UNIVERSITY OF CALIFORNIA, SAN DIEGO

Systems Evaluation of Regulatory Components in Bacterial Transcription Initiation

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Donghyuk Kim

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Eric E. Allen
Professor Victor Nizet
Professor Shankar Subramaniam
Professor Kun Zhang

2014

The dissertation of Donghyuk Kim is approved, and

it is acceptable in quality for publication on microfilm:

Chair

University of California, San Diego

2014

Dedication

To Sejin Koo

For your enduring love, and support.

You truly are my best friend and a life-long companion of faith.

To Isabella

For your heart-warming smiles.

May you live in goodness and faith.

Epigraph

我非生以知之者, 好古, 敏以求之者也

孔子

It is the glory of God to conceal a matter, but the glory of kings is to search out a matter.

Proverbs 25:2

Table of Contents

List of Figures and Tables

Acknowledgements

I owe numerous people greatly for their help, which has contributed to the drafting of this dissertation. This work would not have been possible without their support, both big and small.

First of all, I appreciate many researchers who have laid the scientific groundwork tirelessly, for decades, upon which this thesis stands. In particular, I thank Professor Bernhard O. Palsson for cultivating a collaborative and innovative team of researchers in Systems Biology Research Group. I, particularly, thank him for his care in mentoring and guiding me through my doctoral program and my scientific development. I will be forever grateful for the training I have received under his guidance, with respect to framing meaningful research questions, tailoring paper writing efforts for the appropriate audiences, and for guiding me in my career development. His scientific vision has been enlightening and will influence my lifelong scientific career. I cannot thank enough.

I also thank the many researchers with whom I have collaborated with during my stay in the lab. Over the past several years, I've enjoyed the company and collaboration of many people at UCSD. In particular, Dr. Byung-Kwan Cho, Dr. Sang Woo Seo, Dr. Jay S.J. Hong, Dr. Youngseob Park, Dr. Joo-Hyun Seo, and Dr. Hojung Nam, have all provided hours of deep discussion and fruitful collaboration. I also thank Dr. Karsten Zengler, Dr. Pep Charusanti, Dr. Yu Qiu, Dr. Adam Feist, and Dr. Harish Nagarajan, who have all been senior mentors and role models to me. In addition, I enjoyed the researchers with whom I have extensively collaborated, including Ali Ebrahim, Aarash Bordbar, Steve Federowicz, Haythem Latif, Edward O'Brien, Eric Knight, Richard Szubin, and the various undergraduates that have worked with me. Without the support of Marc Abrams and Kathy Andrews, much of my work would have been, at best, delayed.

I also appreciate my thesis committee members collectively for their candor and for requiring my best efforts. I also thank the other advisors that prepared me for my doctoral research.

Lastly, none of this would have been possible without the loving support of my family. I thank my parents, Wonchae Kim and Misuk Yoon, who instilled in me a love for learning and a desire to work hard and serve others. I also am grateful for the daily loving smiles from my little sweetheart, Isabella Yoonha Kim. Her smiles always gave me the greatest motivation toward my research. Most importantly, I am grateful for by best friend, and wife, Sejin Koo, because without doubt her daily love and support got me to this point. I could not have finished so well without her.

Chapter 3, in total, is a reprint of the material as it appears in Kim, D.*, Hong, J.S.*, Qiu, Y., Nagarajan, H., Seo, J.H., Cho, B.K., Tsai, S.F., Palsson, B.Ø. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. PLoS Genetics, 8(8):e1002867 (2012). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 4, in total, is a reprint of the material as it appears in Cho, B.K.*, Kim, D.*, Knight E.M., Zengler, K., Palsson, B.Ø. Genome-scale Reconstruction of Sigma Factor Network in *Escherichia coli*. BMC Biology, 12(1):4 (2014). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 5, in total, is a reprint of the material as it appears in Seo, S.W.*, Kim, D.**, O'Brien, E., Latif, H., Szubin, R., Palsson, B.Ø. Deciphering Fur regulatory network highlights its role in iron homeostasis of *Escherichia coli*. *Submitted*. I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 6, in total, is a reprint of the material as it appears in Kim, D., Ebrahim, A., Seo, S.W., Bordbar, A., Palsson, B.Ø. Elucidating transcriptional regulation of nitrogen metabolism with systems approaches. *In preparation.* I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Vita

2006	B.S., Computer Science and Engineering, Seoul National University
2006	B.S., Biological Sciences, Seoul National University
2014	Ph.D., Bioengineering, University of California, San Diego

Publications

1. Lewis, N.E.*, Lee, D.H.*, Rutledge, A., Conrad, T.M., **Kim, D.**, Chaudhari, A., Bisariya, R., Barrett, C., Adkins, J.A., Smith, R.D., Palsson, B.Ø. E. coli learns to grow optimally on a non-native carbon substrate through laboratory evolution. *In preparation.*

2. Ebrahim, A., O'Brien, E., Lerman J.A., **Kim, D.**, Feist, A., Palsson, B.Ø. Parametrizing a Genome-Scale Model of Metabolism and Expression in *E. coli* with Multi-omic Data. *In preparation*.

3. Seo, S.W.*, **Kim, D.***, O'Brien, E., Latif, H., Szubin, R., Palsson, B.Ø. Deciphering Fur regulatory network highlights its role in iron homeostasis of *Escherichia coli*. *Submitted*.

4. **Kim, D.**, Ebrahim, A., Seo, S.W., Bordbar, A., Palsson, B.Ø. Elucidating transcriptional regulation of nitrogen metabolism with systems approaches. *In preparation.*

5. Federowicz, S., **Kim, D.**, Ebrahim, A., Lerman, J.A., Nagarajan, H., Zengler, K., Cho, B.K., Palsson, B.Ø. Determining the control circuitry of redox metabolism at the genome-scale. *PLoS Genet*, 10(4):e1004264 *(2014)*

6. Cho, B.K.*, **Kim, D.***, Knight, E.M., Zengler, K., Palsson, B.Ø. Genome-scale Reconstruction of Sigma Factor Network in *Escherichia coli*. *BMC Biology,* 12(1):4 (2014).

7. Chang, R.L., Andrews, K., **Kim, D.**, Li, Z., Godzik, A., Palsson, B.O. Structural Systems Biology Evaluation of Metabolic Thermotolerance in *Escherichia coli*. *Science,* 340(6137):1220-3 (2013).

8. Seo, J.H.*, Hong, J.S.*, **Kim, D.**, Cho, B.K., Huang, T.W., Tsai, S.F., Palsson, B.O., Charusanti, P. Multiple-omic data analysis of *Klebsiella pneumoniae* MGH 78578 reveals its transcriptional architecture and regulatory features. *BMC Genomics,* 29;13:679 (2012).

9. Nam, H.*, Lewis, N.E.*, Lerman, J.A., Lee, D.H., Chang, R.L., **Kim, D.**, Palsson, B.O. Network context and selection in the evolution to enzyme specificity. *Science,* 337(6098):1101-4 (2012).

10. **Kim, D.***, Hong, J.S.*, Qiu, Y., Nagarajan, H., Seo, J.H., Cho, B.K., Tsai, S.F., Palsson, B.Ø. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genetics*, 8(8):e1002867 (2012).

11. Cho, B.K., Federowicz, S.A., Embree, M., Park, Y.S., **Kim, D**, Palsson B.Ø. The PurR regulon in *Escherichia coli* K-12 MG1655. *Nucleic Acids Research*, 39(15):6456-64 (2011).

12. Kim, Y.K., Yu, J., Han, T.S., Park, S.Y., Namkoong, B., **Kim, D.H.**, Hur, K., Yoo, M.W., Lee, H.J., Yang, H.K., Kim, V.N. Functional links between clustered microRNAs: suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer. *Nucleic Acids Research*, 37(5):1672-81 (2009).

* Authors contributed equally

ABSTRACT OF THE DISSERTATION

Systems Evaluation of Regulatory Components in Bacterial Transcription Initiation

by

Donghyuk Kim

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2014
Professor Bernhard Ø. Palsson, Chair

In bacterial transcription, transcription initiation is arguably the most important regulatory point, because transcribing unnecessary genes into RNA could be a waste of energy, time and resources. There are multiple components which are involved in bacterial transcription initiation: RNA polymerase, σ-factors, transcription factors, and transcription start sites. Each component has been intensively investigated, however in a limited scope and mostly with low-throughput methods. New technologies, such as hybridization on microarray and deep-sequencing, enabled researchers to study each component in a systems level, in a combination of two or more components, and in comparison between different species. In order to facilitate the analysis, integration, and comparison, software, MetaScope, was developed to accommodate multiple genome-scale datasets to visualize, analyze, integrate, and compare. TSS-seq, modified 5'-RACE with deep-sequencing, gave a genome-scale landscape of transcription start sites, and comparison of TSSs of conserved genes between closely-related species, *E. coli* and *K. pneumoniae*, showed significantly different usage of promoters, which implies different regulation of orthologous genes. To further investigate properties of promoters which were identified by TSS-seq, ChIP-chip experiments were performed for σ-factors in *E. coli* to determine σ-factor regulons. From the reconstructed σ-factor network, extensive overlaps between regulons were observed. σ^{70} and σ^{38} share the largest set of genes in *E. coli*, and additional experiments revealed that those σ-factors work in competition and utilize the negative regulation by σ^{38}. ChIP-exo, which applies exonuclease to present better resolution of DNA-binding, and RNA-seq implemented more detailed identification of Fur regulon in *E. coli*. Reconstruction of Fur regulon completed the previous knowledge of bacterial response to iron change, and also enabled its role over iron metabolism. In order to understand how bacteria respond to nitrogen limitation, the same methods were used under conditions that were predicted from model-based prediction, and resulted in reconstruction of regulons for major transcription factors, NtrC and Nac. Determination of those regulons expanded the current knowledge of nitrogen metabolism and how it is regulated in bacteria. Thus, systems approaches enabled a genome-scale assessment of regulatory components in in multiple levels, and contributed to expansion of the current knowledge of bacterial transcription initiation.

Chapter 1: Genome-wide assessment of transcriptional regulatory components

Bacteria live in an ever-changing environment, where they have to respond and adapt again and again to these changes[1]. For instance, enterobacteria, such as *Escherichia coli*, live in the gut of mammals, where they are constantly exposed to changing available nutrients, available oxygen, pH, neighboring other bacterial species. A simplistic definition of a phenotype for bacteria is the composite of an observable characteristics or traits, such as its morphology, biochemical or physiological properties. A phenotype results from the expression of genes in a bacterial genome as well as the influence of environmental factors and the interactions between the two. Thus, in response to environmental change, bacteria changes its phenotype by adjusting expression of genes encoding enzymes that are necessary to implement its phenotype, and the information encoded in those genes comprises genotype.

Phenotype-genotype relationship

Since it is a necessity for those bacteria to regulate the expression of genes and synthesize enzymes accordingly to their needs, it becomes of interests to understand how genotype governs phenotype in bacteria (Figure 1). At the heart of this phenotype-genotype relationship, there are multiple levels of regulation on gene expression, such as transcriptional, post-transcriptional, translational, and post-translational regulation. Transcriptional regulation mostly depends on how transcription initiation is regulated in bacteria, while post-transcriptional regulation includes modification or processing of primary transcripts, and degradation of them. During translation process on transcribed products, there are multiple steps to regulate initiation of translation itself, the speed of translation, and where and when to stop translation. After translation, there is post-translational regulation, which is to change the activity of translated enzymes or to degrade no-longer needed proteins.

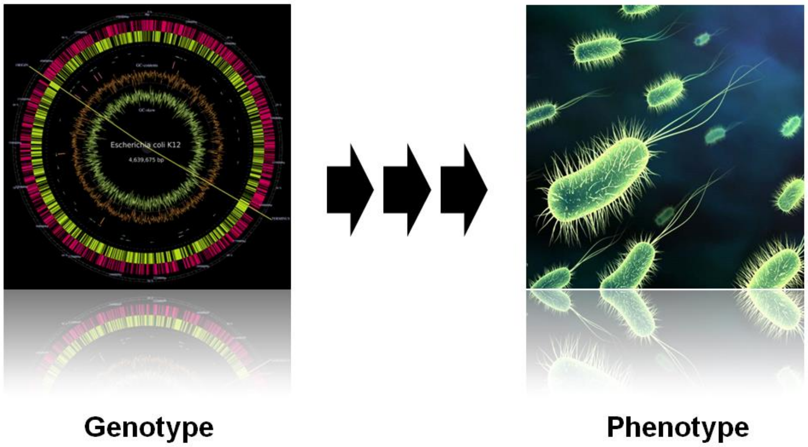


Figure 1. Genotype-Phenotype relationship

Regulatory components in transcription initiation

Among many regulatory points listed above, arguably regulation of transcription, more specifically regulation of transcription initiation, is the most important and efficient way of the regulatory process in the genotype-phenotype relationship. Thus, controlling the process of transcription is fundamental to gene expression, and gene regulation[1]. In all organisms of three kingdoms, the process of transcription is performed by DNA-dependent RNA polymerase, which transcribes information of genes in the DNA genome onto RNA transcripts. This RNA polymerase (RNAP) of bacteria is a complicated protein complex ($\alpha_2\beta\beta'\omega$), and it requires specificity factor, which is called σ-factor, in addition to the RNA-synthesizing machinery to recognize specific sequences in the promoter DNA, and guide where to initiate transcription from (Figure 2). There have been hundreds of σ-factors, or σ subunits, identified in a broad range of bacterial species so far[2]. Amongst them, primary σ-factors, such as *E. coli* σ^{70}, are responsible for housekeeping jobs and maximum growth during exponential phase[1], and in many species it is essential[3]. Different from housekeeping σ-factors, alternative σ-factors function during other growth conditions, such as stationary phase, and/or under stress conditions. For instance, *E. coli* σ^{38} is activated upon entering stationary phase[4], *E. coli* σ^{54} is known for expressing a wide set of genes in response to nitrogen-limiting environment, and *E. coli* σ^{24} is a minor σ-factor, specializing in response to stresses in general including heat shock, and stress on membrane[5].

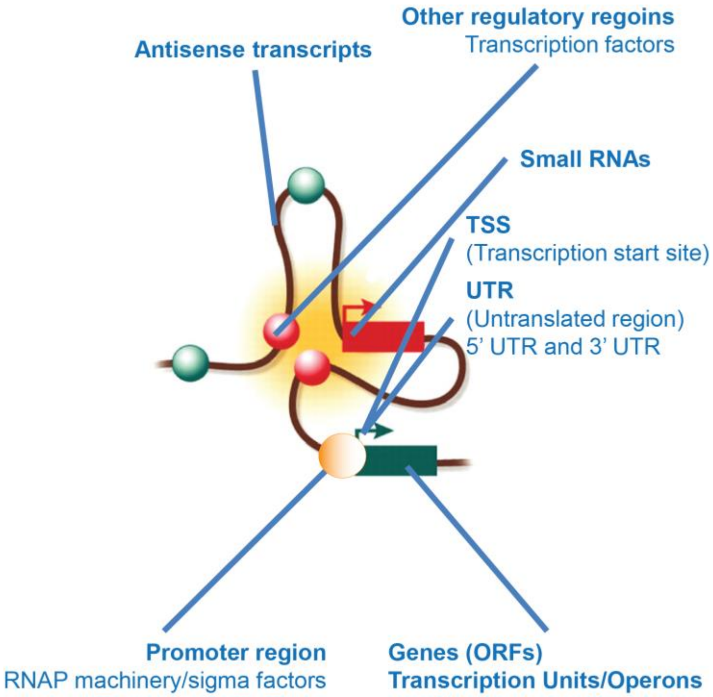


Figure 2. Multiple components of transcription regulation

Besides σ-factors, there is another group of DNA-binding proteins, which is called transcription factor (TF). TF is a protein that binds to specific sequences on genomic DNA, and it controls the process of transcription. TFs perform their functions to promote transcription of down-stream or target genes as an activator by facilitating the recruitment of RNAP towards promoter regions, or to repress transcription of target genes as a repressor by blocking the accessibility of RNAP to the promoters[6]. There are approximately 300 TFs in *E.*

coli K-12 MG1655, and they play pivotal roles in regulating gene expression to change the phenotype from genotype.

The transcription initiation site (TSS) is where transcription begins, and is the +1 position of primary transcript[7]. The promoter, which is recognized and bound by RNAP in association with σ-factors and TFs, governs the ability to initiate transcription and control the expression of genes, and it is directly upstream of the TSS, mostly within 50 bp upstream of TSS. Thus, TSS is another key component of regulation in transcription initiation. Determination of the precise locations of TSSs by experimental approaches is, thus, the necessity to accurately annotate the promoter region and the untranslated region[7]. There have been computational approaches to predict genomic positions of TSSs, however the sequence elements in promoters are short and not fully conserved in the sequence, thus it is likely to find similar sequence elements outside the actual promoter regions[8]. So it is important to determine TSS locations with experimental methods. The conventional way of identifying TSS is 5'-RACE (Rapid Amplification of 5' cDNA Ends), and this method has been extended with advent of sequencing technology, enabling genome-wide identification of TSSs in many organisms[8-13].

Experimental approaches to investigate regulatory components

The most popular way of identifying DNA binding events of DNA-binding proteins including σ-factors, and TFs, is chromatin immunoprecipitation (ChIP) and its variations[14]. ChIP-chip, also known as ChIP-on-chip, is using high-density microarray chip for hybridization with DNA libraries from immunoprecipitated chromatin samples. ChIP-chip has been widely used to investigate interactions between DNA-binding proteins and DNA *in vivo*. In *E. coli*, ChIP-chip has been intensively used to identify binding sites of RNAP[3, 11, 15], σ-factors[3], TFs[16-21], and NAPs[22, 23] (nucleoid-associated protein). ChIP-seq is more advanced version of ChIP-chip, which utilizes next-generation technology, and basically is to sequence DNA libraries generated from ChIPed DNA samples. ChIP-exo which is also called chromatin immunoprecipitation with exonuclease treatment is the newest technology of identifying binding sites of DNA-binding proteins. In brief, ChIP-exo applies a 5'-3' strand-specific exonuclease to a chromatin immunoprecipitated sample. Deep sequencing of an exonuclease-treated ChIP sample enables detection of exonuclease stop sites with near 1-bp resolution[24, 25]. ChIP-exo has much better resolution over the other long-established ChIP methods, and also has improved sensitivity, resulting in more accurate and more detection of binding sites of DNA-binding proteins.

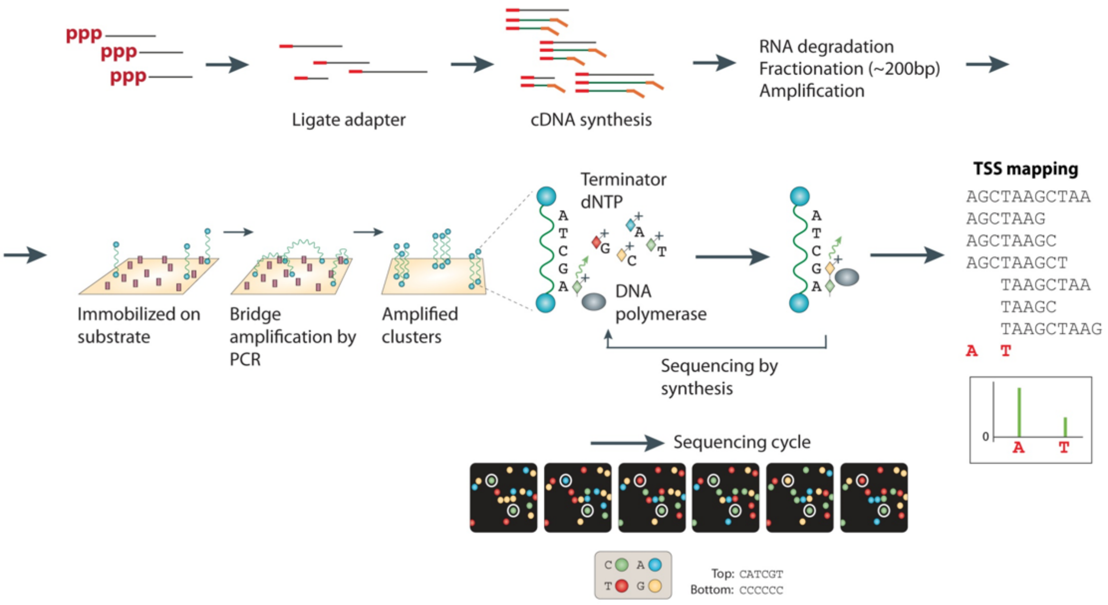


Figure 3. Experimental procedures of TSS-seq

TSS-seq (Transcription Start Site with sequencing) is an experimental method to identify genomic locations of TSSs with next-generation sequencing, which can generate genome-wide determination of transcription start sites[7, 12, 26] (Figure 3). Primary transcripts of bacteria have tri-phosphate groups at the 5' ends of them, while processed and degraded transcripts has mono-phosphate groups at those ends. Removal of processed or degraded products, monophosphate-dependent exonuclease was treated to chop out those products from 5’-3’ direction. Intact primary transcripts, then, are treated with pyrophosphatase to remove two phosphate groups leaving mono-phosphates at the 5’ ends. After that, RNA adaptor is ligated to pyrophosphate-treated RNA samples. Adaptor-ligated sample is then used as a template for cDNA synthesis with random primers. cDNA sample was amplified with PCR reaction to build sequencing library for deep-sequencing. The sequence reads were mapped onto reference genomic sequence to identify 5’ ends of those reads, which are TSSs.

Chapter 2: MetaScope: a genome browser with embedded functions for analysis and integration of multiple omics datasets

The tremendous amount of novel genomic information is inspiring new understandings of genomes on a global scale. With the publication of the first full genome sequence in the mid-1990s[27], it became possible, in principle, to identify all the gene products involved in complex biological processes in a single organism. In practice, almost 15 years later, this has proved difficult to accomplish using sequence information alone. Therefore, establishing the organization structure defined as the metastructure of a genome is a challenging task[11]. The organizational components include promoters, transcription start and termination sites, open reading frames, regulatory noncoding regions, untranslated regions, operons, and transcription units. Measurement of the components has been intensively supported by microarray or sequencing-based technologies on a genome-scale. Ultimately, integrating these multiple data types leads to the metastructure of genomes[11, 12]. However, many of the high-throughput genome-scale data types give rise to representational and computational challenges. For this reason researchers utilize genome browsers as a standard tool for exploring genomes, facilitating analysis and integration of genome-anchored data[28].

MetaScope aims to provide visualization software with embedded functions such as data manipulation and integration for various datasets, and build an improved annotation based on them. In addition, MetaScope is designed to support interactive environment where molecular biologists with minimal computational skills can visualize and process multiple omics data mapped onto the genome, and share their biological findings.

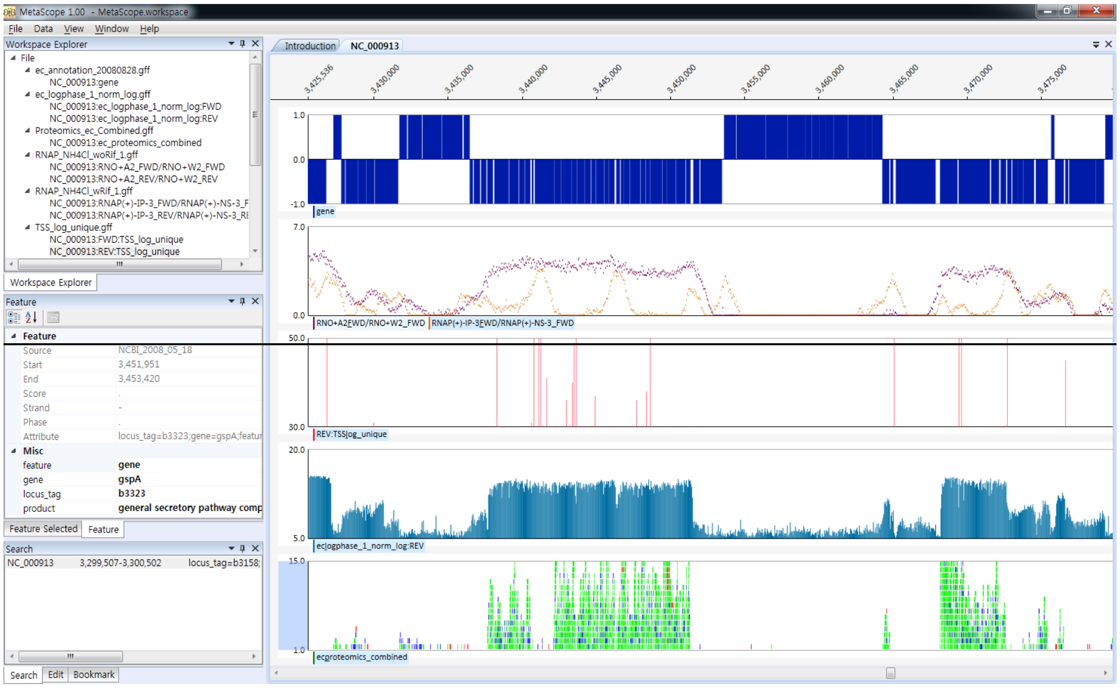
Implementation

The MetaScope is implemented in C# programming language, and runs on any computer platform with .NET framework 4.0 or higher.

Overview

The current range of tools for visualizing and analyzing genome-scale data includes software such as NimbleGen SignalMap (http://www.nimblegen.com), Integrated Genome Browser[29], Argo Genome Browser[30] and Gaggle Genome Browser[31]. Dealing with large volume data requires visualization performance as offered by SignalMap, and data operability and flexible rich user interface as provided by Integrated Genome Browser, Argo Genome Browser or Gaggle Genome Browser. For example, elucidating the transcription unit architecture of *E. coli* requires capturing organizational components of the genome[11]. Elucidating these components necessitates handling various data types of significant volume: ChIP-chip, expression profiling, transcription start site and proteomic data under various growth conditions. In order to analyze and integrate these datasets through cross-referencing different data types, software with data operation functions embedded within the visualization is required to facilitate this process.

