P-3219459	+	patA>	
P-3248888	+	yqjC>	
I-3249267	+	yqjD>	Contained within Operon
P-3250976	+	yqjG>	
P-3252139	-		
P-3298908	+	yhbO>	
P-3303412	+	yhbW>	
AS-3313034	+	argG>	
AS-3345199	-		
AS-3367618	-		Contained within Operon
P-3367769	-		
I-3370087	-		
S-3380653	+	degQ>	
P-3386169	-		
AS-3398471	+	acuI>	
AS-3441061	+	gspC>	
S-3466747	-		
P-3478501	+	yheS>	
I-3491489	-		Contained within Operon
P-3491648	-		
I-3528874	+	hslR>	
P-3586854	-		
P-3619164	-		
P-3634742	+	yhiM>	
P-3637396	+	pitA>	
S-3637578	+	pitA>	Contained within Operon
P-3639846	-		
P-3640838	+	dtpB>	
P-3653936	+	slp>	

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
O-3654112	+	dctR>	Contained within Operon
I-3656309	-		Contained within Operon
P-3656791	-		
P-3656960	+	hdeD>	
AS-3658054	-		
S-3658242	+	gadE>	Contained within Operon
S-3658992	+	mdtE>	Contained within Operon
S-3659013	+	mdtE>	Contained within Operon
I-3659240	+	mdtF>	Contained within Operon
I-3663479	+	gadY>	Contained within Operon
I-3664740	-		Contained within Operon
I-3665158	-		Contained within Operon
S-3665841	-		
P-3667608	-		
P-3669546	+	treF>	
S-3673332	+	yhjD>	
I-3674177	+	yhjE>	Contained within Operon
S-3678380	-		
I-3686203	-		
S-3696393	-		
AS-3712290	+	tag>	
P-3713010	-		
P-3719431	+	yiaG>	
I-3756037	-		Contained within Operon
S-3756535	-		
AS-3766091	-		
AS-3767679	-		
I-3771724	-		Contained within Operon
S-3771811	-		
AS-3801038	+	waaA>	
P-4012886	+	metE>	
I-4013045	+	metE>	Contained within Operon
I-4014980	+	udp>	Contained within Operon
S-4016202	-		
AS-4020921	-		
I-4107371	+	pfkA>	
I-4111252	-		Contained within Operon
P-4214233	+	metA>	
I-4234742	+	yjbE>	Contained within Operon
P-4259179	+	yjbJ>	
P-4278099	-		

RpoS-Dependent Promoters (278)	Strand	Gene	Comments
P-4325766	-		
S-4363330	-		
AS-4366893	+	fxsA>	
P-4375565	+	efp>	
S-4376510	+	ecnB>	
P-4377746	-		
P-4414249	+	aidB>	
I-4415864	+	yjfP>	Contained within Operon
S-4416370	-		Contained within Operon
P-4416869	-		
S-4436564	+	cysQ>	
P-4439134	+	ytfK>	
P-4449915	+	ytfQ>	
I-4461547	-		Contained within Operon
P-4462800	-		
P-4496233	-		
P-4538902	-		
I-4608389	+	yjjG>	
P-4611152	+	osmY>	
P-4616180	+	deoC>	
I-4619270	+	deoB>	Contained within Operon