MetaScope addresses these challenges by providing various data operation and integration functions, visualization performance, and highly flexible and configurable user interface. As enumerated in Table 1, MetaScope supports assorted data processing functions for tracks, several operation functions for data features, and integration functions, which can be used to generate genomic annotations, such as transcription units. Additionally, MetaScope provides zoomable and scrollable view and is contrived to visualize large amount of heterogeneous data types including unprocessed datasets, and processed datasets such as known genomic annotation, genomic association between DNA-binding proteins and genome from ChIP-chip data, and contiguous transcript signals from expression profiling data. MetaScope provides a customizable user interface by enabling every window to be moved and organized to satisfy user preference, as shown in Figure 4.


MetaScope 1.00 - MetaScope.workspace
File
Data
View
Window
Help
Workspace Explorer
File
ec_annotation_20080828.gff
NC_000913:gene
ec_logphase_1_norm_log.gff
NC_000913:ec_logphase_1_norm_log:FWD
NC_000913:ec_logphase_1_norm_log:REV
Proteomics_ec_Combined.gff
NC_000913:ec_proteomics_combined
RNAP_NH4Cl_woRif_1.gff
NC_000913:RNO+A2_FWD/RNO+W2_FWD
NC_000913:RNO+A2_REV/RNO+W2_REV
RNAP_NH4Cl_wRif_1.gff
NC_000913:RNAP(+)-IP-3_FWD/RNAP(+)-NS-3_F
NC_000913:RNAP(+)-IP-3_REV/RNAP(+)-NS-3_RE
TSS_log_unique.gff
NC_000913:FWD:TSS_log_unique
NC_000913:REV:TSS_log_unique
Workspace Explorer
Feature
Feature Selected
Feature
Search
Search
Edit
Bookmark
Introduction
NC_000913
3,425,536
3,430,000
3,435,000
3,440,000
3,445,000
3,450,000
3,455,000
3,460,000
3,465,000
3,470,000
3,475,000
1.0
0.0
-1.0
gene
7.0
0.0
RNO+A2FWD/RNO+W2_FWD
RNAP(+)-IP-3FWD/RNAP(+)-NS-3_FWD
50.0
30.0
REV:TSS_log_unique
20.0
5.0
ec_logphase_1_norm_log:REV
15.0
1.0
ecproteomics_combined

Feature	
Source	NCBI_2008_05_18
Start	3,451,951
End	3,453,420
Score	.
Strand	-
Phase	.
Attribute	locus_tag=b3323;gene=gspA;featur
Misc	
feature	gene
gene	gspA
locus_tag	b3323
product	general secretory pathway comp

NC_000913 3,299,507-3,300,502 locus_tag=b3158;

Figure 4. MetaScope showing multiple datasets.

Streamlined workflow

Any data file in GFF format can be uploaded in MetaScope by simple drag and drop or by using the application menu. MetaScope also supports a workspace file, which stores the list of data files uploaded, track settings including height, scaling information, color of visual data points and display type. This workspace file can be opened in the same manner data files are uploaded.

Upon loading one or more data files, MetaScope categorizes all datasets based on chromosome ID and data type, generating separate tabs for each chromosome ID and displaying data types on corresponding tracks. The workspace explorer window shows what data files are uploaded, what chromosome IDs and data types are recognized. The feature property window displays all the information about the feature, which the mouse cursor is hovering over. In a similar fashion, selected feature property window shows information on multiple features selected by mouse dragging. Search window supports looking up datasets uploaded with the keyword input, and allows prompt navigation to the genomic position by double-clicking the search result. MetaScope also supports bookmark function for user-friendly browsing of the genome-scale data.

The main window of MetaScope visualizes genome annotation and datasets, which researchers can navigate by zooming and scrolling over genomic positions. Each tracks on the main window shows each data type and can be scaled to focus on the range of signal intensity of interest to the user. MetaScope supports four display styles for the data tracks; bar, point and line display styles are good for displaying transcriptomic datasets, and stack display style is suitable for showing proteomic data (Figure 4). Two or more tracks can be overlapped together, allowing researchers to compare and analyze multiple datasets. For example, overlapping RNA polymerase ChIP-chip data with expression profiling data gives a genome-wide landscape of RNA polymerase occupancy and transcription level. Similarly, expression profiling data can be overlapped onto transcription start site data in order to determine starting and ending positions of transcripts.

MetaScope supports assorted data manipulation function in data track and feature levels. Combination of these functions enable researchers to compare and analyze datasets from biological replicates, validate and process by cross-referencing between different data types, and integrate and build a new genomic annotation based on canonical annotations and experimental datasets. In this context, MetaScope supports designated integration functions for assembling transcription unit annotation harboring organizational elements of the bacterial genomes from the genome-scale data in a step-by-step manner.

Performance

MetaScope x86 version can accommodate large volume of datasets up to around 750 MB in size, and work readily on a desktop computer with 2 GHz single core CPU and 3GB RAM. MetaScope x64 version can handle larger datasets, and tested up to 3GB datasets.

Table 1. MetaScope feature list.

Feature category	Features
Input file format	GFF2, GFF3
Navigation	Zooming, scrolling, jumping to a certain genomic position, bookmarking, and searching by gene
Track display style	Bar, point, line and stack
Track operation	Overlapping, scaling, averaging, differencing, summing, merging and filtering
Feature operation	Selecting, uniting, merging, filtering, moving, copying, creating, editing, and deleting features
Integration function	Building transcription unit annotation
Others	Customizing application layout, editing history, splitting data window and managing workspace files for different projects

Chapter 3: Comparison of bacterial regulatory elements by transcription start site profiling

Escherichia coli K-12 MG1655 and *Klebsiella pneumoniae* MGH78578 belong to the same enteric family of bacteria of the class gammaproteobacteria. While E. coli K-12 MG1655 represents an extensively studied laboratory strain that is not known to be pathogenic, *K. pneumoniae* MGH78578 is a well-known pathogenic strain isolated from a patient with pneumonia[32]. There have been many comparative genomics approaches used to understand the similarities of closely related species in a wide range of genera such as *Escherichia*, *Klebsiella*, *Salmonella*, and *Listeria*[33-37]. These comparative genomics studies have mostly focused on comparing the gene contents, either shared or specific for each genome. However, it is also important to investigate the similarities and differences in non-coding regulatory elements including promoter, 5' untranslated region (5' UTR), and small RNA (sRNA), due to their influence on transcriptional and post-transcriptional processes.

The transcription start site (TSS) is where transcription begins and is the +1 position of the 5' untranslated region (5' UTR) of mRNA. The promoter, which governs the ability to initiate transcription and control the expression of genes, is directly upstream of the TSS. The identification of promoter elements in DNA by computational methods depends on the statistical analysis of consensus sequences as overrepresented regions[38, 39]. Regulatory sequence elements have been studied by computational methods based on the genomic sequence of the non-coding upstream region[40-42], however those sequence elements in promoters are short and not fully conserved in the sequence. Thus, there is a high probability of finding similar sequence elements outside the promoter regions. In the case of the TSS, the region is not overrepresented enough by any consensus sequences and is thus difficult to predict by computational efforts. However, when the TSS is known, the DNA region most likely to contain regulatory binding sites is circumscribed, and the effectiveness of searching sequence motifs of interest is greatly enhanced[8]. Thus, determining the precise locations of TSSs by experimental methods is necessary to accurately annotate the promoter region and the untranslated region. Knowledge of the 5' UTR region is important for studying the sequence and structure of the 5' end of mRNA (which is associated with transcription regulation, mRNA transcript stability, and translational efficiency) because translational efficiency in bacteria is often controlled by RNA-binding proteins, noncoding regulatory RNAs, endoRNases, the 30S subunit of ribosome, and structural rearrangements within 5' UTR[43].

Genome-wide identification of TSSs with the aid of deep sequencing has allowed researchers to reveal a landscape of TSSs across the whole genome in many microorganisms, including E. coli[9-11, 44], H. pylori[26], G. sulfurreducens[12], and other species[8, 13]. In these studies, experimental TSS datasets were used to understand the transcription architecture, to appreciate the complexity of genomic structure, and to analyze regulatory elements for each species. Comparison of regulatory elements, which can be addressed by experimentally determined TSSs under the same growth condition, is expected to elucidate any regulatory similarities or differences, based not only on the genomic sequence, but also on the transcriptional context of compared species as well.

Here, we carried out the genome-wide TSS profiling experiments for *E. coli* K-12 MG1655 and *K. pneumoniae* MGH78578 to accurately determine the boundaries in the regulatory regions between the promoter region and the 5' UTR. The upstream regulatory regions between those two closely related species were then compared to investigate whether those regions are conserved and organized in similar manners. In addition, we used the TSS dataset to identify sRNAs in *K. pneumoniae*, because very little is known about them. We then compared the *K. pneumoniae* sRNAs to orthologous sRNAs in *E. coli*, in terms of sequence conservation and their target sites. The range of sequence conservation or diversion between non-coding regulatory elements in interspecies microorganisms could lead to insights about regulatory features that may also play similar roles in the respective species.

Experimental identification of TSSs in E. coli and K. pneumoniae

Primary mRNA transcripts in prokaryotes are triphosphorylated at the 5’ ends. We isolated total RNA from *E. coli* and *K. pneumoniae* cells growing in mid-exponential phase, and enriched primary mRNAs by removing any monophosphorylated ribosomal 23S, 16S rRNA, tRNA, and any degraded mRNAs by treatment with terminator exonuclease[12, 26]. By using a modified 5’RACE (rapid amplification of cDNA ends) followed by deep sequencing as previously described[12], libraries were prepared and sequenced to determine potential TSSs for each strain of *E. coli* K-12 MG1655 and *K. pneumoniae* MGH78578. These TSS libraries yielded > 11.6 million and > 2.4 million sequence reads for *E. coli* and *K. pneumoniae,* respectively. 15.70% and 19.60% of those sequence reads were uniquely mapped with 36 bp read length onto the *E. coli* and *K. pneumoniae* reference genomes respectively. Unique sequence reads that perfectly matched the respective genome sequence were mapped to annotate a total of 3,746 and 3,143 TSSs for the *E. coli* K-12 MG1655 and *K. pneumoniae* MGH78578 genome, respectively. The average number of TSS reads of *E. coli* and *K. pneumoniae* TSSs was 107.8 and 78.5, respectively. The lower number of identified TSSs for *K. pneumoniae* could be due to a lesser number of sequence reads, and this factor was taken into account in further analysis.

To verify the quality of the TSS data, we compared our experimental E. coli TSS data with previously published *E. coli* TSS datasets[10, 11]. There is no public genome-wide TSS dataset available for *K. pneumoniae*, which is why only TSS data for *E. coli* was used for this analysis. In RegulonDB, there are 1258 upstream sense TSSs annotated for *E. coli*, generated by 5’ triphosphate enrichment method. 624 (49.6%) TSSs out of 1258 matched exactly with TSSs of this study, and 257 (20.4%) TSSs matched within 3 bp tolerance. Thus, 70.0% of known TSSs from RegulonDB agreed with the TSSs from our study. From the TSS dataset generated without 5’ triphosphate enrichment method, 3661 TSSs were reported for the exponential growth condition. 1603 (43.8%) TSSs matched exactly with TSSs of this study, and 527 (14.4%) TSSs matched within 3 bp tolerance. In sum, 58.2% of TSSs were found in TSSs of this study. A comparison of our TSS dataset with two other datasets suggested TSS datasets generated by a similar method were in better agreement, and *E. coli* TSSs determined by an independent experiment were matched by TSSs used in this study.

A genome-wide TSS landscape of *E. coli* and *K. pneumoniae* was built by assigning TSSs to the nearest downstream gene including ORFs and sRNAs, but excluding TSSs located beyond 700 bp from the translation start site of the closest ORF in a strand specific manner (Figure 5A). In E. coli, TSSs were assigned to 2654 genes, while TSSs in *K. pneumoniae* were assigned to 2301 genes (2175 genes in the main chromosome, and 126 genes in the plasmids).

Identification of small RNAs in K. pneumoniae

While over 80 sRNAs have been identified and experimentally verified in *E. coli*, very little is known about *K. pneumoniae* sRNAs. Identifying the occurrence of sRNAs and determining their boundaries in a genome-wide manner is challenging, especially for less studied organisms, because sRNAs generally have no clear-cut signatures unlike protein-coding genes, which are specified by a genetic code. In order to overcome limitations of previous experimental approaches, and to interrogate sRNAs in a genome-wide manner, a deep-sequencing approach was applied and proved successful[45]. Before investigating the possible presence of sRNAs in *K. pneumoniae*, *E. coli* TSS datasets were analyzed to assess how many currently annotated sRNAs in *E. coli* could be identified under the experimented condition, and how well TSS signals were matched with 5’ ends of those sRNAs. In addition, TSS datasets generated with the 5’ triphosphate enrichment method in this study were compared to four other TSS datasets generated by different methods[9-11] in the light of using 5’ triphosphate enrichment. Many sRNAs are subjected to post-transcriptional processing, however, which results in an accumulation of shorter products with 5’ monophosphate[46-48]. Therefore, only unprocessed sRNAs or precursor transcripts of sRNAs, which have 5’ triphosphate and can be detected by this method, were analyzed.

Of 81 annotated sRNAs in *E. coli*, 58 (71.6%) had corresponding TSSs, and were thus considered to be expressed during exponential growth. Expression profiling data taken from the previous study[11] also supported the expression of those sRNAs under the experimented condition, although *rprA* showed no significant expression according to that data. This could be because *E. coli* RprA transcript is subject to specific endoribonuclease cleavage[46], resulting in the accumulation of processed shorter form, which is not long enough to be detected by the tiling array. TSS signals were well matched with the 5’ ends of unprocessed or precursor transcripts of 58 sRNAs including *rprA* (Figure 5B). In comparison, TSS datasets generated by deep-sequencing without 5’ triphosphate enrichment[11] presented TSS signals for 44 sRNAs (54.3%). Three other TSS datasets, generated by other methods[9, 10], were obtained from RegulonDB database (http://regulondb.ccg.unam.mx/). They showed TSSs assigned to 11 (13.6%), 6 (7.4%), and 0 (0%) sRNAs for each method (Figure 5B). Thus, experimental TSS generated by deep sequencing is a practical indicator that shows the occurrences of sRNAs in *E. coli* and determines the genomic positions of their 5’ ends. Additionally, our TSS dataset detected the largest number of annotated sRNAs in E. coli, compared to previous methods. We believe this result shows that the TSS dataset for *K. pneumoniae*, generated with the same method, can be used to detect possible sRNAs in that species and to determine the 5’ ends of those sRNA candidates.

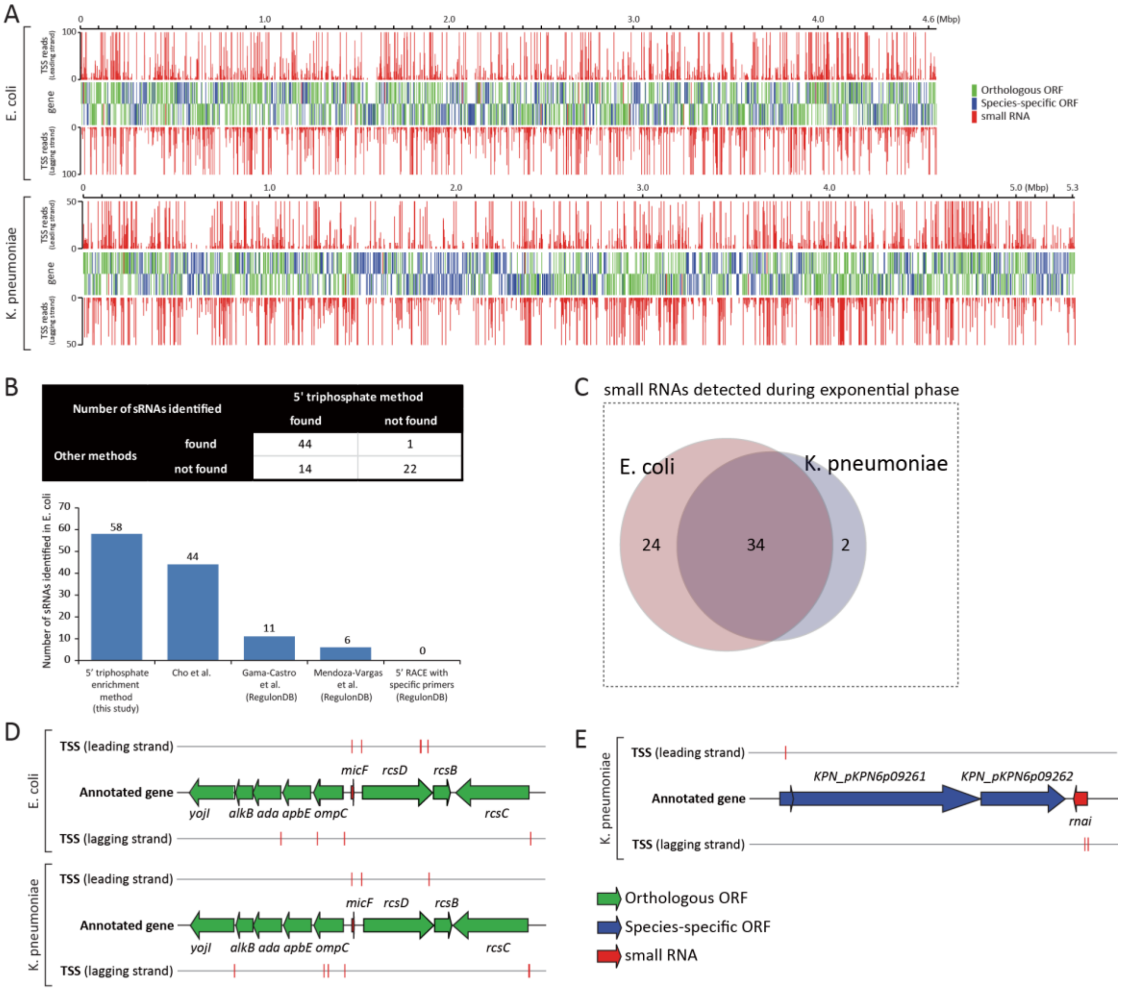


Figure 5. Experimentally determined TSSs and their association with annotated genes. (A) Genome-wide TSS mapped onto *E. coli* and *K. pneumoniae* genome annotation. (B) Number of *E. coli* sRNAs detected with 5 TSS datasets generated by different methods. (C) Number of sRNAs detected from *E. coli* and *K. pneumoniae* during the exponential growth. (D) Schematic drawing of annotated TSSs assigned to orthologous *micF* sRNA and coding genes surrounding *micF* in *E. coli* and *K. pneumoniae*. (E) Schematic drawing of annotated TSSs assigned to *K. pneumoniae* sRNA, *rnai*, and coding genes near *rnai*.

In order to identify and confirm the occurrence of sRNAs in *K. pneumoniae* by experimentally determined TSSs, tentative sRNA candidates should first be predicted by computational methods. A number of computational algorithms have been developed over the last decade for the purpose of predicting sRNAs in bacterial genomes, and primary sequence conservation in closely related species is one of the most useful data types for predicting whether a genomic sequence corresponds to a sRNA[49]. Since a majority of *E. coli* annotated coding genes (63.7%) have homologs in the *K. pneumoniae* genome, and conserved sRNAs are frequently identified adjacent to conserved coding genes in other organisms, we looked up the closest orthologous ORFs to annotated sRNAs of *E. coli*, and then searched tentative sRNA sequences in *K. pneumoniae* genomic sequences bound to those neighboring orthologous genes. For example, in E. coli, *micF* sRNA is surrounded by *ompC* and *rcsD*, both of which are conserved coding genes between the two species. The *K. pneumoniae* genomic sequence bound to *ompC* and *rcsD* orthologous ORFs was used for searching the genomic sequence of *micF* by sequence alignment (Figure 5D, detailed method described in Methods section). This approach was supplemented by running Infernal algorithm[50] with sRNA models from the Rfam database 10.1 (http://rfam.sanger.ac.uk/). Using this combined approach, we identified 48 tentative sRNAs in the *K. pneumoniae* genome, and 36 of them were expressed by associated TSSs (Figure 5C). Expression of those sRNAs was also supported with expression profiling data, with the one exception being *rprA*. *rprA* of *K. pneumoniae* showed no significant level of transcription according to the expression profiling data, however *rprA* had an assigned TSS with 1865 reads, which was also observed in *E. coli rprA* with an assigned TSS of 3012 reads. This indicates a possibility of post-transcriptional processing of *K. pneumoniae* RprA transcript as is the case in *E. coli*. 47 of 48 putative sRNAs were located in the main chromosome (NC_009648) of *K. pneumoniae*, while one sRNA, *rnai*, was found in the plasmid (NC_009652) (Figure 5E).

Of 36 small RNAs detected during the exponential phase in *K. pneumoniae*, 34 had orthologous sRNAs in *E. coli*, leaving 2 non-orthologous sRNAs, *rnai*, and *ryhB-2*. Their expression was supported by TSS and expression profiling. *ryhB-2* was so-named because another orthologous *ryhB* sRNA was identified in a position between orthologous ORFs *yhhX* and *yhhY*. *rnai* non-coding RNA is an antisense repressor of the replication of some E. coli plasmids[51]. While *E. coli* K-12 MG1655 does not have any plasmid, *K. pneumoniae* MGH78578 has 5 plasmids, one of which (NC_009652) contains *rnai* sRNA.

Similar usage of regulatory features

The majority of *E. coli* annotated genes, 1945 (73.5%), were annotated with a single TSS, and the remaining 26.5% had multiple TSSs mainly ranging from 2 to 7, allowing alternative transcripts (Figure 6A). Similar to the complex organization of promoter regions and usage of multiple TSSs shown in *E. coli*, 534 (22.8%) of *K. pneumoniae* annotated genes had multiple TSSs, leaving a large fraction of genes, 1802 (77.2%), which were assigned to a single TSS (Figure 6A).

In order to investigate other regulatory features shared by *E. coli* and *K. pneumoniae,* the length distribution of the 5’ UTR bounded by experimental TSS and translational start site was calculated, and possible sequence motifs were examined with the MEME motif search algorithm[52]. The length of the 5’ UTR ranged from 0 to 700 nucleotides, with the most abundant length found to be between 25 to 35 bp for both bacterial species (Figure 6B). For 18 genes from *E. coli* and 10 genes from *K. pneumoniae*, leaderless mRNAs with the TSSs corresponding exactly to the start codon were found. The leaderless mRNAs encoded proteins of various functions.

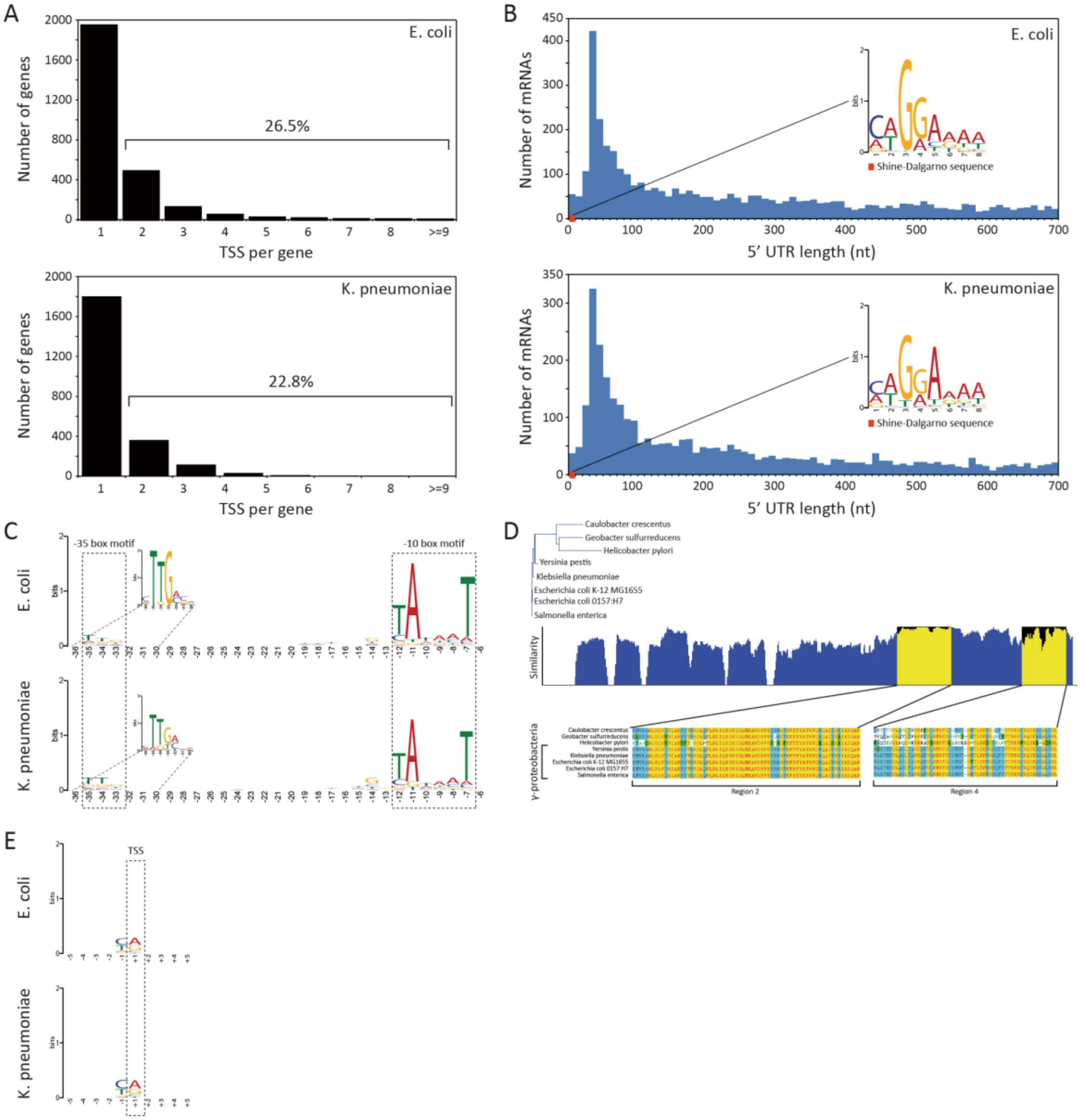


Figure 6. TSS annotation and structure of promoter region and 5' UTR. (A) Number of TSSs assigned per annotated genes. (B) Distribution of 5' UTR lengths for *E. coli* and *K. pneumoniae*, and the Shine-Dalgarno sequence motif. (C) Sequence motif of promoter region containing -10 and -35 boxes. (D) Conservation of RpoD amino acid sequences of 5 species in gammaproteobacteria and 3 other species belonging to proteobacteria. (E) Di-nucleotide preference near the TSS site.

Experimentally determined TSSs in *E. coli* and *K. pneumoniae* were used to detect the Shine-Dalgarno (SD) sequence of the ribosome binding site (RBS). Expecting to find that motif within the boundaries of the 5' UTR, which are defined by the TSS and translation start site of the downstream ORF, we took sequences from 5’ UTR regions in *E. coli* and *K. pneumoniae* and searched for consensus motifs. A conserved caGGaaaa sequence motif (lower-case characters indicate an information content <1 bit) was found in *E. coli*, and an identical conserved caGGaaaa motif was also found in the 5’ UTR of *K. pneumoniae*. The most dominant distance between the SD sequence motif and translational start site was 6 nucleotides in both species. Motif logos for both species are illustrated in Figure 6B.

Bacterial promoters usually contain specific sequences, which RNA polymerase-associated sigma factors can recognize and to which they can bind. For example, the E. coli housekeeping sigma factor σ70 (rpoD, b3067) is known to recognize -10 (TATAAT) and -35 (TTGACA) boxes[53]. Although sequence motifs of major *E. coli* sigma factors have been investigated by experimental and computational approaches, less is known for *K. pneumoniae* sigma factors and their binding motifs. *E. coli* and *K. pneumoniae* are closely related, and they share major sigma factors, such as *rpoD*, *rpoS*, *rpoH*, *rpoN*, and *rpoE* with a high level of amino acid sequence conservation over 95%, with the exception of *rpoN* that has 89.8% amino acid sequence similarity. Since sigma factor σ^{70} is housekeeping during exponential growth in *E. coli* and presumably in other gammaproteobacteria including *K. pneumoniae* as well, conservation of subregions 2 and 4 of bacterial sigma factor σ^{70}, which are known to recognize the -10 and -35 boxes, can give insights toward understanding the promoter structure of *K. pneumoniae*. Thus, amino acid sequences of *rpoD* of 5 strains belonging to gammaproteobacteria and 3 strains belonging to other classes were aligned and analyzed (Figure 6D). Notably, region 2, which recognizes the -10 box, was perfectly conserved among species in gammaproteobacteria, and region 4, which recognizes the -35 box, was almost conserved as well. Since the conservation of sigma factor σ^{70} subregions recognizing sequence motifs in the promoter and the expression of housekeeping *rpoD* in *E. coli* and *K. pneumoniae* was confirmed with the TSS dataset and expression profile, it is likely that the promoter structure of those species are identical. Thus, TSSs in *E. coli* and *K. pneumoniae* identified in this study were used to find sequence motifs of the promoter region, which includes the -10 and -35 boxes, in order to see whether two closely related bacteria share similar or identical promoter sequence motifs. We extracted 50 bp long sequences directly upstream of the TSSs, which are long enough to cover the -10 and -35 boxes, and ran the MEME motif search algorithm. As a result, the consensus sequence of the extended Pribnow box motif (tgnTAtaaT) including the -10 box was obtained, and the -35 box sequence motif (cTTgaca) was also found, as expected (Figure 6C). Moreover, the most dominant distances between the -10 box and TSS and between the -10 and -35 boxes were also the same in both bacteria. Although the sequence motif obtained herein is based on genome-wide TSS profiles generated only under exponential growth and other sigma factors having different binding sequence motifs may play a minor role in transcription regulation under the experimented condition, overrepresented sequence motifs of promoter regions in *E. coli* are in accordance with prior knowledge, and the two species in this study showed identical sequence motifs of the promoter. Thus, these closely related species seem to share identical promoter structures, reflecting a high conservation of major sigma factors.

Previous studies have shown evidence of a purine (A/G) preference at the TSS in *E. coli*[54]. Here, we investigated if the experimentally derived TSS data provide insights into any such nucleotide preference at the TSS. Thus, nucleotide preferences from -5 to +5 sites surrounding the TSSs for *E. coli* and *K. pneumoniae* were calculated. The current experimentally derived TSSs in both species also showed a significant dinucleotide preference at the +1 TSS and -1 site (Figure 6E). In *E. coli*, 78.6% of the TSSs were represented by purine base (45.2% A and 33.4% G) at the TSS. Similarly, 79.4% of *K. pneumoniae* TSSs presented the purine base (48.0% A and 31.4% G) at that site. Interestingly, another nucleotide preference at the -1 site, the nucleotide before the TSS and the last nucleotide that is not transcribed, was observed in both species. In *E. coli*, 80.2% showed the pyrimidine base (35.4% T and 44.8% C) preference at the -1 site. Likewise, in *K. pneumoniae,* 81.5% of cases also showed the pyrimidine base (31.0% T and 50.5% C) at the -1 site. Flanking regions ranging from +2 to +5 sites and -2 to -5 sites showed no significant nucleotide preference (Figure 6E). Thus, both species showed the purine preference at the +1 TSS and the pyrimidine preference at the -1 site. In accordance with this observation, *H. pylori*, which belongs to a different class of alphaproteobacteria, also showed purine preference at the TSS (66.0% A or G) and pyrimidine preference at the -1 site (68.3% T or C)[26]. Similar to the dinucleotide sequence preference at +1 and -1 sites found in bacteria, transcription from the *S. cerevisiae* promoter[55] and the mammalian[44] promoter preferentially starts with a purine at position +1, having a preference for pyrimidine at position -1.

These results suggest that *E. coli* and *K. pneumoniae* share many regulatory features at the transcriptional and translational level. They have a conserved promoter structure reflecting preserved sigma factors, use multiple TSSs that extensively increase transcriptome complexity by resulting in alternative transcripts, and show dinucleotide preference near the TSS position. In addition to this similarity in transcriptional features, *E. coli* and *K. pneumoniae* exhibit conserved Shine-Dalgarno sequence motifs, the same distance from Shine-Dalgarno motif to translation start site, and 5’ UTR length distribution, suggesting similarity in regulatory features of translation.

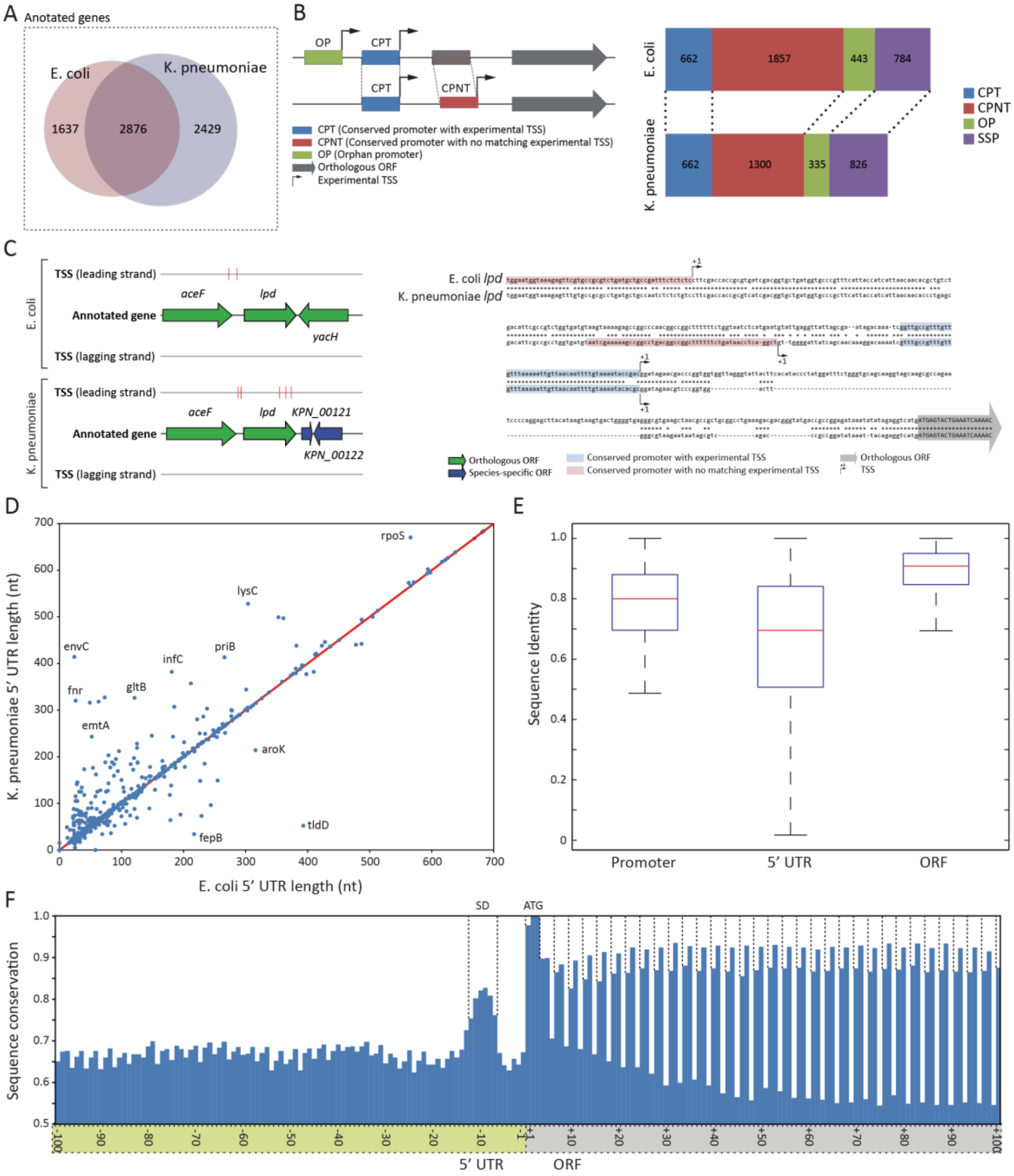


Figure 7. Different organization of upstream regulatory region between *E. coli* and *K. pneumoniae.* (A) Venn diagram showing orthologous genes and species-specific genes between *E. coli* and *K. pneumoniae*. (B) 4 different types of promoter regions, and their numbers identified in two species. (C) Schematic drawing of annotated TSSs and sequence comparison of regulatory region upstream of *lpd*. (D) Length difference between the pairs of comparable 5' UTR. (E) Comparison of sequence conservation of promoter, 5' UTR, and ORF regions. (F) Sequence conservation of genomic regions surrounding translation start sites.

Different organization of the upstream regulatory region

While *E. coli* and *K. pneumoniae* share several regulatory features, it is still unknown whether the two species use them to regulate gene expression in the same manner. Thus, we analyzed the usage of regulatory elements upstream of orthologous genes between two strains in order to investigate whether those conserved genes are regulated in a similar or different manner. The orthologous genes present in *E. coli* and *K. pneumoniae* were selected by reciprocal alignments using a threshold of 50% amino acid sequence similarity and 50% alignment length between the encoded proteins, resulting in a set of 2,876 orthologs (Figure 7A). 2962 (79.1%) *E. coli* TSSs were assigned to orthologous genes defined herein, and in *K. pneumoniae*, 2317 (73.1%) of TSSs were assigned to orthologous genes. Considering 63.7% (2876 out of 4513) of genes in *E. coli* and 54.2% (2876 out of 5305) of genes in *K. pneumoniae* were orthologous, detection of over 79.1% of TSSs in *E. coli* and 73.1% in *K. pneumoniae* assigned to orthologous coding genes implies over 73% of primary transcripts were expressed from operons or transcription units having orthologous genes at the first position. In *E. coli*, the average number of genes in an operon is about 1.5 as reported previously[11], and operons containing orthologous genes in *E. coli* have a tendency to keep their sequential position in *K. pneumoniae*, suggesting possible conservation of operon structures. This result suggests that the majority of primary transcripts were expressed from operons containing conserved orthologous genes during exponential growth in both species. Thus, further analysis of regulatory regions upstream of orthologous genes with genome-wide TSSs covers a majority of expressed gene contents under the experimented condition.

Despite the fact that orthologous genes were used to express the majority of primary transcripts during exponential growth, regulatory regions upstream of those conserved coding genes were organized in a different manner with multiple TSSs (Figure 7B). In order to perform a detailed investigation comparing promoter regions between two species, each TSS was used to define a promoter region. A promoter region was defined as 50 bp long nucleotide sequences upstream of each TSS, which was long enough to include most of the regulatory elements identified, including the -1 site, -10 box, and -35 box, but not too long to exclude unnecessary sequences. Then, the promoter region was categorized into one of four groups, based on sequence conservation of the promoter region and presence of an experimental TSS: conserved promoter region with TSS (CPT), conserved promoter region with no matching TSS (CPNT), orphan promoter region (OP), or species-specific promoter (SSP). CPT was defined as a promoter region with a conserved sequence in both strains with a matching experimental TSS, and was used to define the promoter region and 5' UTR, which were comparable between the two species. CPNT was defined as a promoter region with a conserved sequence in both strains, however with experimentally determined TSSs in only one species. Similarly, OP was defined as a promoter region with no conserved sequence between *E. coli* and *K. pneumoniae*, and with experimental TSSs in only one species. SSP was defined as a promoter region upstream of non-orthologous genes. (More details described in Methods section)

If the sequence of regulatory regions upstream of orthologous coding genes is also conserved, then conserved promoter (CPT) should be the most frequent type of promoter region. However, an exhaustive comparison of promoter regions resulted in only 662 conserved promoters (CPT) between *E. coli* and *K. pneumoniae*, which covered 17.7% of TSSs and corresponding promoter regions in *E. coli* and 21.1% in *K. pneumoniae*. An unexpectedly small portion of conserved promoter regions with matching TSSs in two species under the exponential growth supports a different organization of regulatory regions containing multiple TSSs and their associated promoters between those two closely related species. Interestingly, in both species, the promoter type with the largest number was the conserved promoter with no matching TSS (CNPT). In *E. coli*, 49.6% of TSSs were associated with promoters with conserved sequence, and no matching TSSs were found upstream of corresponding orthologous genes in *K. pneumoniae*. Similarly, 41.3% of TSSs of *K. pneumoniae* were associated with that type of promoter. A smaller number of TSSs was detected in *K. pneumoniae* versus *E. coli*, despite *K. pneumoniae* having the larger genome. This was possibly due to fewer raw reads being obtained from the *K. pneumoniae* TSS library. Thus, it is arguable that the portion of conserved promoters with matching TSSs could increase as the coverage of TSS reads goes up. However, over 40% of promoters with conserved sequences had TSSs in one species, but had no matching TSS in the other species. Thus, the regulatory regions upstream of orthologous genes are organized in a different manner, despite a large portion of promoters having conserved sequences between two species. This suggests different sets of TSSs are used to express those orthologous genes. For example, *lpd* had two experimental TSSs in *E. coli* and *K. pneumoniae* (Figure 7C). Proximal TSSs were matching, and had a highly conserved promoter sequence. Distal TSSs of *lpd* had conserved sequences, but were in different locations. Moreover, promoters with no conserved sequence and TSS in one species (OP, orphan promoter) also support that interpretation. Thus, while two closely related species may share identical transcriptional machineries including sigma factors and RNA polymerase, upstream regulatory regions are organized differently, so that even conserved genes can be regulated differently, and in many cases mRNA transcripts from orthologous genes can have different 5’ UTRs, which may have disparate regulatory elements in that region.

To investigate similarities and differences in 5’ UTR regions, their length and sequences were defined by 662 comparable conserved promoter regions with TSSs in both species, as shown in Figure 7D and Figure 7E. The length comparison of the 5’ UTR between *E. coli* and *K. pneumoniae* showed a strong correlation (R2 value of 0.877), and 169 (25.5%) 5’ UTR regions had exactly the same length. However, in general, the *K. pneumoniae* 5’ UTR was longer than that of E. coli, reflecting the bigger size of the genome (Figure 7D). For example, the 5’ UTR length of *rpoS*, which is one of the orthologous genes, was 566 in *E. coli*, while the length of the *K. pneumoniae rpoS* was 670. To investigate the sequence conservation between those comparable 5’ UTR regions, sequences of 5’ UTR regions from two species were aligned and percentage sequence identity for each 5’ UTR pair was calculated. The sequence variation of the 5’ UTR region along with the percentage identity of corresponding promoter and ORF is shown in Figure 7E. Consequently, ORF sequence was found to be the most conserved element, followed by sequence of promoter regions and sequence of the 5’ UTR region as the most diverse regulatory element among them. The averages of sequence identity of orthologous ORFs, comparable conserved promoters, and their 5’ UTR were 88.9%, 79.0%, and 66.0%, respectively. In order to calculate the level of conservation of the regions surrounding translation start site of orthologous genes, sequences of 200 bp long regions around translation start sites were aligned for orthologous genes having clearly aligned translation start sites between *E. coli* and *K. pneumoniae* (Figure 7F). In the 5’ UTR, there was a relatively more conserved regions 6 bp upstream of the translation start site. This region was considered to be the Shine-Dalgarno sequence of the ribosome binding site because in both species the most dominant distance between the Shine-Dalgarno sequence motif and translation start site was 6 nucleotides. In the coding region, the first codon, frequently ATG, was most conserved with slightly less conservation of the first nucleotide of the first codon. This was because the start codon, ATG, was replaced with GTG or TTG in some orthologs. In agreement with the wobble theory[56, 57], the third nucleotide of each codon was least conserved.

Interestingly, however, the second nucleotide was more conserved than the first in every codon analyzed. This might be because conservation of the second nucleotide can contribute to preserving the same amino acids like leucine, or amino acids with a similar property. Accordingly, codon analysis of the coding sequence between orthologous genes of the two species suggested that the majority of substitutions in the first nucleotide of the codon resulted in either keeping leucine or changing amino acids having similar properties, such as leucine/isoleucine, leucine/valine, valine/isoleucine, serine/threonine, glutamine/glutamic acid, or asparagine/aspartic acid.

In addition to species-specific gene content, *E. coli* and *K. pneumoniae* also exhibited differences in the organization of regulatory regions upstream of conserved orthologous genes. Different usage of TSSs and their promoter regions can contribute to varied regulation of genes downstream of those promoters, resulting in transcripts with different 5' UTR. Moreover, both species extensively use multiple TSSs, which increase the complexity and diverse nature of regulatory regions. Thus, *E. coli* and *K. pneumoniae*, which are closely related, have regulatory regions of orthologous genes organized in a different manner.

Comparison of regulatory non-coding small RNAs

The investigation of regulatory features of coding genes based on genome-wide TSSs and their comparison between two closely related enterobacteria showed that the two species share almost identical regulatory features. However, they deploy those regulatory features upstream of conserved or orthologous coding genes in a different manner, suggesting a variation of transcriptional regulation by using multiple TSSs and post-transcriptional regulation by having different 5' UTRs, generated from a different set of TSSs. Since small regulatory RNAs can function in post-transcriptional control of gene expression in many processes including stress responses, metabolic reactions, and pathogenesis[57, 58], and identification of sRNAs in *K. pneumoniae* resulted in 34 orthologous sRNA pairs between two species, we compared sequences of those conserved sRNAs and investigated whether they would regulate their target genes in the same manner. This was done, similarly in previous studies[48, 59, 60].

The conserved RNA-binding protein Hfq, first discovered in *E. coli*, is a pleiotropic regulator that modulates the stability or the translation of an increasing number of mRNAs[61, 62]. Thus, knowledge of *hfq* in *K. pneumoniae* is preliminary in terms of analyzing and comparing sRNAs between two species. Similar to *E. coli* and other *K. pneumoniae* strains[63, 64], *hfq* of *K. pneumoniae* MGH78578 (KPN_04570), existed between conserved *miaA* and *hflX* in the genome. *E. coli* K-12 MG1655 *hfq* (b4172) and *K. pneumoniae* MGH78578 *hfq* (KPN_04570) had one TSS detected upstream of *hfq* and in the coding region of *miaA*, with the genomic position of 4,397,824 and 5,000,510, respectively (Figure 8A). Similar to the high level of sequence conservation of the *hfq* ORF, sequences of promoter regions defined by experimental TSSs were perfectly conserved. 5' UTR sequences were also highly conserved; preserving sequences for the Shine-Dalgarno sequence of the ribosome binding site 6 bp upstream of translation start sites in both species (Figure 8B). This result supports the existence of a sequence of *K. pneumoniae hfq* ORF in the genome and is expressed with matching TSSs. Furthermore, sequence conservation of the promoter region and 5' UTR indicates they could be regulated in a similar way, at least during the experimented condition.

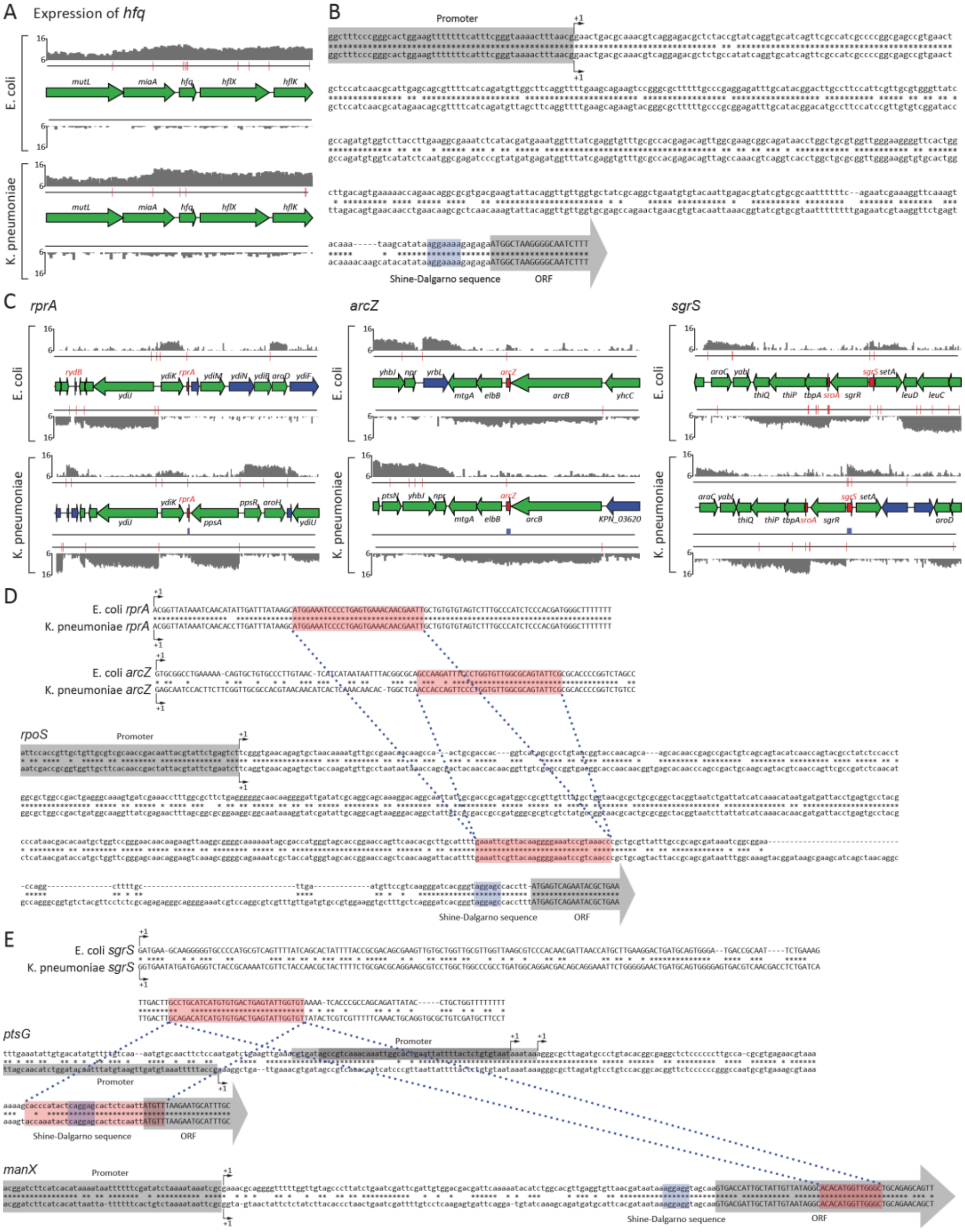


Figure 8. Comparison analysis of orthologous sRNAs. (A) Expression of RNA-binding protein *hfq* (B) Sequence conservation of regulatory region upstream of *hfq* ORF, including promoter, TSS and 5’ UTR. (C) Conservation and expression of non-coding regulatory sRNAs, *rprA*, *arcZ* and *sgrS*. (D) Sequence comparison analysis of *rprA* and *arcZ* regulating translation of *rpoS*. (E) Sequence comparison analysis of *sgrS* regulating translation of *ptsG* and *manX*.

With the occurrence of *hfq* in both species, we further investigated expression of orthologous sRNAs, compared their sequence, and analyzed possible working mechanisms in *K. pneumoniae* with prior knowledge of those sRNAs from *E. coli*. 34 orthologous sRNA candidates were confirmed to be expressed during exponential phase by TSS signals matching their 5' ends. Their expression was also supported by expression profiling data. However, those 34 expressed orthologous sRNAs showed different degrees of sequence conservation levels, ranging from 47.3% to 98.8% with an average of 83.1%. *rybB* has the most conserved sequence, whereas *sroH* has the least. Essentially, no sRNAs of *K. pneumoniae* had perfect sequence conservation compared to those of *E. coli*, which raised the question as to whether *K. pneumoniae* sRNA would work in a similar way as the *E. coli* sRNA. Thus, we compiled known target sites of *E. coli* sRNAs from the EcoCyc database[64], and mapped them onto the corresponding genomic sequence of *K. pneumoniae* (Figure 8D, Figure 8E).

3 sRNAs, *rprA*, *arcZ*, and *dsrA* were known to regulate the expression of *rpoS* by making base paring in the middle region of 5' UTR of *rpoS* mRNA with the aid of Hfq protein[65-67]. *rprA* and *arcZ* in *E. coli* target and bind to the same region of 5' UTR of *rpoS*. Thus, based on the fact that those two sRNAs are expressed in *K. pneumoniae*, if the sequence of the target site in *rpoS* mRNA and the sequence of the corresponding region of sRNA which binds to that target site are conserved, then one can hypothesize that *rprA* and *arcZ* sRNAs of *K. pneumoniae* would regulate the expression of *rpoS* in a similar manner as in *E. coli*. As expected, *rprA* and *arcZ* of *K. pneumoniae* were expressed during exponential growth with TSSs, which match the TSSs of *E. coli*'s *rprA* and *arcZ* (Figure 8C). Furthermore, regions that bind to the target site of *rpoS* were also conserved (Figure 8D). Analogously, *rpoS* of *E. coli* and *K. pneumoniae* was expressed with TSS at 2,866,139 in *E. coli* and 3,401,901 in *K. pneumoniae*, and the sequence of the promoter region of *rpoS* was highly conserved between

the two species, although the 5’ UTR defined by those TSSs showed significantly different lengths with a long nucleotide addition in the 5’ UTR of *K. pneumoniae rpoS*. However, the sequence of the target site of *rprA* and *arcZ* was almost perfectly conserved with one nucleotide replacement (Figure 8D). Considering that *rpoS* was expressed from conserved promoters in both species, and *rprA* and *arcZ* targeting the conserved regions of the 5’ UTR of *rpoS* transcript were also conserved and expressed, it is quite likely those sRNAs in *K. pneumoniae* regulate the expression of *rpoS* in the same manner as in *E. coli*.

Furthering the analysis of *rpoS*-targeting *rprA* and *arcZ*, another sRNA *sgrS* was similarly analyzed. In *E. coli*, *sgrS* was shown to regulate expression of two metabolic transporters, *ptsG* and *manX*, by base-pairing dependent manner[68, 69]. Although *sgrS* sRNA of *E. coli* and *K. pneumoniae* was expressed during exponential phase (Figure 8C), the sequence conservation was quite low at 56.3%. *ptsG* and *manX,* which are targeted and regulated by *sgrS* in *E. coli,* were also expressed in *K. pneumoniae*. However, *E. coli* and *K. pneumoniae ptsG* was expressed by different promoters, resulting in *ptsG* transcripts with different 5’ UTR, while *manX* was expressed by a conserved promoter with matching TSSs. Although the overall conservation level of *sgrS* between *E. coli* and *K. pneumoniae* was quite low, the region, which is known to bind to *ptsG* and *manX* in *E. coli,* was highly conserved (Figure 8E). Besides its target sites in *ptsG* and *manX* transcripts were also highly conserved, which suggests *sgrS* would regulate the expression of *ptsG* and *manX* in a similar way as in *E. coli*.

Comparisons of sRNAs and their working mechanisms should be performed not only with just the sequence of sRNAs and their target sites, but also with the working context, including expression of those sRNAs, transcripts of genes containing target sites, and occurrence of Hfq, if the sRNA requires the protein. In depth comparisons of sRNAs and their working context suggest that many of the orthologous sRNAs identified with the TSS dataset of this study could work in a similar way as *in E. coli* since target sequences of those sRNAs and sequences of sRNAs known to bind to the target sites are conserved and exist in the primary transcript of target genes under the given condition. Moreover, high conservation of regions that bind to the target sites despite poor conservation of whole sRNA sequences suggests that sequence comparison and conservation can determine which region may be more important in terms of regulation and their working context.

Disucssion

E. coli K-12 MG1655 is an extensively studied laboratory strain with a wealth of genome-wide studies. As such, genome-wide TSS determination of the E. coli genome with single base pair resolution by deep sequencing has been performed by a number of studies[9-11]. However, *K. pneumoniae* has only recently been studied using genome-wide approaches[34]. A number of TSSs have been reported and investigated with specific focus on genes involved mostly in virulence[70-77] and nitrogen metabolism[78-88] in other strains of *K. pneumoniae*. In addition to previously known TSSs of *K. pneumoniae*, this study extended the scope of knowledge by adding over 3,000 experimental TSSs for that species and by performing an in-depth look at the intergenic region of the *K. pneumoniae* genome.

Regulatory features discussed herein with *E. coli* and *K. pneumoniae* are not limited to the class of gammaproteobacteria. *G. sulfurreducens* in deltaproteobacteria[12], *H. pylori* in epsilonproteobacteria[26], *C. crescentus* in alphaproteobacteria[8] and methanogenic archaea *Methanosarcina mazei*[13] were also shown to have a significant amount of multiple TSS usage. Similarly, a mammalian promoter was also reported to have multiple TSSs[89, 90]. Thus, extensive use of multiple TSSs is a common strategy in a wide range of living organisms, exploiting alternative transcripts and providing complexity in gene expression and regulation. The level of multiple TSS usage differs by organism, however. *E. coli*, a generalist which can adjust and live in a wider range of environments, showed extensive usage of multiple TSSs, suggesting more complicated transcription regulation. On the other hand, *G. sulfurreducens* or *M. mazei*, a specialist that thrives in a more specific niche, showed lesser multiple TSSs. A significant fraction of operons had multiple TSSs in both *E. coli* and *K. pneumoniae* and encoded genes with essential functions, e.g., genes involved in amino acid biosynthesis, central metabolism, and transport, similar to *G. sulfurreducens*[12]. In addition to the usage of multiple TSSs, bacterial strains in different classes of proteobacteria show a similar distribution of 5' UTR length. Like *E. coli* and *K. pneumoniae*, the preferred length of the 5' UTR of *H. pylori* and *G. sulfurreducens* is 20-40 nucleotides in length. A distinctive regulatory function of the 5' UTR was reported in yeast[91], however no correlation between the 5' UTR length and function was found in *E. coli* or *K. pneumoniae*. However, when we compared the distribution of the 5' UTR length between *E. coli* and *K. pneumoniae* belonging to the same COG functional group, most of the COG groups showed similar preferences for 5' UTR length. Only the "Transcription" group showed significant differences (p-value of Wilcoxon rank sum test was 1.76x10-5).

Multiple promoters upstream of a gene can be regulated by transcription factors in different ways. For example, *rpoD* of *E. coli* which encodes sigma factor σ^{70} has multiple TSSs, and each promoter is recognized by σ^{70}, σ^{32}, or σ^{24}, enabling expression of *rpoD* under conditions including exponential growth, heat shock, or other stresses[92-94]. Other transcription factors also contribute to the increasing complexity of promoter region structure. Transcription of an essential cell division protein operon of *E. coli*, *ftsQAZ*, is under the control of the two core promoters with two TSSs which are separated by 125 bp. Binding of the quorum sensing regulator, SdiA, activates the distal core promoter while it represses the proximal one[95]. *K. pneumoniae* also has that conserved operon *ftsQAZ*, and the TSSs of that particular operon were observed in both species and a similar regulation on expression of the *ftsQAZ* operon may be happening in *K. pneumoniae*. Another example is the *ure* operon (*ureDABCEFG*) in *K. pneumoniae*, which has two core promoters with distinct TSSs. One core promoter is NAC (nitrogen assimilation control protein) dependent, and the other is not[80]. For this operon, one TSS at the genomic position of 3,790,095 was identified during the mid-exponential growth in *K. pneumoniae* from this study. Thus, two closely related species in gammaproteobacteria, *E. coli* and *K. pneumoniae*, showed extensive usage of multiple TSSs, however they exhibited diverse organization of the regulatory region with different sets of promoters and associated TSSs. In addition to the presence of species-specific genes, this usage of multiple TSSs could potentially confer divergent regulation of orthologous genes, thereby contributing to phenotypic differences between two closely related species

Another advantage of having multiple TSSs is a transcript from each TSS has the different 5’ UTR upstream of coding region. Comparative analysis of 5’ UTRs between two species may provide insight into understanding similar or different roles of the 5’ UTR in the regulation of gene expression. One good example is a comparison of orthologous sRNAs and their binding onto the 5’ UTR region of target genes. Many orthologous sRNAs, including *rprA*, *arcZ*, and *sgrS*, showed enough evidence to postulate that their regulatory mechanism by the base pair dependent manner proven in *E. coli* may work similarly in *K. pneumoniae*. This conclusion is further supported by phylogenetic analysis of sRNA evolution in *E. coli* and *Shigella* genomes[96]. In the previous study, it is claimed that core or conserved sRNAs are more tightly integrated into cellular genetic regulatory networks, and over 80% of genes targeted by Hfq-associated core sRNAs have been transferred intact. 90% of orthologous sRNAs identified in our study were also categorized as core or conserved sRNAs in the previous study[96], supporting conserved regulatory mechanisms of those orthologous sRNAs.

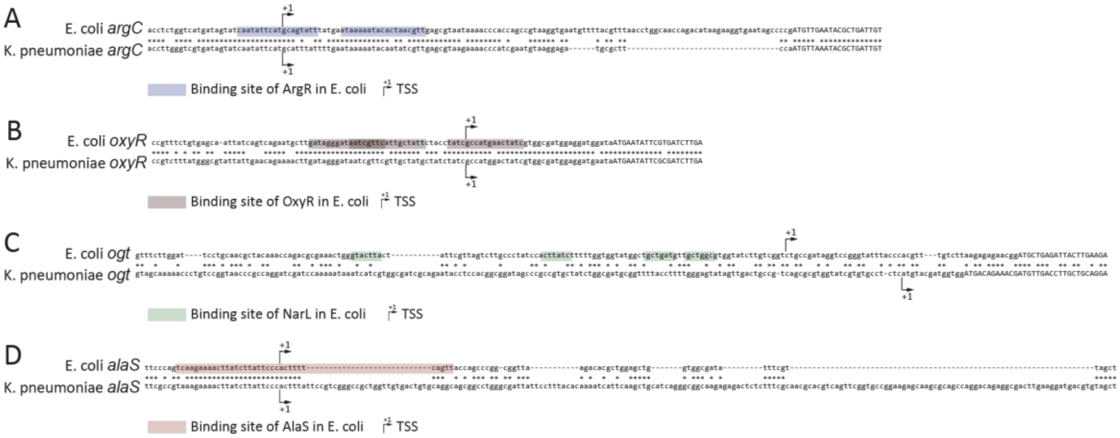


Figure 9. Comparison analysis of transcription factor binding sites upstream of orthologous genes. (A) ArgR binding on *argC* (B) OxyR binding on *oxyR* (C) NarL binding on *ogt* (D) AlaS binding on *alaS*

An interesting additional attribute of the 5' UTR sequence is that it can potentially serve as a transcription factor binding site, thereby contributing to the transcriptional regulation of a downstream gene or operon. For example, the ArgR (b3237) transcription regulator is known to bind to the promoter region and 5’ UTR of *argC* in *E. coli*[97, 98] (Figure 9A). The conserved promoter regions and TSSs were identified from the analysis of this study, and the sequence alignment of the upstream regulatory region and ORF suggests the binding regions of ArgR upstream of *argC* were highly conserved. Similarly, OxyR (b3961) transcription regulator, which auto-regulates its transcript by binding the promoter region and upstream regulatory region[99], also had conserved promoter regions and 5’ UTR defined by experimental TSSs identified in this study (Figure 9B). The binding regions of OxyR upstream of *oxyR* were also highly conserved. Considering those two transcription regulators, ArgR and OxyR, had a high level of amino acid sequence similarity of 94.2% and 96.1% respectively, and their target genes were expressed with conserved promoters and matching TSSs, it is likely that *K. pneumoniae*, ArgR, and OxyR may regulate *argC* and *oxyR* in a similar manner as in *E. coli*. However, unlike ArgR and OxyR regulation, NarL and AlaS transcription regulators showed opposite tendencies. In *E. coli*, NarL (b1221) regulates *ogt* by binding the upstream region of *ogt*[100] (Figure 9C). AlaS (b2697) of *E. coli* auto-regulates transcription of *alaS* by binding to the region covering parts of promoter region and the 5' UTR[101] (Figure 9D). A high conservation level of those transcription factors (94.9% amino acid sequence similarity for NarL and 91.1% for AlaS) suggests the sequence motifs of their binding sites would be similar between *E. coli* and *K. pneumoniae*. However, sequence alignments of upstream regulatory regions of *ogt* and *alaS* between two species showed NarL and AlaS of *K. pneumoniae* may not regulate *ogt* and *alaS* by binding the same regions of binding sites. Thus, comparative analysis of upstream regions including the promoter region and the 5' UTR of two closely related species can hint that the possibility of regulatory mechanisms of a lesser-studied microorganism, *K. pneumoniae*, by transferring ample knowledge from a well-studied microorganism such as *E. coli*.

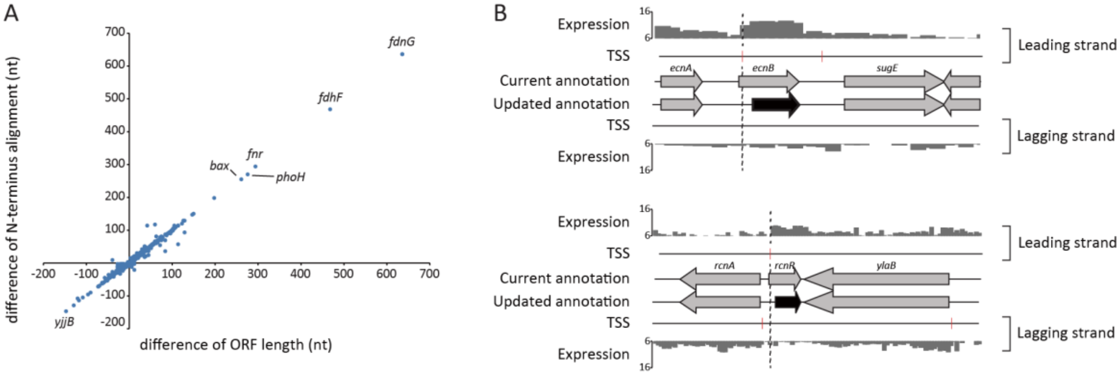


Figure 10. Comparative analysis on possible misannotation in the current annotation of *K. pneumoniae.* (A) Comparison of the length difference between the whole orthologous ORFs and their N-terminus alignments. (B) Updated annotation of *ecnB* and *rcnR* by TSS

Analysis of upstream regulatory regions performed in this study was based on the assumption that the current gene annotation of analyzed species is correct. Unlike the current gene annotation of *E. coli*, which is the well-studied microorganism, the current gene annotation of *K. pneumoniae* was built by the computational methods and has not been fully confirmed with proteomic data, leaving the possibility of incorrect annotation of protein coding genes. Analyzing the sequence of coding regions and upstream regions of orthologous coding genes by sequence alignment suggested that many *K. pneumoniae* orthologous genes were longer than *E. coli* orthologous genes. The longer length of *K. pneumoniae* genes was mostly due to the fact that the current annotation of those genes had longer sequences at the N-terminus side (Figure 10A). When the genomic position of translation start sites were changed based on sequence alignment analysis of the coding region and upstream flanking region, 8 TSSs were found upstream of those changed translation start sites of *K. pneumoniae* orthologous coding genes (Figure 10B).

TSS profiling through high-throughput sequencing techniques provides a comprehensive source of experimentally derived information related to the initiation of transcription. Annotation of non-coding regions in bacterial genomes, including the promoter regions and untranslated regions of transcripts, allows for the comparison of regulatory elements of transcription and translation between closely related species, and the identification of a spectrum from highly conserved to diverse regulatory elements. Direct comparison between cross species of bacteria also assists in transferring regulatory information of lesser-studied bacterial species and significantly improves annotation of regulatory regions. Thus, the comparative approach of this study provides a starting point for the determination of conserved and specific features of the transcriptional output of closely related bacteria at single nucleotide resolution.

Methods

Bacterial Strains, media, and growth conditions: *Escherichia coli* K-12 MG1655 and *Klebsiella pneumoniae* subsp. pneumoniae MGH78578 were grown in glucose (2 g/L) minimal M9 medium containing 2 ml/L 1M MgSO4, 50 l/L 1M CaCl2, 12.8 g/L Na2HPO4.7H2O, 3 g/L KH2PO4, 0.5 g/L NaCl, 1 g/L NH4Cl and 1 ml trace element solution

(100X) containing 1 g EDTA, 29 mg ZnSO4.7H2O, 198 mg MnCl2.4H2O, 254 mg CoCl2.6H2O, 13.4 mg CuCl2, and 147 mg CaCl2. Glycerol stocks of the *E. coli* and *K. pneumoniae* strains were inoculated into the minimal medium supplemented with glucose and cultured at 37 ℃ with constant agitation overnight. The cultures were diluted 1:100 into 50 mL of the fresh minimal medium and then cultured at 37 ℃ to an appropriate cell density.

Total RNA isolation: Three milliliters of cells from mid-log (OD=0.6) phase culture were mixed with 6 ml RNAprotect Bacteria Reagent (Qiagen). Samples were mixed immediately by vortexing for 5 seconds, incubated for 5 minutes at room temperature, and then centrifuged at 5000 g for 10 minutes. The supernatant was decanted and any residual supernatant was removed by inverting the tube once onto a paper towel. Total RNA samples were then isolated using RNeasy Plus Mini kit (Qiagen) in accordance with the manufacturer's instruction. Samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific) and quality of the isolated RNA was checked by visualization on agarose gels and by measuring the sample's A260/A280 ratio (≥1.8).

Modified 5'RACE for 5'tri-phosphorylated mRNA profiling: TSS determination protocol previously described[12] was adapted for the bacteria strains in the current study. To enrich intact 5' tri-phosphorylated mRNAs from the total RNA, 5' mono-phosphorylated ribosomal RNA (rRNA) and any degraded mRNA were removed by treatment with a Terminator 5'-Phosphate Dependent Exonuclease (Epicentre) at 30oC for 1 hr. The reaction mixture consisted of 10 μg purified total RNA, 1 μL terminator exonuclease, reaction buffer and RNase-free water up to total 20 μL. The reaction was terminated by adding 1 μL of 100 mM EDTA (pH 8.0). Intact tri-phosphorylated RNAs were precipitated by adding 1/10 volume of 3 M sodium acetate (pH 5.2), 3 volume of ethanol and 2 μL of 20 mg/mL glycogen. RNA was precipitated at -80oC for 20 min and pelleted, washed with 70% ethanol, dried in Speed-Vac for 7 minutes without heat and resuspended in 20 µL nuclease free water. The tri-phosphorylated RNA was then treated with RNA 5’-Polyphosphatase (Epicentre) to generate 5’-end mono-phosphorylated RNA for ligation to adaptors. The RNA sample from the previous step was mixed with 2 µL 10 reaction buffer, 0.5 µL SUPERase-In (Ambion), 1 µL RNA 5’-Polyphosphatase and RNase-free water up to 20 µL. The mixture was incubated at 37oC for 30 minutes and reaction was stopped by phenol-chloroform extraction. Ethanol precipitation was carried out for isolating the RNA as described above. To ligate 5′ small RNA adaptor (5’ GUUCAGAGUUCUACAGUCCGACGAUC 3’) to the 5′-end of the mono-phosphorylated RNA, the enriched RNA samples were incubated with 100 µM of the adaptor and 2.5 U of T4 RNA ligase (New England Biolabs). cDNAs were synthesized using the adaptor-ligated mRNAs as template using a modified small RNA RT primer from Illumina (5’ CAAGCAGAAGACGGCATACGANNNNNNNNN 3′) and Superscript II Reverse Transcriptase (Invitrogen). The RNA was mixed with 25 µM modified small RNA RT primer and incubated at 70oC for 10 min and then at 25oC for 10 min. Reverse transcription was carried out at 25oC for 10 min, 37oC for 60 min, 42oC for 60 min and followed by incubation at 70oC for 10 min. A reaction mixture for reverse transcription consisted of the following components: 5 1st strand buffer; 0.01 M DTT; 10 mM dNTP mix; 30 U SUPERase•In™ (Ambion); and 1500 U SuperScript™ II (Invitrogen). After the reaction, RNA was hydrolysed by adding 20 µL of 1 N NaOH and incubation at 65oC for 30 min. The reaction mixture was neutralized by adding 20 µL of 1 N HCl. The cDNA samples were amplified using a mixture of 1 µL of the cDNA, 10 µL of Phusion HF buffer (NEB), 1 µL of dNTPs (10 mM), 1 µL SYBR green (Qiagen), 0.5 µL of HotStart Phusion (NEB), and 5 pmole of small RNA PCR primer mix. The amplification primers used were 5’AATGATACGGCGACCACCGACAGGTTCAGAGTTCTAC AGTCCGA3’ and 5’

CAAGCAGAAGACGGCATACGA 3’. The PCR mixture was denatured at 98 ℃ for 30 s and cycled to 98 ℃ for 10 s, 57 ℃ for 20 s and 72 ℃ for 20 s. Amplification was monitored by a LightCycler (Bio-Rad) and stopped at the beginning of the saturation point. Amplified DNA was run on a 6% TBE gel (Invitrogen) by electrophoresis and DNA of size ranging from 100 to 300 bp were size fractionated. Gel slices were dissolved in two volumes of EB buffer (Qiagen) and 1/10 volume of 3 M sodium acetate (pH 5.2). The amplified DNA was ethanol-precipitated and resuspended in 15 μL DNase-free water (USB). The final samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific).

Sequencing, data processing and mapping: The amplified cDNA libraries from two biological replicates for each *E. coli* and *K. pneumoniae* were sequenced on an Illumina Genome Analyzer. Sequence reads for cDNA libraries for *E. coli* and *K. pneumoniae* were aligned onto the *E. coli* K-12 MG1655 genome (NC_000913) and *K. pneumoniae* subsp. pneumoniae MGH78578 genome with 5 plasmids (NC_009648, NC_009649, NC_009650, NC_009651, NC_009652, NC_009653), respectively, using Mosaik (http://code.google.com/p/mosaik-aligner) with the following arguments: hash size = 10, mismatach = 0, and alignment candidate threshold = 30 bp. Only reads that aligned to unique genomic location were retained. Two biological replicates were processed separately, and only sequence reads presented in both biological replicates were considered for further process. The genome coordinates of the 5’-end of these uniquely aligned reads were defined as potential TSSs. Among potential TSSs, only TSSs with the strongest signal within 10 bp window were kept to remove possible noise signals, and. TSSs with greater than or equal to 50% of the strongest signal upstream of an annotated gene were considered as multiple TSSs.

Transcriptome analysis: Transcriptome dataset with oligonucleotide tiling microarrays for E. coli grown in glucose minimal media to the mid-exponential phase was taken from the previous study[11]. In order to get the transcriptome dataset for *K. pneumoniae*, the protocol previously described[12] was adapted for the *K. pneumoniae* in the current study. Briefly, 10 g of purified total RNA sample was reverse transcribed to cDNA with amino-allyl dUTP. The amino-allyl labeled cDNA samples were then coupled with Cy3 Monoreactive dyes (Amersham). Cy3 labeled cDNAs were fragmented to 50 ~ 300 bp range with DNase I (Epicentre). High-density oligonucleotide tiling arrays consisting of 379,528 50-mer probes spaced 30 bp apart across the whole *K. pneumoniae* genome and 5 plasmids were used (Roche Nimblegen). Hybridization, wash and scan were performed in accordance with manufacturer's instruction. Two biological replicates were utilized for mid exponential growth under glucose minimal media. Probe level data were normalized with RMA (Robust Multiarray Analysis) algorithm[102] without background correction, as implemented in NimbleScan 2.4 software.

Defining orthologous ORF in E. coli and K. pneumonia: Genome annotation for *E. coli* K-12 MG1655 and *K. pneumoniae* MGH 78578 were obtained from the NCBI Genome Database. The refseq ID of *E. coli* K-12 MG1655 is NC_000913, and the refseq IDs of *K. pneumoniae* MGH 78578 genome and 5 plasmids are NC_009648, NC_009649, NC_009650, NC_009651, NC_009652 and NC_009653. In order to define orthologous ORFs between *E. coli* and *K. pneumoniae*, we performed reciprocal alignment for exhaustive pairs of amino acid sequences of ORFs in both strains, by using ClustalW2 software[103]. From the reciprocal alignment, we calculated percentage identity and percentage aligned scores, and used 50% as a cutoff for both percentage identity and percentage aligned scores of amino acid sequence alignment.

Data processing, visualization and availability: Graphs representing the number of uniquely mapped reads per nucleotide were stored in GFF (generalized feature format) format files and visualized using MetaScope (http://sbrg.ucsd.edu/Downloads/MetaScope) and SignalMap software from Nimblegen (http://www.nimblegen.com/products/software/). Motif logos were calculated and drawn by MEME[52] and Venn diagrams and histograms were prepared by Microsoft Excel software. Experimental data were formatted to GFF format and visualized in MetaScope. The raw TSS reads for *E. coli* and *K. pneumoniae* and expression profiling dataset for *K. pneumoniae* have been deposited in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/), GSE35822. Processed experimental data in this study are available at http://www.sbrg.ucsd.edu.

Identification of potential sRNA: Potential sRNAs in *K. pneumoniae* were predicted by two methods. The first method is sRNA sequence search in the target region bound by neighboring orthologous genes. With the list of orthologous genes between *E. coli* and *K. pneumoniae*, closest orthologous genes neighboring each sRNA in *E. coli* were searched. Then, target region in *K. pneumoniae* genome was decided by neighboring orthologous genes. The sequence of *E. coli* sRNA was used to search conserved sequence in the target region in *K. pneumoniae* genome. For example, *E. coli glmY* is surrounded by *glrK* (b2556) and *purL* (b2557) orthologous genes. Thus, the target region in *K. pneumoniae* was chosen with the boundaries by *glrK* (KPN_02881) and *purl* (KPN_02882). Then, the sequence of *E. coli glmY* was searched in the target region, by sequence alignment between *glmY* and the target region with ClustalW2. This approach resulted in 48 putative sRNA candidates. The potential sRNA candidates were supplemented with prediction with Infernal[50] (http://infernal.janelia.org). Rfam database 10.1 was used as model for sRNA prediction. Hits with E-value less than 10^{-5} were mapped to TSS dataset previously identified, and hits with the 5’ end matching to experimental TSSs were considered as potential sRNAs. The Infernal and Rfam approach resulted in 41 sRNA candidates. In sum, a total 50 number of sRNA candidates were prediction in a combination of two approaches.

Categorization of promoter region based on conservation and presence of TSS: Each TSS experimentally identified was considered to be associated with one promoter region, so 50 bp long genomic region directly upstream of TSS was defined as promoter region. With the list of orthologous genes between *E. coli* and *K. pneumoniae*, TSS and its associated promoter region was categorized as species-specific promoter region (SSP), if that TSS was not assigned to any of orthologous genes. TSSs assigned to orthologous genes were categorized as one of three groups: conserved promoter region with TSS (CPT), conserved promoter region with no matching TSS (CPNT) or orphan promoter region (OP). For each orthologous gene, all TSSs assigned to that gene in both species were used to define promoter regions. Then the sequence of each promoter region of one species was aligned onto the sequence of 800 bp long genomic region upstream of the orthologous gene in the other species, in order to see the 3' end of the alignment match with any TSS of the other species with 2 bp tolerance. If there is a matching TSS, the promoter region of that TSS was aligned again back onto the 800 bp upstream region of the first species, and if the 3' end of the second alignment matched with the first TSS, then those two TSS in both species were categorized as CPT. If the 3' end of the first alignment didn't match any TSS of the other species, then the sequence of alignment of the other species was aligned back onto the upstream region of the first species. If the 3' end of the second alignment matched with the first TSS, then the promoter region defined by that TSS was categorized as CPNT. If the 3' end of the second alignment did not match with the first TSS, then the promoter region was categorized as OP.

Acknowledgements

Chapter 3, in total, is a reprint of the material as it appears in Kim, D.*, Hong, J.S.*, Qiu, Y., Nagarajan, H., Seo, J.H., Cho, B.K., Tsai, S.F., Palsson, B.Ø. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide

transcription start site profiling. PLoS Genetics, 8(8):e1002867 (2012). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 4: Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states

The RNAP core enzyme (E) for bacterial transcription is a catalytic multi-subunit complex ($\alpha_2\beta\beta'\omega$) capable of transcribing portions of the DNA template into RNA transcripts. At the beginning of the transcribing process, the E requires a σ-factor to recognize the genomic location where the process initiates [104-106] (Figure 11a). σ-factor, a single dissociable subunit, binds to the E, forming a holoenzyme ($E\sigma^x$, x for each σ-factor) and orchestrates the promoter-specific transcription initiation [104]. To date, one housekeeping σ-factor σ^{70} (*rpoD*) and six alternative σ-factors σ^{54}, σ^{38}, σ^{32}, σ^{28}, σ^{24}, and σ^{19} (*rpoN, rpoS, rpoH, fliA, rpoE,* and *fecI*, respectively) have been described in *E. coli*. Although the importance of σ-factors and their role in the function of the RNAP and bacterial transcription are well known, we do not yet have a genome-wide understanding of the network of regulatory interactions that the σ-factors comprise in any species. With systems biology and genome-scale science emerging and describing the phenotypic functions of bacteria, it is now possible to comprehensively elucidate the structure of the σ-factor network. Here, we present the results from a systems approach that integrates multiple genome-scale measurements to reconstruct the regulatory network of σ-factor-gene interactions in *E. coli*. This reconstruction is provided here as a resource for the community.

Determination of the genome-wide map of holoenzyme binding

To capture the first step of transcription cycle, which is the formation of the $E\sigma^x$-promoter complex, we obtained genome-wide location profiles and integrated the identified RNAP and σ-factor binding sites, leading to a reconstruction of a genome-scale Eσ-binding region map (Figure 11b). A genome-wide static map of entire $E\sigma^x$-binding sites ($E\sigma^x$-map) was determined by employing chromatin immunoprecipitation coupled with microarrays (ChIP-chip) of rifampicin-treated cells (Figure 11c), revealing the active promoter regions *in vivo* across the *E. coli* genome[107, 108] (see Methods). A total of 2,129 Eσ^x-binding regions were identified, consisting of 727 (34.1%) for leading strand, 755 (35.5%) for lagging strand, and 647 (30.4%) for both strands (i.e., divergent promoter regions) (Figure 12).

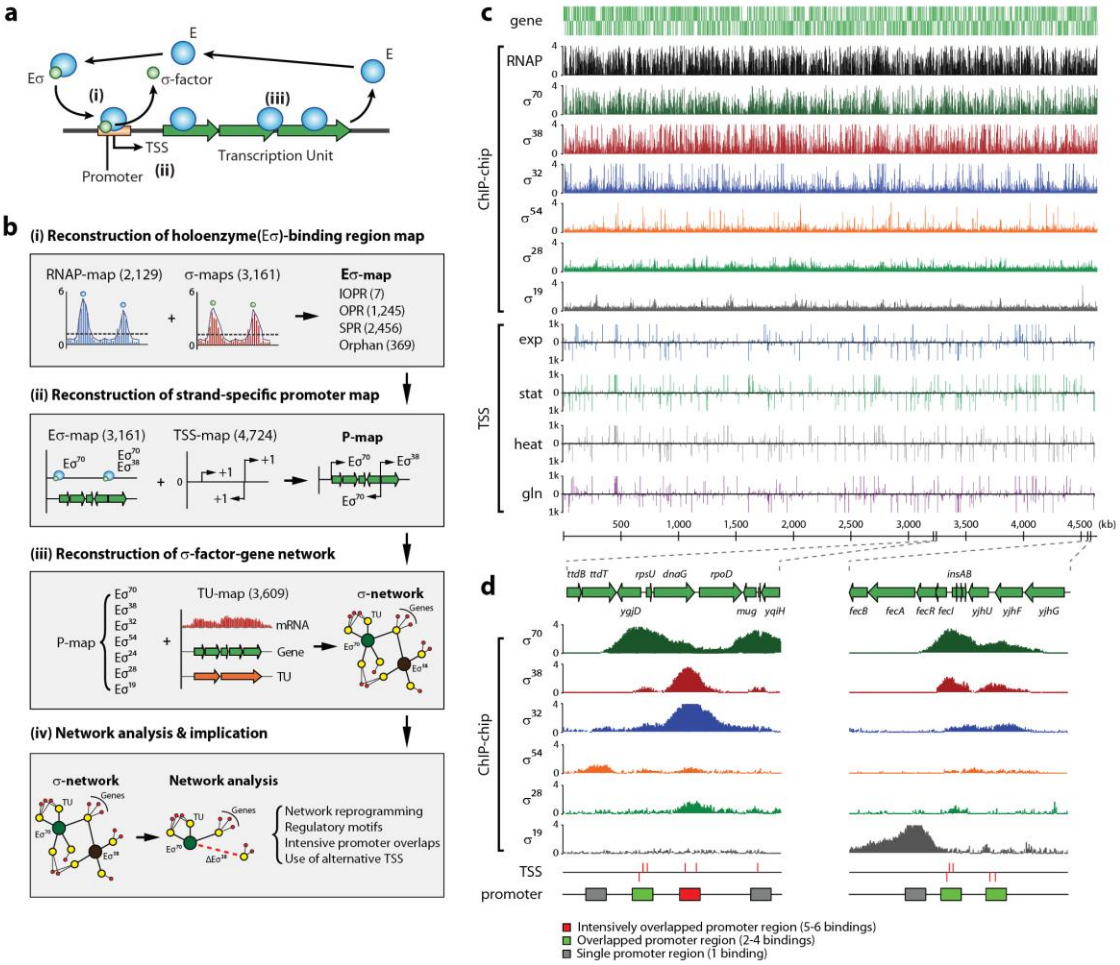


Figure 11. Molecular basis of transcription and a reconstruction of σ-TUG network from multi-omic experimental datasets. (a) A diagram shows bacterial transcription process by a RNAP core enzyme and an associated σ-factor. (b) Four-step process of multi-omic data integration to reconstruct σ-TUG network. (c) Datasets used for σ-TUG network reconstruction: ChIP-chip dataset with RNAP and 6 σ-factors and TSS dataset. (d) Zoomed-in examples of *rpoD* , *fecI* and *fecRAB.*

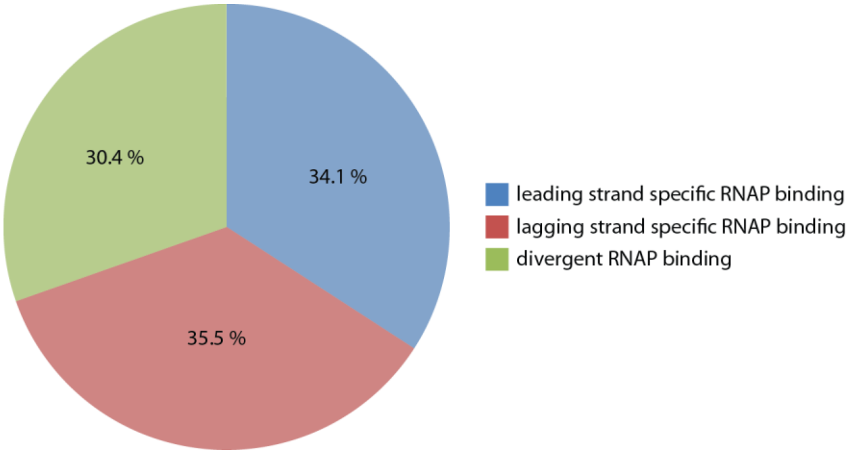


Figure 12. Strand specificity of RNAP bindings.

While the construction of the Eσ^x-map is informative, is not sufficient to give the σ-specific Eσ-binding map where the promoter-specific role of the σ-factor is detailed[109]. We thus deployed ChIP-chip assays for the direct identification of locations of σ-factor binding across the *E. coli* genome. We analyzed *E. coli* cells grown to mid-logarithmic phase or to stationary phase under multiple growth conditions. Using data from biological duplicate or triplicate experiments for each σ-factor ChIP-chip (36 experiments in total), we identified 1,643 targets for σ^{70}, 903 targets for σ^{38}, 312 targets for σ^{32}, 180 targets for σ^{54}, 51 targets for σ^{28}, and 7 targets for σ^{19} (Figure 11c, Figure 13a). We were not able to get dataset for σ^{24}, and the missing dataset was supplemented by incorporating 65 σ^{24} promoter regions from RegulonDB[109]. For validation, we compared the σ-factor binding regions with the previously reported promoters regulated by each σ-factor [109] (Figure 11d). Overall, we identified 86% of previously reported binding sites and 2,465 new σ-factor binding regions, extending our current knowledge by over 300%.

By integrating the entire Eσ^x and σ-factor binding regions, we obtained the genome-wide Eσ-binding region map (Eσ-map) comprising 3,161 binding regions. Next, each Eσ-binding site was classified into one of three categories depending on the number of σ-factors recruited to that site: single Eσ-binding promoter region (SPR), overlapped Eσ-binding promoter region (OPR), and intensively overlapped Eσ-binding promoter region (IOPR) (Figure 11b, Figure 11d, and Figure 13b). For instance, all σ-factors except σ^{19} were detected at the promoter region of the *rpoD* gene, which encodes σ^{70}, however only σ^{19} was found to bind to the promoter region of the *fecABCDE* operon which encodes the ferric citrate outer membrane receptor and the ferric citrate ABC transporter (Figure 11d). Over 48% of Eσ-binding regions identified in this study were overlapped or extensively overlapped binding regions, indicating that Eσ-switching, or binding of alternative Eσ, at the same promoter region may be needed to ensure continued gene expression in response to environmental changes [105] (Figure 13a).

Determination of the genome-wide promoter map

69% of the Eσ-binding regions exhibited strand specificity, with the balance being observed as divergent promoter regions. Although the assignment of the RNAP-binding regions to each strand was achievable using the expression profiles [11], it is difficult to directly assign σ-factors to the promoter regions because information on the *cis*-acting sequence elements, such as -10 and -35 boxes in the promoter regions, is not yet fully elucidated for each σ-factor. To identify the promoter elements more precisely with strand specificity and a better resolution than ChIP-chip, we performed transcription start site (TSS) profiling at the genome scale with a single nucleotide resolution. A genome-wide TSS-map was generated from TSS profiling by the rapid amplification of cDNA ends followed by deep-sequencing after 5’ triphosphate enrichment[7, 12, 26] for 3 conditions: stationary phase, heat-shock, and alternative nitrogen source with glutamine. TSS profiling for exponential phase was taken from the previous study[7], and it was processed together with the other three datasets. TSS-map was then integrated with the Eσ-map to build a strand-specific promoter map (P-map) (Figure 11b, Figure 11c, Figure 11d).

Reconstruction of sigma factor regulons and their overlaps

The P-map was combined with the transcription unit (TU) map [11], resulting in the σ-factor-TU-gene network (σ-TUG network) (Figure 13d, Figure 13e). A network of interactions among the σ-factors was extracted from the σ-TUG network (Figure 13c). σ^{70} and σ^{24} are the only σ-factors that auto-regulate themselves, and σ^{70} and σ^{38} regulate most of the other σ-factors, reflecting their roles as housekeeping σ-factors under exponential and stationary phase [104]. Gene essentiality data is available for *E. coli* [110], and only *rpoD* has been found to be an essential σ-factor. This network feature is consistent with the fact that σ^{70} regulates the highest number of σ-factors, including itself. In addition, σ^{70} has the biggest regulon that cannot be substituted for by the other σ-factors (Figure 13d).

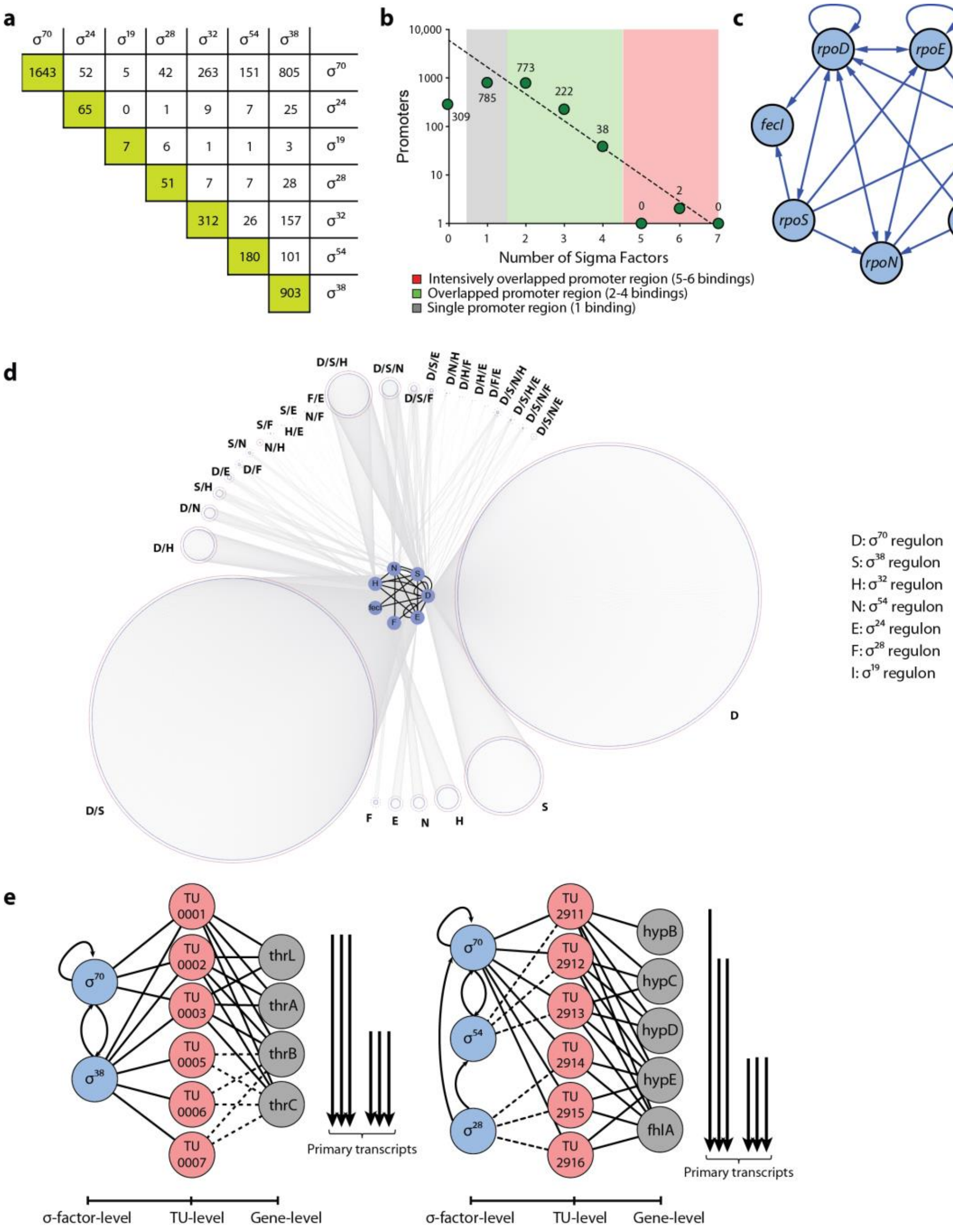


Figure 13. Properties of the reconstructed σ-factor network in *E. coli.* (a) Extensive overlapping between σ-factor binding sites. (b) Number of promoters bound by multiple σ-factors shows complex overlap between different σ-factors, indicating complicated alternative σ-factor usage. (c) A regulatory network between σ-factors in E. coli, where σ^{70} and σ^{38} regulate expression of most of 7 σ-factors. (d) Reconstruction of 3-layered network of σ-factors, transcription units, and genes. (e) Examples of *thrLABC* and *hypBCDE-fhlA* transcription units that are differently regulated by multiple σ-factors, and result in different TUs containing different sets of genes.

The significant overlap of σ-factor regulons leads to fundamental questions: what is the molecular basis for the overlap and what are the consequences of having a complicated σ-factor network? Due to the individual ability of each σ-factor to recognize *cis*-acting sequence elements in the promoter region (such as -10 box or -35 box) the sequence motifs of promoter regions were analyzed. Like previous studies[111-113], the sequence motifs of σ^{70} and σ^{38} showed a similar -10 box sequence (TAtaaT and CTAtacT), however, unlike the σ^{70} sequence motif, the σ^{38} did not have a distinctive -35 box. The similarity in the -10 box sequence motifs of the σ^{70}- and σ^{38}-specific promoters and the degenerate nature of the -35 box sequence of the σ^{38}-specific promoters explains, in part, how a large overlap between σ^{70} and σ^{38} regulons is possible.

With the structure and molecular details of the σ-TUG network in hand, we can begin to study its functional states. Due to the limited number of E complexes in a growing *E. coli* cell [104], each σ-factor should compete to achieve association with an E complex to initiate transcription. Thus, it becomes important which $E\sigma^{x}$ binds to and how frequently [114]. We find that the promoter sets specific to each σ-factor overlap extensively, and a large number of promoters bound by multiple σ-factor share the same TSS (Figure 13a, Figure 13d). These findings raise a question about the molecular mechanism of σ-factor competition for binding to E complex and subsequently to the promoter, and how that affects the transcription initiation.

Sigma factor competition in overlapped promoters

σ-factors are believed to act predominantly as a positive effector, since they recognize the *cis*-acting elements in promoters that enable the $E\sigma^{x}$ to bind. Interestingly, however, σ^{38} does have a negative effect on expression level of some genes, even though it acts mainly as a positive effector[115, 116]. To shed light on the molecular mechanisms of σ-factor competition by

σ^{38}, we performed ChIP-chip experiments for RpoB with WT and its isogenic *rpoS* knock-out strain to obtain differential Eσ^{x} binding to the genome. The differential binding intensity of the Eσ^{x} to the promoters of 1139 genes, whose transcription is directly affected by σ^{38}, is shown in Figure 14a. If σ^{38}-specific promoters were bound only by σ^{38}, then the E complex recruited onto those promoters would be very scarce. However, the majority of σ^{38}-specific promoters showed significant levels of signals for Eσ^{x} binding in the σ^{38} deletion strain, indicating recruitment of the Eσ^{x}, implying rescue of transcription activity (Figure 14a). To confirm that the detected binding of the Eσ^{x} leads to transcription, we performed expression profiling with WT and *rpoS* knock-out strain cells under stationary phase conditions (Figure 14b). Most genes having σ^{38}-specific promoters were expressed. Among 1139 genes with σ^{38}-specific promoters, 178 genes (16%) showed up-regulated expression when *rpoS* was removed and expression of 291 genes (26%) was down regulated more than 2-fold (t-test p-value ≤ 0.05). The remaining 58% of genes showed no statistical significance in expression (fold change less than 2) or were not expressed in either strain. In the absence of *rpoS,* σ^{38}-specific promoters became active in transcription, leading to expression of the corresponding genes, but at a different level for 469 (41%) of these 1139 genes.

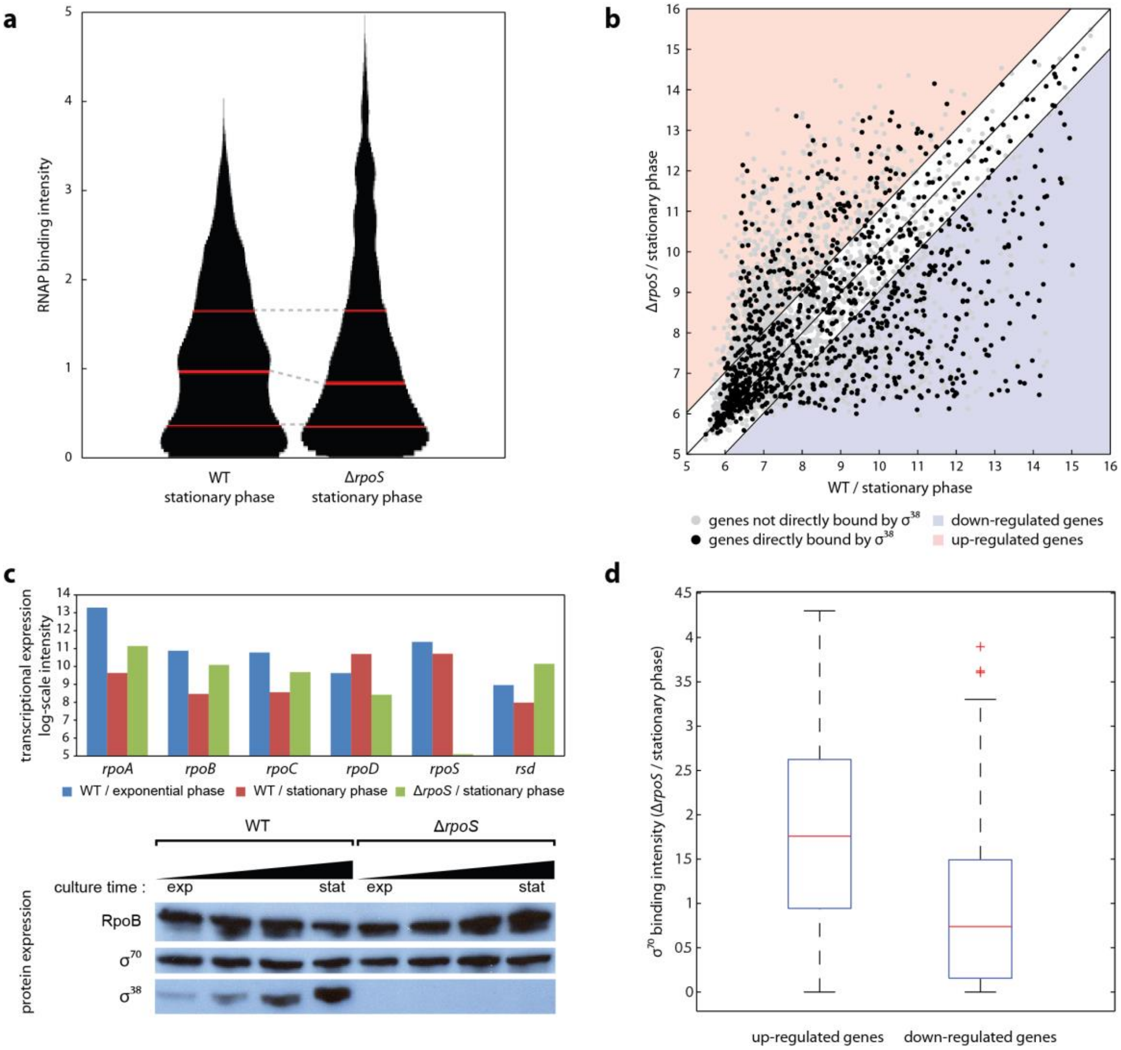


Figure 14. Competition between σ^{70} and σ^{38} in overlapping promoter regions. (a) Recruitment of RNAP core enzyme to promoters upstream of 1139 σ^{38}-specific genes was recovered when *rpoS* is knocked-out. RNAP binding intensity on the y-axis was the ChIP-chip intensity, and 3 red lines represent the first, second, and third quantiles. (b) Comparison of transcriptional expression of genes in WT and *ΔrpoS* strains. Among 1139 genes with σ^{38}-specific promoters, transcription of 178 genes was up-regulated (red background) and that of 291 genes was down-regulated (blue background) (c) Expression level of σ^{70} and σ^{38} was measured in both transcriptional and translational level. The amount of σ^{70} is abundant in exponential and stationary phase, and so it is absent of *rpoS*. (d) Up-regulated genes upon rpoS knock-out were more strongly bound by σ^{70} than down-regulated genes.

The expression of genes with σ^{38}-regulated genes was recovered when *rpoS* was knocked out; however it is still unknown which, among the other σ-factors, is replacing the role of σ^{38}. Since σ^{70} shared the largest portion of promoters with σ^{38}, it is reasonable to assume that σ^{70} would replace σ^{38} when σ^{38} is missing. In *E. coli* MC4100, it was reported that the amount of σ^{70} is in abundance during stationary phase [117]. Like that strain, *E. coli* K-12 MG1655 also showed high protein expression of σ^{70} during stationary phase in WT and *ΔrpoS* strain (Figure 14c). In addition, we examined how many genes bound by σ^{38} in the WT strain were bound by σ^{70} when *rpoS* was deleted. About 89% of those genes were found to be bound by σ^{70} when σ^{38} was missing, (Figure 16). This surprisingly high rate of σ-factor substitution explains how the majority of genes directly bound by σ^{38} recovered their expression when *rpoS* is knocked out (Figure 14b). However, it is still unclear how some of those genes were up-regulated. Since ~89% of them were bound by σ^{70}, we measured the intensity of σ^{70} binding in Δ*rpoS* during stationary phase with ChIP-chip experiments, and compared the binding intensity between up-regulated genes and down-regulated genes (Figure 14d, Figure 15). This measurement showed that up-regulated genes were bound more strongly by σ^{70} (p-value of Wilcoxon rank sum test was $4.80\text{x}10^{-18}$), suggesting that strong σ^{70} binding resulted in increased transcription. This finding indicates that the presence of σ^{38} actually contributed to repressing the transcriptional expression of some genes, presumably by competition for shared promoters between σ^{70} and σ^{38}.

Comparative analysis of the sigma factor network in close related species

With the detailed reconstruction of the σ-TUG network in *E. coli*, we are in a position to address the issue of the difference between such networks in closely related species. Genome-wide identification of transcription start sites (TSSs) of two gamma-proteobacteria, *E. coli* and *K. pneumoniae*, revealed promoter regions upstream of orthologous genes are differently organized in the two species, resulting in different usage of TSSs[7]. Since σ-factors recognize sequence elements of promoters, and that they are directly upstream of TSSs, it becomes important to determine any differences in σ-factor binding patterns. While the *E. coli* genome contains 7 σ-factors, *K. pneumoniae* is known to have only 5 σ-factors, missing *fliA* and *fecI* that are found in *E. coli*. The other 5 σ-factors which the two species have in common are highly conserved in terms of amino acid sequence similarity: 95.9% (*rpoD*), 98.5% (*rpoS*), 89.8% (*rpoN*), 95.1% (*rpoH*), and 96.3% (*rpoE*). Promoter sequence motifs examined from the TSSs were found to be identical between *E. coli* and *K. pneumoniae* suggesting that the sequence motifs for each orthologous σ-factor are identical[7, 118]. However, the different organization of upstream regulatory regions of the two species and a different pattern of transcription initiation indicates the possibility of significantly diverse σ-factor binding.

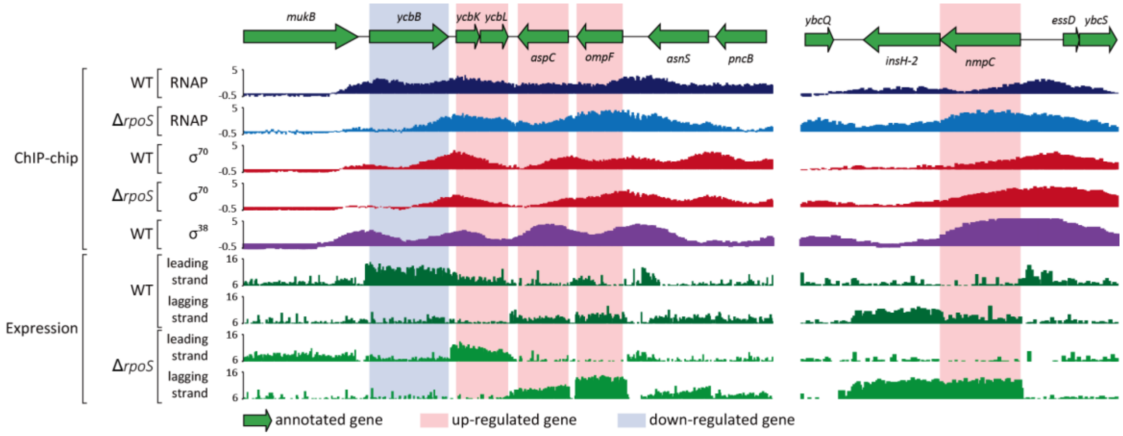


Figure 15. Examples of up-regulated and down-regulated genes when *rpoS* is knocked out. *ycbB* is an example of a down-regulated gene upon *rpoS* knock-out, and *ycbK*, *ycbL*, and *nmpC* are up-regulated genes. *ycbB* was not bound by σ^{70} when σ^{38} was missing, which resulted in no significant recruitment of RNAP enzyme complex, which was supported by no transcriptional expression of the particular gene. On the other hand, *nmpC* was more strongly bound by σ^{70} when σ^{38} was absent, which resulted in more RNAP binding and stronger expression.

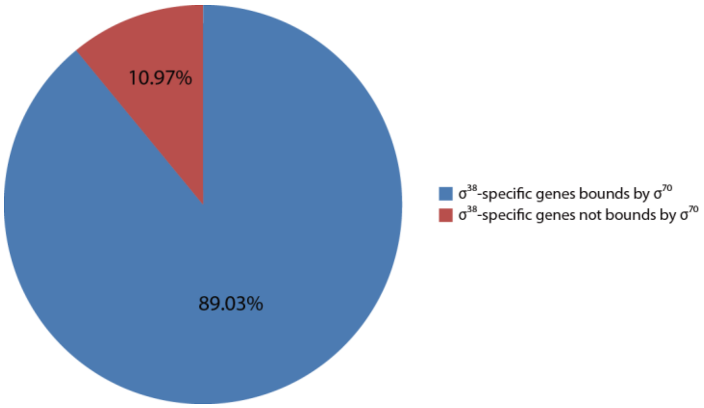


Figure 16. The majority of σ^{38}-specific promoters were bound by σ^{70} when *rpoS* is missing.

To investigate binding patterns of two major σ-factors, *rpoD* and *rpoS*, we analyzed ChIP-chip datasets for σ^{70} under exponential phase and σ^{38} under stationary phase grown in glucose minimal media from the previous study[118]. *E. coli* and *K. pneumoniae* have 4513 and 5305 genes, respectively, and 2876 coding genes were defined as orthologs by two-way reciprocal alignment. Then binding of σ^{70} and σ^{38} under specified conditions upstream of those orthologous genes were analyzed and clustered (Figure 17a). Among 2876 orthologous genes, 60% showed the same binding patterns (584 for both bound, 213 for σ^{70} bound, 102 for σ^{38} bound, and 847 for not bound). The two closely related bacteria, *E. coli* and *K. pneumoniae,* share the majority of gene contents with highly conserved sequences of most ORFs. However, conserved genes showed significantly different σ-factor binding patterns, indicating diverse gene regulation by different transcription initiation (Figure 17c, and Figure 17d). Interestingly, in some cases, altered binding of σ-factors was associated with changes in TU organization, suggesting even more diverse regulation between the two species. Although two major σ-factors were found to bind differently upstream of orthologous genes, regulation between σ-factors remained unchanged, except for the two missing σ-factors, *fliA* and *fecI*, in *K. pneumoniae* (Figure 18). Thus, regulation of gene expression by σ-factors may evolve faster than regulation among the σ-factors themselves.

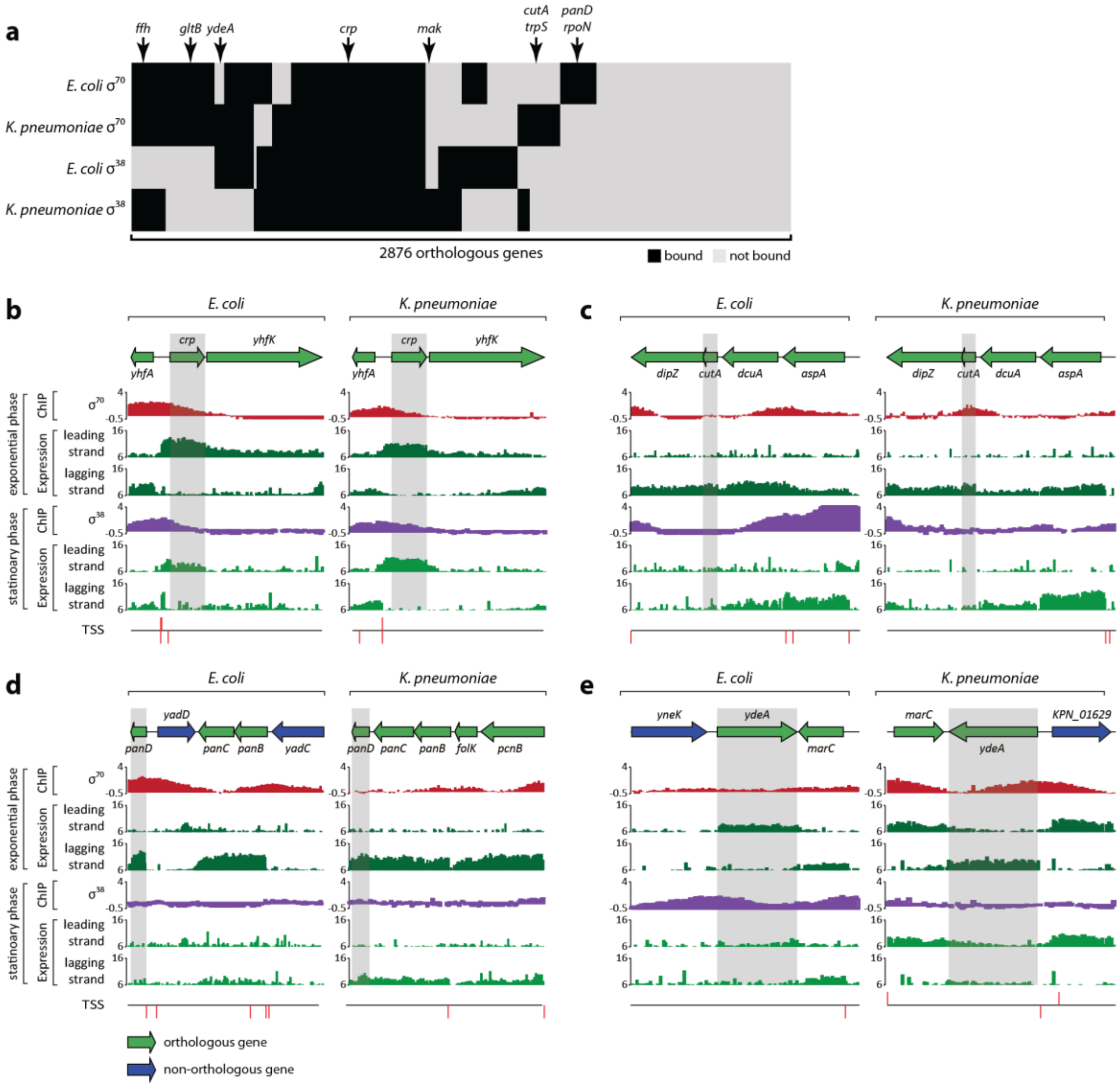


Figure 17. Conservation and divergence in transcriptional regulation by σ-factors. (a) Clustering σ-factor binding patterns revealed conserved and divergent transcriptional regulation of 2876 orthologous genes. (b) *crp* is regulated by σ^{70} and σ^{38} in both species, showing regulation conservation. (c) In *E. coli*, *cutA* is a part of *dcuA*-*cutA*-*dipZ* transcription unit and is regulated by σ^{70} and σ^{38}, while *cutA* in K. pneumoniae is the first gene in its transcription unit and is directly bound by σ^{70}. (d) In *K. pneumoniae*, *panD* is a part of *panBCD* transcription unit and that transcription unit is regulated by σ^{70}, however in E. coli *panD* is separated from *panBC* by *yadD*, making another distinct transcription unit. Those two transcription units are both regulated by σ^{70}. (e) A genomic region containing *ydeA* and *marC* in both species was inverted, and this genomic inversion was accompanied with transcription regulation switch between σ^{70} and σ^{38}.

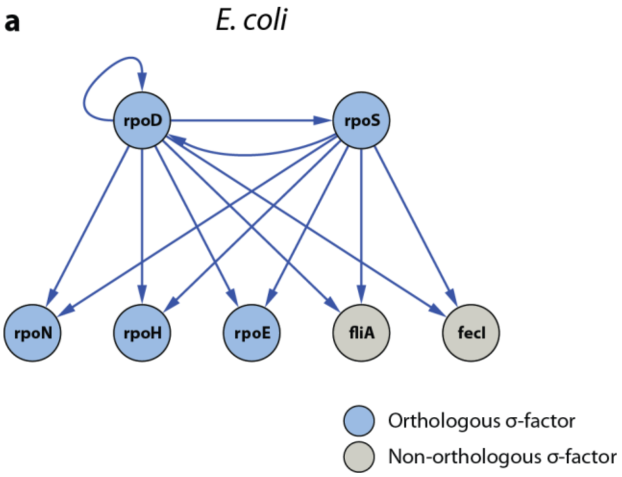


Figure 18. Comparison of transcriptional regulation by two major σ-factors, σ^{70} and σ^{38}, in two closely-related bacteria.

Conclusions

Genome-scale measurements have enabled the reconstruction of the σ-TUG network in *E. coli* K-12 MG1655. This network is at the core of transcriptional regulation in bacteria. Its reconstruction has enabled the assessment of its topological characteristics, functional states, and limited comparison with related species. With the integration of a growing body of experimental data on transcription factor binding and activity, the resource provided here open up the possibility of developing a comprehensive reconstruction of the entire transcriptional regulatory network in *E. coli* that would simultaneously describe the function of σ-factors and transcription factors that produce the entire expression state of the organism.

Methods

Bacterial strains, media, and growth conditions: *E. coli* K-12 MG1655 and its isogenic knock-out strains were used in this study. The deletion mutants (*ΔrpoS and ΔrpoN)* were generated by a λ Red and FLP-mediated site-specific recombination system[119]. *E. coli* cells were harvested at mid-exponential phase (OD_{600nm} ~ 0.5) with the exception of stationary phase experiments (OD_{600nm} ~ 1.5). Glycerol stocks of *E. coli* strains were inoculated into M9 or W2 minimal media[120] (for nitrogen-limiting condition) with glucose (2 g/L) and cultured at 37 ^{o}C with constant agitation overnight. Cultures were then diluted 1:100 into 50 mL of fresh minimal media and cultured at 37 ^{o}C to appropriate cell density. For heat-shock experiments, cells were grown to mid-exponential phase at 37 ^{o}C and half of the culture was used as a control, while the remaining culture was transferred into pre-warmed (50 ^{o}C) media and incubated for 10 min. For nitrogen-limiting condition, ammonium chloride in the minimal media was replaced by glutamine (2 g/L).

Total RNA isolation: Three milliliters of cell culture were mixed with 6 mL RNAprotect Bacteria Reagent (Qiagen). Samples were mixed immediately by vortexing for 5 seconds, incubated for 5 minutes at room temperature. Then they were centrifuged at 5000 ×*g* for 10 minutes. The supernatant was decanted and any residual supernatant was removed by inverting the tube once onto a paper towel. Total RNA samples were then isolated using RNeasy Plus Mini kit (Qiagen) in accordance with the manufacturer’s instructions. Samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific) and the quality of the isolated RNA was checked by visualization on agarose gels and by measuring the sample’s A_{260}/A_{280} ratio (>1.8).

Transcriptome analysis: Transcriptome dataset with oligonucleotide tiling microarrays for *E. coli* K-12 MG1655 wild type grown under 4 conditions, exponential phase, stationary phase, heat-shock, and nitrogen-limiting condition, were taken from the previous study[11]. In order to get the transcriptome dataset for *E. coli* deletion mutant *ΔrpoS*, the protocol previously described [7] was adapted for the deletion mutant in the current study. Briefly, 10 μg of purified total RNA sample was reverse transcribed to cDNA with amino-allyl dUTP. The amino-allyl labeled cDNA samples were then coupled with Cy3 monoreactive dyes (Amersham). Cy3 labeled cDNAs were fragmented to 50 ~ 300 bp range with DNase I (Epicentre). High-density oligonucleotide tiling arrays consisting of 371,034 50-mer probes spaced 25 bp apart across the whole *E. coli* genome were used (Roche Nimblegen). Hybridization, wash, and scan were performed in accordance with manufacturer's instructions. Three biological replicates were utilized for stationary phase under glucose minimal media. Probe level data were normalized with RMA (Robust Multiarray Analysis) algorithm without background correction, as implemented in NimbleScan 2.4 software.

TSS-seq by modified 5' RACE and deep-sequencing: The raw TSS dataset for exponential phase was taken from the precious study[7]. For the other 3 conditions, stationary phase, heat-shock, and nitrogen-limiting condition, TSS determination protocol previously described [7] was adapted for *E. coli* K-12 MG1655. To enrich intact 5' tri-phosphorylated mRNAs from the total RNA, 5' mono-phosphorylated ribosomal RNA (rRNA) and any degraded mRNA were removed by treatment with a Terminator 5'-Phosphate Dependent Exonuclease (Epicentre) at 30°C for 1 hr. The reaction mixture consisted of 10 μg purified total RNA, 1 μL terminator exonuclease, reaction buffer, and RNase-free water up to total 20 μL. The reaction was terminated by adding 1 μL of 100 mM EDTA (pH 8.0). Intact tri-phosphorylated RNAs were precipitated by adding 1/10 volume of 3 M sodium acetate (pH 5.2), 3 volumes of ethanol, and 2 μL of 20 mg/mL glycogen. RNA was precipitated at -80 °C for 20 min and pelleted, washed with 70% ethanol, dried in Speed-Vac for 7 minutes without heat, and resuspended in 20 μL nuclease free water. The tri-phosphorylated RNA was then treated with RNA 5'-Polyphosphatase (Epicentre) to generate 5'-end mono-phosphorylated RNA for adaptor ligation. The RNA sample from the previous step was mixed with 2 μL 10× reaction buffer, 0.5 μL SUPERase-In (Ambion), 1 μL RNA 5'-Polyphosphatase, and RNase-free water up to 20 μL. The mixture was incubated at 37°C for 30 minutes and reaction was stopped by phenol-chloroform extraction. Ethanol precipitation was carried out for isolating the RNA as described above. To ligate 5′ small RNA adaptor (5’-GUUCAGAGUUCUACAG UCCGACGAUC-3’) to the 5′-end of the mono-phosphorylated RNA, the enriched RNA samples were incubated with 100 μM of the adaptor and 2.5 U of T4 RNA ligase (New England Biolabs). cDNAs were synthesized using the adaptor-ligated mRNAs as template using a modified small RNA RT primer from Illumina (5’-CAAGCAGAAGACGGCATACGANNNNNNNNN -3’) and Superscript II Reverse Transcriptase (Invitrogen). The RNA was mixed with 25 μM modified small RNA RT primer and incubated at 70 ^{o}C for 10 min and then at 25 ^{o}C for 10 min. Reverse transcription was carried out at 25 ^{o}C for 10 min, 37 ^{o}C for 60 min, 42 ^{o}C for 60 min, and followed by incubation at 70 ^{o}C for 10 min. A reaction mixture for reverse transcription consisted of the following components: 5× 1st strand buffer; 0.01 M DTT; 10 mM dNTP mix; 30 U SUPERase•In (Ambion); and 1500 U SuperScript II (Invitrogen). After the reaction, RNA was hydrolysed by adding 20 μL of 1 N NaOH and incubation at 65 ^{o}C for 30 min. The reaction mixture was neutralized by adding 20 μL of 1 N HCl. The cDNA samples were amplified using a mixture of 1 μL of the cDNA, 10 μL of Phusion HF buffer (NEB), 1 μL of dNTPs (10 mM), 1 μL SYBR green (Qiagen), 0.5 μL of HotStart Phusion (NEB), and 5 pmole of small RNA PCR primer mix. The amplification primers used were 5’-AATGATACGGCGACCAC CGACAGGTTCAGAGTTCTACAGTCCGA-3’ and 5’-CAAGCAGAAGACGGCATACGA-3’. The PCR mixture was denatured at 98 °C for 30 s and cycled to 98 °C for 10 s, 57 °C for 20 s, and 72 ℃ for 20 s. Amplification was monitored by a LightCycler (Bio-Rad) and stopped at the beginning of the saturation point. Amplified DNA was run on a 6% TBE gel (Invitrogen) by electrophoresis and DNA of size ranging from 100 to 300 bp were size fractionated. Gel slices were dissolved in two volumes of EB buffer (Qiagen) and 1/10 volume of 3 M sodium acetate (pH 5.2). The amplified DNA was then ethanol-precipitated and resuspended in 15 μL DNase-free water (USB). The final samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific).

Sequencing, data processing and mapping: The data processing and mapping of the sequencing results to get potential TSSs was performed in the identical way it was done in the previous study[7]. In brief, the amplified cDNA libraries from two biological replicates for each condition were sequenced on an Illumina Genome Analyzer. Sequence reads for cDNA libraries were aligned onto the *E. coli* K-12 MG1655 genome (NC_000913) using Mosaik (http://code.google.com/p/mosaik-aligner) with the following arguments: hash size = 10, mismatach = 0, and alignment candidate threshold = 30 bp. Only reads that aligned to a unique genomic location were retained. Two biological replicates were processed separately, and only sequence reads presented in both biological replicates were considered for further process. The genome coordinates of the 5'-end of these uniquely aligned reads were defined as potential TSSs. Among potential TSSs, only TSSs with the strongest signal within 10 bp window were kept to remove possible noise signals. TSSs with greater than or equal to 40% of the strongest signal upstream of an annotated gene were considered as multiple TSSs. The strongest signal was defined as the potential TSSs with the highest number of reads among the TSSs upstream of an annotated gene. For further analysis, among TSSs, ones that lie within RNAP binding regions (**Table S3**) were used for integration with σ-factor binding information.

Chromatin immunoprecipitation and microarray analysis: Briefly, the immunoprecipitated RNAP-associated DNA fragments were fluorescently labeled and hybridized to a high-density oligonucleotide tiling microarray representing the entire *E. coli* genome[108]. To identify *in vivo* binding regions of RNAP complex and 6 σ-factors, σ^{70}, σ^{54}, σ^{38}, σ^{32}, σ^{28}, and σ^{19}), we isolated DNA fragments bound to those RNAP subunits from formaldehyde-crosslinked *E. coli* cells through chromatin immunoprecipitation with 6 different antibodies that specifically recognize each subunit (NeoClone). *E. coli* strain harboring RpoH-8myc was constructed in the way previously described[16, 23] and used for the σ^{38} ChIP-chip with anti-c-myc antibody (9E10, Santa Cruz biotech). Cells were grown under appropriate conditions and harvested. The IP DNA and mock-IP DNA were hybridized onto high-resolution whole-genome tiling microarrays, which contained a total of 371,034 oligonucleotides with 50-bp probes overlapping 25 bps on both forward and reverse strands. Tiling microarrays were hybridized, washed, and scanned in accordance with the manufacturer's instructions (Roche NimbleGen). To increase depth of the number of promoter regions identified, datasets were generated under multiple growth conditions with a total number of 45 ChIP-chip experiments (36 for σ-factors and 9 for RNAP), and analyzed (**Table S1**). We were not able to perform ChIP-chip experiment for σ^{24}. This could be because the expression level of σ^{24} was not high enough, or we were not able to find an appropriate condition to activate σ^{24}. To remedy the missing dataset, we deployed known binding information for σ^{24} from the public database[121].

ChIP-chip data analysis: We used the peak-finding algorithm built in the NimbleScan software from Roche Nimblegen, and following analysis was performed in the way previously described[11, 17]. In brief, transcription factor–binding regions were identified by using peak-finding algorithm, which is built in the NimbleScan software (Roche NimbleGen). Processing of ChIP-chip data was performed in three steps: normalization, IP/mock-IP ratio computation (in $\log_2$ scale) and enriched-region identification. The $\log_2$ ratios of each spot in the microarray were calculated from the raw signals obtained from both Cy5 and Cy3 channels, and then the values were scaled by Tukey biweight mean. The $\log_2$ ratio of Cy5 (IP DNA) to Cy3 (mock-IP DNA) for each point was calculated from the signals. Then, the biweight mean of this $\log_2$ ratio was subtracted from each point. Each log-ratio dataset (from duplicate or triplicate samples) was used to identify transcription factor–binding regions using the software (width of sliding window = 300 base pairs). Our approach to identify the transcription factor–binding regions was to first determine binding locations from each dataset and then combine the binding locations from at least five of six datasets to define a binding region using the recently developed MetaScope visualization software and genome browser (http://systemsbiology.ucsd.edu/Downloads/MetaScope).

Western blotting: *E. coli* K-12 MG1655 and Δ*rpoS* deletion mutant cells were grown in M9 minimal media with 0.2% glucose, and were harvested from mid-exponential phase to stationary phase every 2 hours. Cells were pelleted by centrifugation, and were lysed with lysozyme in a lysis buffer containing 10 mM Tris-HCl (pH 7.5), 100 mM NaCl, and 1 mM EDTA. The supernatant was taken after centrifugation to remove unlysed cells. The concentration of total protein in the lysate was measured with Qubit Protein Assay Kit (invitrogen), and 5 μg of total protein samples were mixed with 4X SDS-PAGE sample loading buffer (Invitrogen) and 10 mM DTT, and then boiled at 90 ^{o}C for 5 min. Boiled samples were separated by electrophoresis with 10% Bis-Tris Gel in MOPS buffer, and transferred onto Hybond-ECL membrane (Amersham Biosciences). The membrane was briefly washed in TBS buffer with 0.1% Tween-20 (1X TBS-T) for 5 min on a rocker, and then treated with 2% skim milk in TBS-T buffer for 1 hr with mild shaking. The membrane was washed twice with TBS-T for 5 min each on a rocker, and then it was sliced into 3 pieces having RpoB, σ^{70} and σ^{38} in each slice. Sliced membranes were treated with anti-RpoB, anti-σ^{70}, and anti-σ^{38} antibodies (NeoClone) in 1/10,000 dilution for 1 hour on a rocker. Each membrane slices were washed in TBS-T for 15 min once and 5 min three times each, and then was treated with HRP-conjugated anti-mouse IgG (Amersham Bioscience) in 1/10,000 dilution for 30 min on a rocker, followed by washing in TBS-T for 15 min once and 5 min three times each. Chemiluminescent detection was applied to peroxidase conjugates on membrane to detect the amount of RpoB, σ^{70}, and σ^{38}.

Acknowledgements

Chapter 4, in total, is a reprint of the material as it appears in Cho, B.K.*, Kim, D.*, Knight E.M., Zengler, K., Palsson, B.Ø. Genome-scale Reconstruction of Sigma Factor Network in *Escherichia coli*. BMC Biology, 12(1):4 (2014). I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 5: Deciphering the Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*

Iron is essential for many fundamental cellular processes, including N_2 fixation, DNA synthesis, the tricarboxylic acid (TCA) cycle, and respiration[122]. Its function depends on its incorporation into proteins either as an isolated ion or in a more complex form such as iron-sulfur (FeS) clusters or a heme group. Unfortunately, although iron is essential for most organisms, it can also be extremely toxic under oxic environments. Its ability to interact with superoxide and hydrogen peroxide can generate the highly reactive and damaging hydroxyl radical species by Fenton or Haber-Weiss reactions[123]. Thus, the amount of cellular free iron should be carefully managed to protect cells from iron-induced toxicity.

In most gram-negative bacteria, including *Escherichia coli*, ferric uptake regulator (Fur) regulates iron metabolism to precisely control cytoplasmic iron levels. Although classical Fur regulation involves the binding of Fur-Fe^{2+} to the promoter region as a repressor, recent studies have demonstrated that Fur-Fe^{2+} can function as an activator[124] and even Fur without an iron cofactor can act as both in some pathogenic bacteria[122, 125-127]. In *E. coli*, the general role of Fur in iron metabolism has been extensively investigated from *in vitro* DNA-binding experiments and related mutation analysis[128-130]. However, much less is known about genome-scale *in vivo* Fur-binding events and the regulatory network they comprise. A complete reconstruction of the Fur transcriptional regulatory network in response to iron availability will reveal detailed modes of Fur regulation by emphasizing direct regulatory mechanisms and distinguishing them from indirect regulation. Furthermore, a better understanding of the Fur regulatory network can shed light on unanswered questions about its role in fundamental cellular processes, other than direct iron metabolism, that need to be coordinated when *E. coli* responds to iron availability.

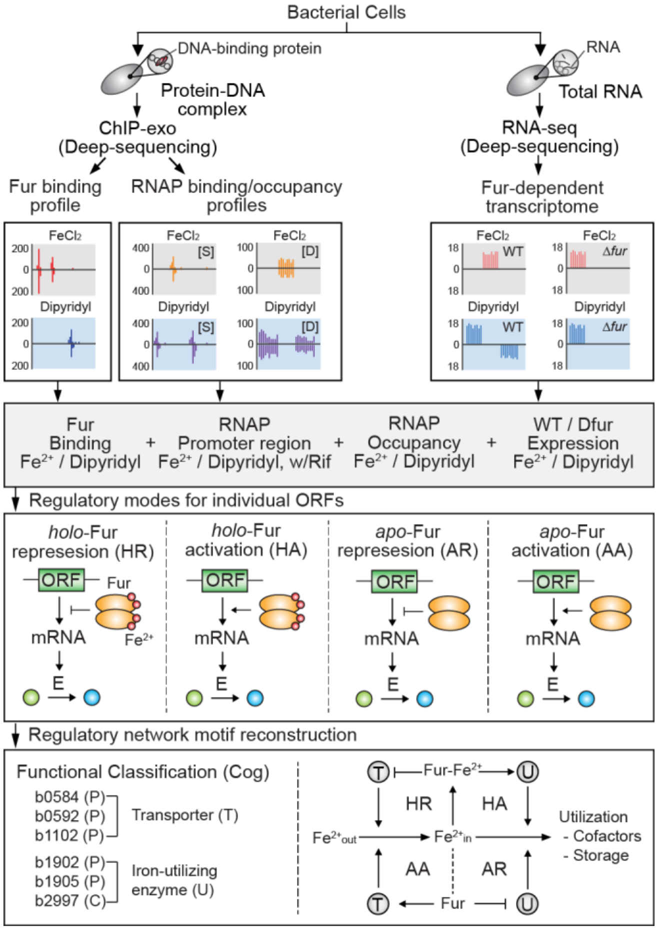

Bacterial Cells
DNA-binding protein
Protein-DNA complex
ChIP-exo (Deep-sequencing)
Fur binding profile
RNAP binding/occupancy profiles
RNA
Total RNA
RNA-seq (Deep-sequencing)
Fur-dependent transcriptome
FeCl2
Dipyridyl
200
0
400
100
18
[S]
[D]
WT
Δfur
Fur Binding Fe2+ / Dipyridyl
+
RNAP Promoter region Fe2+ / Dipyridyl, w/Rif
RNAP Occupancy Fe2+ / Dipyridyl
WT / Dfur Expression Fe2+ / Dipyridyl
Regulatory modes for individual ORFs
holo-Fur represion (HR)
holo-Fur activation (HA)
apo-Fur represeion (AR)
apo-Fur activation (AA)
ORF
Fur
Fe2+
mRNA
E
Regulatory network motif reconstruction
Functional Classification (Cog)
b0584 (P)
b0592 (P)
b1102 (P)
Transporter (T)
b1902 (P)
b1905 (P)
b2997 (C)
Iron-utilizing enzyme (U)
T
U
Fur-Fe2+
HR
HA
AA
AR
Fe2+out
Fe2+in
Fur
Utilization
- Cofactors
- Storage

Figure 19. Flowchart of the method. The *in vivo* genome-wide Fur-binding maps along with the changes in RNAP bindings (S, static map) and occupancies (D, dynamic maps) and Fur-dependent transcriptomic data were generated under both iron-replete and iron starvation conditions. Combined data sets were used to determine direct Fur regulon and the regulatory mode for individual ORFs governed by Fur. The Fur regulatory network was reconstructed by connecting iron transport and utilization regulatory motifs with negative-feedback loops.

In this study, we applied a systems biology approach by integrating genome-scale data from chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) for Fur and RNA polymerase (RNAP) and from strand-specific massively parallel cDNA sequencing (RNA-seq) to decipher the Fur regulatory network in response to iron availability following the workflow shown in Figure 19. We first sought to fully reconstruct the Fur regulon. We examined the Fur-binding sites on the *E. coli* genome and also measured the changes in RNAP bindings/occupancies and mRNA transcript levels on a genome-scale to identify the direct Fur regulon. From this data, we then determined regulatory modes for individual open reading frames (ORFs) subject to Fur regulation and reported distinct mechanisms of *apo*- and *holo*-Fur activation as well as *holo*-Fur repression in *E. coli*. Finally, we identified that the Fur regulatory network maintains intracellular iron concentration by connecting iron transport and utilization enzymes with negative-feedback loop pairs. The reconstruction of the Fur regulatory network provides a comprehensive view of the coordinative genome-wide regulatory role of this important global transcription factor.

Genome-wide identification of Fur binding sites

Previously, Fur-binding sites in *E. coli* have been characterized by *in vitro* DNA-binding experiments and related mutation analysis[128]; however, direct measurement of *in vivo* Fur binding has not been available. We therefore first employed the ChIP-exo method to determine the *in vivo* Fur-binding maps with near 1-bp resolution in *E. coli* under both iron-replete and iron starvation conditions (Figure 20a).

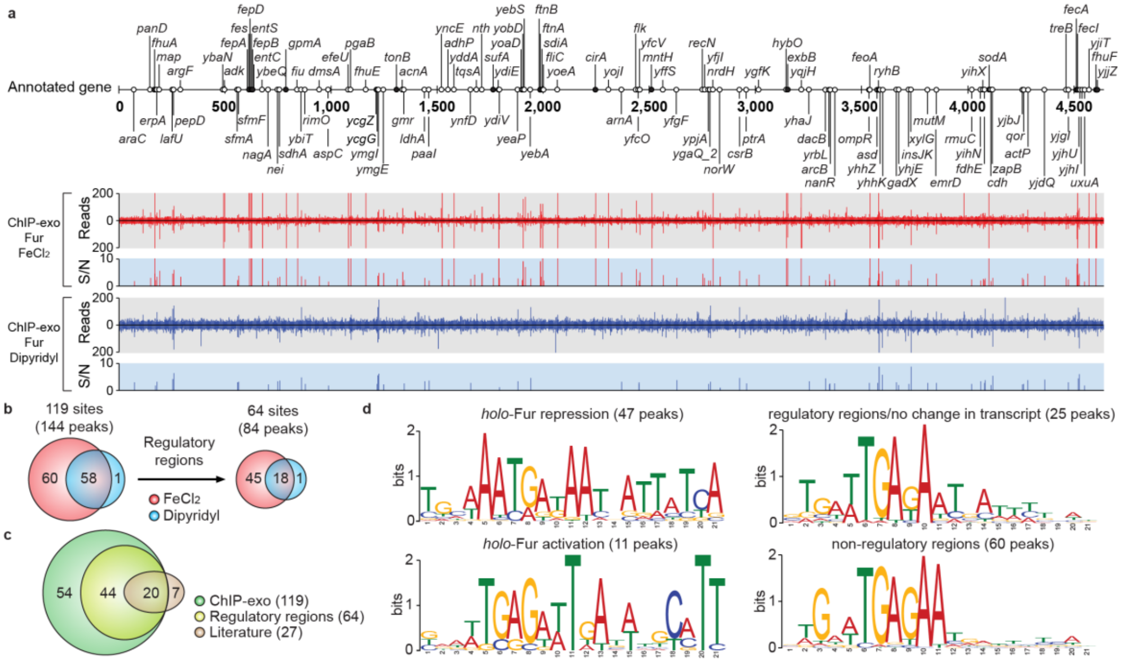


Figure 20. Genome-wide distribution of Fur-binding sites. (a) An overview of Fur-binding profiles across the *E. coli* genome at mid-exponential growth phase under both iron-replete (red) and iron starvation (blue) conditions. Black and white dots indicate previously known and newly found Fur-binding sites, respectively. (b) Overlaps between Fur-binding sites under iron-replete and iron starvation conditions. (c) Comparison of the Fur-binding sites obtained from this study (ChIP-exo) with the literature information. (d) Sequence logo representations of the Fur-DNA binding profiles.

Using a peak finding algorithm, 118 and 59 unique and reproducible Fur-binding sites were identified under iron-replete and iron starvation conditions, respectively (Figure 20a). The high-resolution of ChIP-exo method enabled us to identify multiple binding peaks in several binding sites and separate binding peaks in divergent promoter regions, resulting in 143 and 61 peaks under iron-replete and iron starvation conditions, respectively. Most of the binding sites (58 of 59) under iron starvation condition overlapped with those under iron-replete condition thus giving a total number of binding sites of 119 (Figure 20b). Only 54% of them (64 of 119) were located in putative regulatory regions and the remaining 46% were found in intragenic regions or between two coding regions of convergent genes. In addition, 69% (40 of 58) of the overlapped binding sites were located in non-regulatory regions. One interesting exception was the upstream region of *ycgZ-ymgA-ariR-ymgC* where Fur occupied only under iron starvation condition, indicating possible direct regulation by Fur under this condition. Prior to this study, 27 Fur-binding sites had been identified with strong experimental evidence[128], 74% (20 of 27) of which were also detected in this study (Figure 20c). Collectively, a total of 98 Fur-binding sites were newly identified in this study, 45% (44 of 98) of which are located at putative regulatory regions (Figure 20c), expending the current scope of the Fur regulatory network.

Genome-wide reconstruction of Fur regulon

Currently, a total of 70 genes in 27 transcription units (TUs) have been characterized as members of Fur regulon to be directly regulated by Fur in *E. coli* based on strong experimental evidence[128]. From our ChIP-exo analysis, we significantly expanded the size of the potential Fur regulon to comprise 110 target genes in 64 TUs. To determine causal relationship between Fur binding and transcript level, we compared transcript levels between wild-type and Δ*fur* mutant cells grown under both iron-replete and starvation conditions. Overall, a total of 678 genes were differentially expressed by the Fur deletion under either iron-replete (553) or iron starvation (211) condition, 86 of which overlapped in both conditions (Figure 22). In addition, we also measured the RNAP occupancy on a genome-scale using ChIP-exo to gain a better mechanistic understanding of the transcriptional regulatory roles of Fur. We could identify locations where RNAP occupancy is increased or decreased due to changes in Fur-binding levels and iron availability.

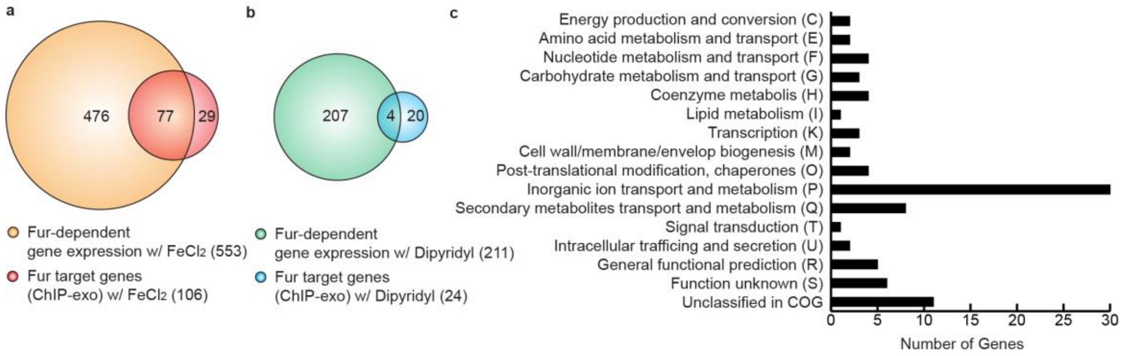


Figure 21. Genome-wide identification of Fur regulon. Comparison of ChIP–exo results and gene expression profiles under (a) iron-replete and (b) iron starvation conditions to distinguish direct and indirect Fur regulon. (c) Functional classification of genes directly regulated by Fur.

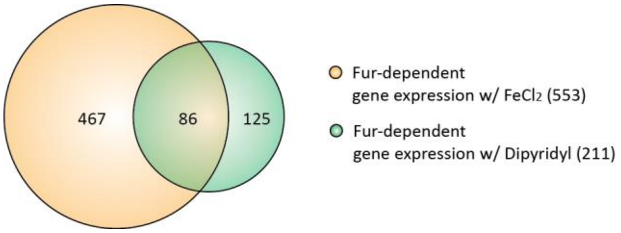


Figure 22. Fur-dependent transcriptome in response to iron availability. We compared transcript levels between wild-type and Δ*fur* mutant under both iron-replete ($FeCl_2$) and starvation (Dipyridyl) conditions.

Combining our ChIP-exo results of Fur- and RNAP-binding/occupancy maps with Fur-dependent transcriptome data, we could clarify target genes for direct Fur regulation depending on iron availability (Figure 21a and Figure 21b). A total of 81 genes in 42 TUs were directly regulated by Fur under either iron-replete (77 genes) or iron starvation (4 genes) condition. Only 13% (81 of 678) of the Fur-dependent genes were directly regulated by Fur. These genes were categorized into clusters of orthologous groups (COG) categories according to their functional annotation (Figure 21c). As expected, the COG category for inorganic ion transport and metabolism (P) was found to be overrepresented. However, they also encompassed a diverse range of COG functional categories, indicating that Fur may play complicated regulatory roles beyond iron metabolism to coordinate associated cellular processes. The other 597 genes would be targets for either indirect Fur regulation (mediated by RyhB small RNA) or other stress-responsive TFs since iron availability can generate different types of damages such as redox imbalance and oxidative stress[131].

Regulatory modes of Fur in response to iron availability

An interesting aspect of the regulatory modes of Fur is its variable response to iron availability. Classical Fur regulation involves the binding of iron-bound Fur to the promoter region as a repressor; however, recent studies have shown cases where Fur functions as an activator or as both even in the absence of its iron cofactor in some pathogenic species[122, 125, 126]. In order to define regulatory modes of Fur in *E. coli*, we classified the regulation of 81 genes into 4 different modes (*holo*-Fur repression, *holo*-fur activation, *apo*-Fur repression, and *apo*-Fur activation) depending on Fur binding with and without iron.

For example, the promoter regions of *fepA* and *fes* TUs, where two divergent promoters exist, were extensively occupied by Fur under iron-replete condition with a decrease in RNAP bindings and transcript levels (Figure 24b, *holo*-Fur repression). We denoted this regulatory mode as *holo*-Fur repression (HR). A total of 65 genes were regulated by this mode, and 23 of them were previously investigated with either weak evidence (*fiu-ybiX*, *efeUOB*, *fhuE*, *yncE*, *yddA-yddB*, *ydiE*, *nrdHIEF*, *yqjH*, and *feoABC*) or no evidence (*ybaN*, *adhP*, *ynfD*, *yoeA*, and *yojI*) (Figure 24a). Interestingly, the *efeUOB* operon, which encodes a ferrous iron transporter complex, was found to be still repressed by iron-bound Fur, even though this operon is cryptic due to a frameshift mutation[132]. In contrast, associations of Fur on the promoter regions of *ftnB* and *ftnA* TUs increased RNAP bindings and transcript levels under iron-replete condition (Figure 24b, *holo*-Fur activation). Thus, we denoted this regulatory mode as *holo*-Fur activation (HA). In the previous study, only *ftnA* was thought to be a direct target for activation by iron-bound Fur[124]. However, we identified 11 more target genes (*argF*-*yagI*, *adk*, *ftnB*, *hybOA*, *zapB*, *uxuAB*, *acnA*, and *yjiT*) for direct activation by iron-bound Fur (Figure 24a). Most of them utilize iron or other divalent metal ions as cofactors. It is known that *acnA*, encoding aconitase not only catalyzing the inter-conversion of citrate and isocitrate in TCA cycle but also sensing iron starvation and oxidative stress, was indirectly regulated by Fur via RyhB-mediated mRNA degradation[130]. Surprisingly, *acnA* was also directly activated by iron-bound Fur. This observation is in agreement with the Northern Blot analysis from the previous study where Δ*ryhB*Δ*fur* double mutant showed much less amount of *acnA* transcript compared to Δ*ryhB* mutant, presumably due to the loss of direct activation by Fur[130]. Another aspect of Fur is its regulation mediated by binding of iron-free form in some pathogenic bacterial species[122, 125, 126]. Although *E. coli* K-12 MG1655 has been regarded not to have this regulatory mode, we observed that the promoter region of the *ycgZ*-*ymgA*-*ariR*-*ymgC* operon is bound by Fur only under iron starvation condition, and this binding was accompanied with an increase in both RNAP binding and transcript level (Figure 24b, *apo*-Fur activation). We denoted this regulatory mode as *apo*-Fur activation (AA). Three genes (*ymgA*, *ariR*, and *ymgC*) in this operon are associated with biofilm formation and one of them (*ariR*) is also related with acid resistance[133]. Thus, Fur could play a key role in suppression of biofilm formation and resistance to acidic stress by activating this particular operon under iron starvation condition. We did not observe the regulatory mode of *apo*-Fur repression (AR) for any TU in *E. coli* K-12 MG1655, although some pathogenic strains have been reported to have this mode[122, 125, 126].

To further analyze the sequences of individual Fur-binding sites, we created four different datasets with footprint sequences based on the binding locations (regulatory or non-regulatory region) and modes of regulation (HR, HA, or no change in transcript level), resulting in 47 binding sites in *holo*-Fur repression, 11 binding sites in *holo*-Fur activation, 25 binding sites in regulatory regions but no change in transcript level, and 60 binding sites in non-regulatory regions. A consensus sequence for AA mode was not identified because it only had one binding site. We also arbitrarily extended the sequence of each site by 20 nt at each end to allow for adjacent sequences to be included in the motif search procedure[134]. The motif search for *holo*-Fur repressed genes yielded a consensus sequence with canonical Fur boxes containing an internal palindromic 7-1-7 sequence[135] (Figure 20d). In contrast, the identified motif for *holo*-Fur activated genes was similar to previously identified Fur boxes but had an incomplete palindromic sequence. Interestingly, the sequence motifs obtained from binding sites in non-regulatory regions and those in regulatory regions without transcript level change resembled the half of the previously known consensus Fur boxes, indicating that Fur binds to these sites with low affinity due to the recognition sequence but not affect the transcription level of the downstream genes.

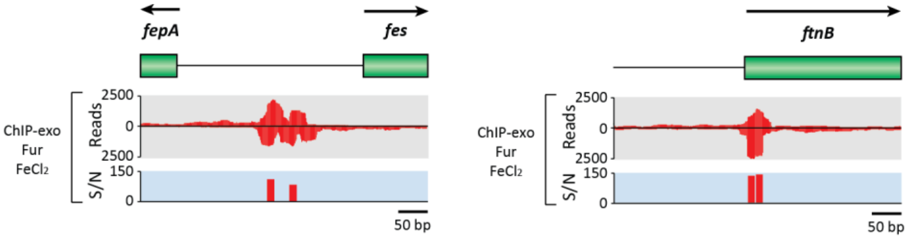


Figure 23. Zoom-in examples of Fur bindings identified by ChIP-exo. The high-resolution of ChIP-exo method enabled us to separate binding peaks in divergent promoter regions between *fepA* and *fes* as well as identify multiple biding peaks for *ftnB*. Arrows indicate the direction of transcription of each gene.

Fur-regulated feedback loop motifs for iron metabolism

Next, we reconstructed the Fur regulatory network in *E. coli* to observe how Fur regulates genes for iron metabolism. After the functional classification of 81 genes directly regulated by Fur, we observed that the functions of 55% (44 genes) of those genes were mainly localized to iron transport and metabolism. To identify the metabolic pathways regulated by Fur, the members of Fur regulon were mapped to the iron uptake/utilization pathways[123, 128, 129, 136] (Figure 25a). Under iron-replete condition, Fur repressed entire enterobactin biosynthesis/transport and iron or iron-complex transport systems including *fhuE*, *feoABC*, and *fiu*. On the other hand, it activated several iron-utilizing enzymes including *ftnB*, *uxuAB*, and *hybOA* under the same condition. As shown in a previous study, one of the FeS cluster assembly systems mediated by *suf* operon is directly regulated by Fur while the other mediated by *isc* operon is indirectly regulated by Fur via RyhB-mediated mRNA degradation[137]. This RyhB-mediated regulation also forces genes associated with iron utilization to be regulated according to the intracellular iron pool (*sdhDC*, *fumA*, *bfr*).

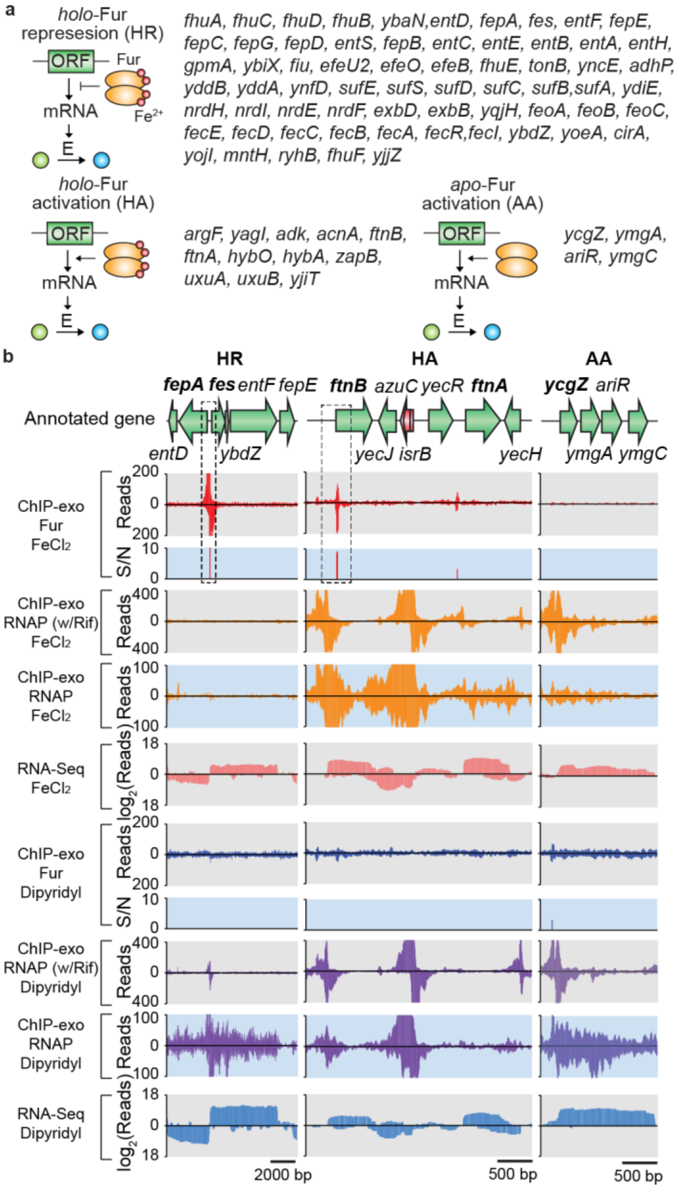


Figure 24. Regulatory modes of individual ORFs governed by Fur in response to iron availability. (a) Classification of the Fur regulatory modes based on the location analysis of Fur and RNAP and gene expression profiling in response to iron availability. (b) Examples of *holo*-Fur repression (HR) mode (*fepA-entD* and *fes-ybdZ-entF-fepE*), *holo*-Fur activation (HA) mode (*ftnB* and *ftnA*), and *apo*-Fur activation (AA) mode (*ycgZ-ymgA-ariR-ymgC*). Boxes with dotted lines are zoom-in examples in Figure 23.

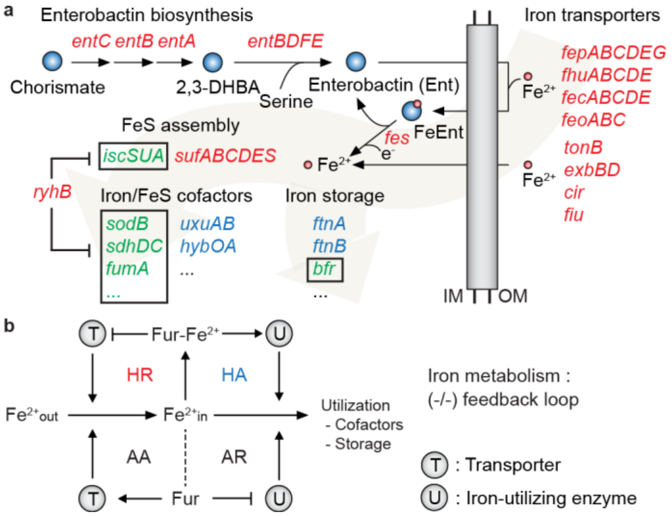


Figure 25. Iron acquisition/utilization pathways directly regulated by Fur and regulatory network motif. (a) The iron acquisition (enterobactin biosynthesis and iron/enterobatin transport), iron utilization (iron storage and iron/FeS cofactors), and FeS assembly pathways are represented. The genes regulated by HR and HA are depicted by red and blue characters, respectively. The genes regulated by RyhB-mediated mRNA degradation are depicted by green characters with black boxes. 2,3-DHBA, 2,3-dihydroxybenzoic acid; IM, inner membrane; OM, outer membrane. (b) Schematic diagram for the Fur regulatory motif reconstruction for iron metabolism.

From this analysis, we were able to connect transport and utilization feedback loop pairs[138] (Figure 25b). In the left loop, Fur regulates transcription of the transport protein (T) in either the presence or absence of iron, facilitating the uptake of the iron (Fe^{2+}_{out}), whereas in the right loop, Fur controls transcription of iron-utilization enzymes (U) that store Fe^{2+}_{in} or use it as cofactors. The possible logical structures of the feedback loop motifs can be characterized depending on how Fur-Fe^{2+} (or Fur) represses or activates both T and U. In this case, Fur regulatory network showed (-/-) motif for iron metabolism, which represses transcription of iron transporter genes (HR) and enhances production of iron-utilizing proteins (HA) under iron-replete condition (Figure 25b). Taken together, these observations demonstrate that *E. coli* Fur in the iron-bound form directly controls transcription of transporters and utilization enzymes in a manner of (-/-) motif to maintain intracellular iron concentration.

Elucidation of complex roles of Fur regulatory network beyond iron metabolism

The genome-scale reconstruction of Fur regulatory network in *E. coli* enabled us to extend the scope of its roles in response to iron availability (Figure 26). We classified them into four different categories: iron metabolism, DNA synthesis, energy metabolism, and nutrient search. Most importantly, Fur regulates iron metabolism comprised of a set of genes involved in iron uptake, storage, and FeS assembly in response to iron availability (Figure 26a). Under iron-replete condition, Fur activated transcription of iron-utilization genes (HA) while repressing that of iron transporter genes (HR). Under iron starvation condition, Fur released these regulations to respond to iron shortage and scavenge available iron.

Fur is also involved in DNA synthesis under iron starvation condition (Figure 26b). It is known that *E. coli* uses iron-dependent (NrdAB) or manganese-dependent (NrdEF) ribonucleotide reductase (Fe-RNR or Mn-RNR) to provide dexoyribonucleotide precursors for DNA synthesis under iron-replete or iron starvation condition, respectively[139]. Prior to this study, *nrdHIEF* operon did not have strong evidence for belonging to the Fur regulon. However, this study found that Fur regulated both *mntH* and *nrdHIEF*, which encode divalent metal cation transporter (preferentially Mn^{2+}) and Mn^{2+}-dependent ribonucleotide reductase system, respectively. Once the iron is scarce, MntH imports Mn^{2+} and NrdEF (Mn-RNR) to utilize this metal ion for DNA synthesis rather than NrdAB (Fe-RNR). The increased intracellular Mn^{2+} would also activate MntR (Mn-MntR) to repress RybA small RNA that down-regulates key genes in the aromatic amino acid biosynthesis pathway[140]. Thus, these series of regulations might increase precursor pools of enterobactin synthesis to efficiently scavenge iron under iron starvation condition.

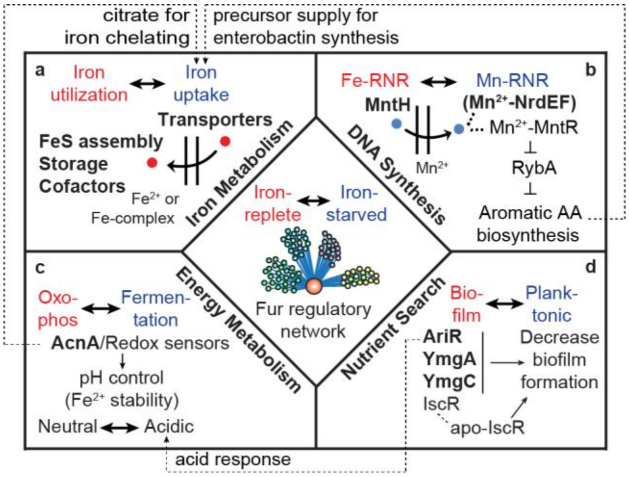


Figure 26. Global coordination roles of the Fur regulatory network in *E. coli*. The Fur regulatory network is involved in many cellular functions required in addition to iron acquisition and utilization. Fur directly regulates genes associated with (a) iron metabolism, (b) DNA synthesis, (c) redirection of metabolism toward fermentative pathways, and (d) biofilm formation for searching nutrients in response to iron availability. These networks are linked through the coordination role that Fur plays.

Another interesting role of Fur regulatory network is its involvement in energy metabolism by rapidly shifting metabolism between oxidative phosphorylation and fermentation in response to iron availability (Figure 26c). Our recently developed genome-scale model of metabolism and gene expression (ME-Model) in *E. coli*[141] predicted that iron starvation leads carbon flux to fermentative pathways rather than oxidative phosphorylation as shown in the previous study using *Staphylococcus aureus*[142] (Figure 27). Under iron starvation, down-regulation of *acnA* is achieved by indirect Fur regulation (mediated by RyhB) that results in temporarily shutting down the TCA cycle and releasing citrate for chelating iron[130]. Based on our results, Fur also activated the transcription of *acnA* under iron-replete condition, presumably to support active TCA cycle and oxidative phosphorylation for faster cell growth. This dual regulation of *acnA* by Fur might enable cells to rapidly change flux of the TCA cycle in response to iron availability. The metabolism shift might also dramatically change redox state (such as NADH/NAD$^+$ ratio) and thus, redox sensors including ArcA regulate a large number of genes associated with respiration[143]. Moreover, redirection of metabolism towards fermentative products would also decrease pH of the environment so that Fe^{2+} is slowly or not being oxidized to Fe^{3+}[142, 144].

Remarkably, Fur is also responsible for nutrient search of cells by controlling biofilm formation in response to iron availability (Figure 26d). It is known that iron starvation leads to the repression of biofilm formation in *E. coli* by *apo*-IscR[145]. The *apo*-Fur activation of *ymgA*, *ymgC*, and *ariR* would also suppress biofilm formation and enable planktonic growth of the cells to find iron-rich environment[133]. Our results also indicate that activation of *ariR* would contribute to expressing genes to endure acidic environments caused by shifting the metabolic state towards fermentative pathways under iron starvation condition.

Taken together, the primary role of Fur regulatory network is to maintain intracellular iron molecules within a narrow range of concentration so that their level can be relatively insensitive to changes in extracellular conditions. To accomplish this, Fur regulates genes associated with iron metabolism for iron uptake, storage, utilization, and FeS assembly. In the meantime, Fur is also directly involved in various fundamental cellular processes such as DNA synthesis, energy metabolism, and biofilm formation to allow cells to survive and adapt to the iron imbalance.

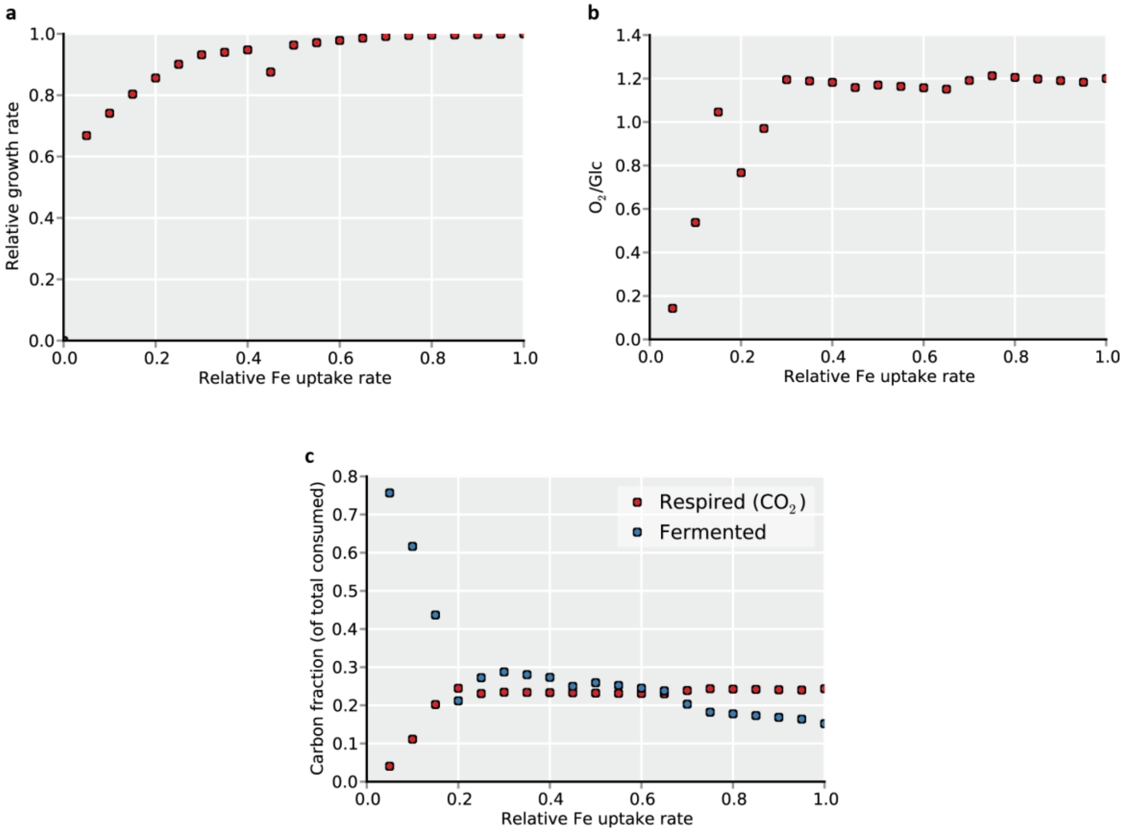


Figure 27. ME-simulation results depending on the iron uptake rate. From ME-model simulation, we predicted (a) relative growth rate, (b) O_2 uptake rate relative to glucose uptake, and (c) fraction of carbon secretion in response to different iron uptake rate. As iron is scarce, both growth rate and O_2 uptake rate are decreased. However, the fraction of fermented carbon dramatically increases while that of respired carbon decreases, indicating metabolism shifts towards fermentative pathways.

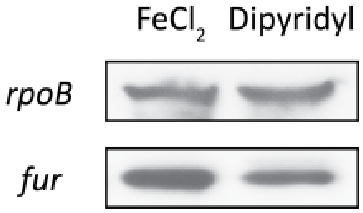


Figure 28. Western blot analysis. The *E. coli* strain harboring Fur-8myc was grown under iron-replete ($FeCl_2$) and iron starvation (Dipyridyl) conditions. Antibodies that specifically recognize myc-tag for Fur and RpoB subunit of RNAP were used.

Discussion

We comprehensively reconstructed the Fur regulon in *E. coli* by combining genome-scale Fur binding maps, RNAP binding/occupancy profiles, and transcriptome data under both iron-replete and iron starvation conditions. We identified (i) a total of 81 genes in 42 TUs directly regulated by Fur under either iron-replete (77 genes) or iron starvation (4 genes) conditions, (ii) regulatory modes for individual genes in the Fur regulon, (iii) the Fur regulatory feedback loop motifs composed of transporters and utilization enzymes in response to iron availability for iron metabolism, and (iv) the additional roles of Fur regulatory network beyond iron metabolism.

From the genome-wide Fur binding maps under both iron-replete and iron starvation conditions, we were able to show that a total of 119 Fur-binding sites were identified. The high number of Fur-binding sites was not surprising, considering the number of binding sites bound by global transcription factors such as ArcA, Crp, Fnr and Lrp specifically bind to the similar number of sites[16, 20, 23, 143]. Among the 119 binding sites, 54% (64 of 119) of them were located within regulatory regions, while the remaining 46% were found within non-regulatory regions such as intergenic or intragenic of two convergent genes (Figure 20). As expected, none of the Fur binding events within non-regulatory regions affected transcription. Surprisingly, 95% (55 of 58) of Fur bindings on the overlapped binding sites in both conditions were not responsible for regulation of transcription either (Supplementary Table S1). Given these binding properties of Fur, it is plausible that either there might be additional undiscovered function for Fur such as maintaining chromosome structure like H-NS and Fis in addition to a promoter-specific regulator[23, 146] or evolution has been slow to eliminate these non-functional DNA binding sites for Fur[147].

Combining Fur binding maps with transcriptome data and RNAP binding/occupancy profiles led us to identify direct Fur regulon and define 4 different regulatory modes of Fur based on the binding patterns of iron cofactor and the way to regulate the target genes (Figure 24). We provided strong evidences that 9 Y-genes (*yncE*, *yddA*-*yddB*, *ydiE*, *yqjH*, *ybaN*, *ynfD*, *yoeA*, and *yojI*) were directly regulated by HR mode. Given their putative functional annotations[128] and the fact that *E. coli* Fur directly controls transcription of transporters and utilization enzymes in a manner of (-/-) motif (HR/HA) for iron metabolism, we suggest that the functions of these genes might be involved in either iron/iron-complex transportation or iron-acquisition processes[129]. In addition, there have been accumulating evidences that Fur in other pathogenic species also have AA mode for transporters and AR mode for utilization enzymes (Supplementary Table S6). Given that the other (-/-) motif can also be implemented by AA/AR pairs (Figure 25b), these pathogenic species might have evolved to more sensitively respond to iron availability with additional iron acquisition/utilization regulatory system.

Comprehensive reconstruction of Fur regulatory network also extended our knowledge of the roles of Fur in response to iron availability. Beyond iron metabolism, Fur regulatory network was directly involved in other various biological processes such as DNA synthesis, energy metabolism, and nutrient search that are essential for cell survival (Figure 26). Remarkably, these networks were also connected to each other. For example, Mn^{2+} uptake (Figure 26b) and temporal TCA shutdown (Figure 26c), driven by Fur under iron starvation condition, consequently could lead to increase the intracellular Fe^{2+} pool, and Fur-mediated repression of biofilm formation (Figure 26d) could contribute to protecting cells from acidic environment caused by temporal TCA shutdown under this condition. Thus, disrupting Fur regulatory network in either iron-replete or iron starvation condition can lead to dramatic changes in transcript levels of genes associated with various stress responses (oxidative stress, redox imbalance, and acidic environment) besides Fur regulon (Supplementary Table S7 and S8); note that only 13% (81 of 678) of the Fur-dependent genes were directly regulated by Fur.

In summary, we have described an integrative analysis of various types of cutting-edge genome-scale experimental data and how this systems approach enabled us to comprehensively understand the complex roles of Fur regulatory network in E. coli. By combining ChIP-exo, which showed higher resolution and lower false-discovery rates than conventional ChIP-chip or ChIP-seq[24], with highly sensitive RNA-seq-based transcriptome analysis[148], we showed an unprecedented view into genome-wide binding of E. coli Fur and its regulon as well as complex regulatory network. In the future, the incorporation of these comprehensive operon structures that account for cellular regulation along with regulatory networks[3] into ME model would make it possible to mechanistically model and predict the complex regulatory interactions and thus, allow us to more accurately compute complex phenotypes[149].

Methods

Bacterial strains, primers, media, and growth conditions: All strains used are *E. coli* K-12 MG1655 and its derivatives. The *E. coli* strain harboring Fur-8myc was generated as described previously (Figure 28)[150]. Deletion mutant (Δ*fur*) was constructed by a λRed-mediated site-specific recombination system[151]. Glycerol stocks of *E. coli* strains were inoculated into M9 minimal medium (47.8 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.6 mM NaCl, 18.7 mM NH_4Cl, 2 mM $MgSO_4$, and 0.1 mM $CaCl_2$) supplemented with 0.2 % (w/v) glucose and cultured overnight at 37^oC with vigorous agitation. For iron-replete condition, the overnight cultures were inoculated into the same fresh M9 medium with 0.1 mM $FeCl_2$ and continued to culture at 37^{o}C with vigorous agitation to mid-log phase. For iron-depleted condition, the overnight cultures inoculated into the same fresh M9 media were supplemented with 0.2 mM 2, 2'-dipyridyl (DPD) at early-log phase and continued to culture at 37^{o}C for additional 2 h with vigorous agitation. For the rifampicin-treated cultures, the rifampicin dissolved in methanol was added to a final concentration of 150 g/mL at mid-log phase and stirred for 20 min.

ChIP-exo: To identify Fur- and RNAP-binding maps *in vivo*, we isolated the DNA bound to Fur protein and RNAP from formaldehyde cross-linked *E. coli* cells by chromatin immunoprecipitation (ChIP) with the specific antibodies that specifically recognizes myc tag (9E10, Santa Cruz Biotechnology) and RpoB subunit of RNAP (NT63, Neoclone), respectively, and Dynabeads Pan Mouse IgG magnetic beads (Invitrogen) followed by stringent washings as described previously[16]. ChIP materials (chromatin-beads) were used to perform on-bead enzymatic reactions of the ChIP-exo method[25] with following modifications. Briefly, the sheared DNA of chromatin-beads was repaired by the NEBNext End Repair Module (New England Biolabs) followed by the addition of a single dA overhang and ligation of the first adaptor (5'-phosphorylated) using dA-Tailing Module (New England Biolabs) and NEBNext Quick Ligation Module (New England Biolabs), respectively. Nick repair was performed by using PreCR Repair Mix (New England Biolabs). Lambda exonuclease- and RecJ$_{f}$ exonuclease-treated chromatin was eluted from the beads and the protein-DNA cross-link was reversed by overnight incubation at 65^{o}C. RNAs- and Proteins-removed DNA samples were used to perform primer extension and second adaptor ligation with following modifications. The DNA samples incubated for primer extension as described previously[25] were treated with dA-Tailing Module (New England Biolabs) and NEBNext Quick Ligation Module (New England Biolabs) for second adaptor ligation. The DNA sample purified by GeneRead Size Selection Kit (Qiagen) was enriched by polymerase chain reaction (PCR) using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The amplified DNA samples were purified again by GeneRead Size Selection Kit (Qiagen) and quantified using Qubit dsDNA HS Assay Kit (Life Technologies). Quality of the DNA sample was checked by running Agilent High Sensitivity DNA Kit using Agilent 2100 Bioanalyzer (Agilent) before sequenced using MiSeq (Illumina) in accordance with the manufacturer's instructions. Each modified step was also performed in accordance with the manufacturer's instructions.

RNA-seq: Total RNA including small RNAs was isolated using the cells treated with RNAprotect Bacteria Reagent (Qiagen) followed by purification using Qiagen RNeasy Plus Mini Kit (Qiagen) in accordance with the manufacturer's instruction. Samples were then quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific) and quality of the isolated RNA was checked by running RNA 6000 Pico Kit using Agilent 2100 Bioanalyzer (Agilent). Paired-end, strand-specific RNA-seq was performed using the dUTP method[148] with the following modifications. The ribosomal RNAs were removed with Ribo-Zero rRNA Removal Kit (Epicentre). Subtracted RNA was fragmented for 2.5 min using RNA Fragmentation Reagents (Ambion). cDNA was generated using SuperScript III First-Strand Synthesis protocol (Invitrogen) with random hexamer priming. The samples were sequenced using MiSeq (Illumina) in accordance with the manufacturer's instructions.

Data analysis: To identify enriched sites in the ChIP-exo data, Illumina sequencing reads were mapped to reference genome (NC_000913) and analyzed for peak calling by using MACE program (https://code.google.com/p/chip-exo/). Signal-to-noise ratio was calculated by setting the noise level as the value at the top 5% of entire signals. The Fur-binding motif analysis was completed using the MEME tool from the MEME software suite[152]. We extended the sequence of each binding site by 20 bp at each end to allow for adjacent sequences to be included in the analysis. We used default settings except for the width parameter fixed at 21 bp. For RNA-seq data analysis, we used Cufflinks/Cuffdiff[153] to identify differentially expressed genes with log2 fold change > 0.5 and a false discovery rate (FDR) value < 0.01. Genome-scale data were visualized using NimbleGen's SignalMap software.

Western blot analysis: Soluble cell lysates were subjected to electrophoresis in a NuPAGE Novex 10% Bis-Tris Gel (Invitrogen) with MOPS running buffer, and the resolved proteins were transferred to a Hybond^{TM}-ECL membrane (Amersham Biosciences) using XCell II Blot Module (Invitrogen). The ECL^{TM} Western detection kit (Amersham Biosciences), antibodies that specifically recognizes myc tag (9E10, Santa Cruz Biotechnology) and RpoB subunit of RNAP (NT63, Neoclone), respectively, and horseradish peroxidase (HRP)-conjugated sheep anti-mouse immunoglobulin G (IgG) (Amersham Biosciences) were used to detect the proteins. The Qubit Protein Assay Kit (Invitrogen) was used to measure the amount of total proteins in the lysates in order to load same amount of proteins in each lane. All experiments were performed in accordance with the manufacturer's instructions.

Acknowledgements

Chapter 5, in total, is a reprint of the material as it appears in Seo, S.W.*, Kim, D.*****, O'Brien, E., Latif, H., Szubin, R., Palsson, B.Ø. Deciphering Fur regulatory network highlights its role in iron homeostasis of *Escherichia coli*. *Submitted*. I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 6: Elucidating transcriptional regulation of nitrogen metabolism with systems approaches

Demystifying transcription regulatory network (TRN) in bacteria is important to understand metabolic flexibility and robustness in response to environmental changes[154]. The most popular way to elucidate TRN at the systems level is integrating multiple 'omics' datasets, such as ChIP (chromatin immunoprecipitation), expression profiling and genome-scale TSS profiling. The first step of such experiments is deciding a relevant growth condition where a transcription factor (TF) of interest is expected to be active. Deciding experimental conditions for TF activation has been frequently based on information from the literature. This approach has been working well, especially for very specific or locally working TFs, which has one or a few activation conditions. However, there could be multiple activation conditions for global TFs, and in many cases it is difficult to assess which condition is better than the others.

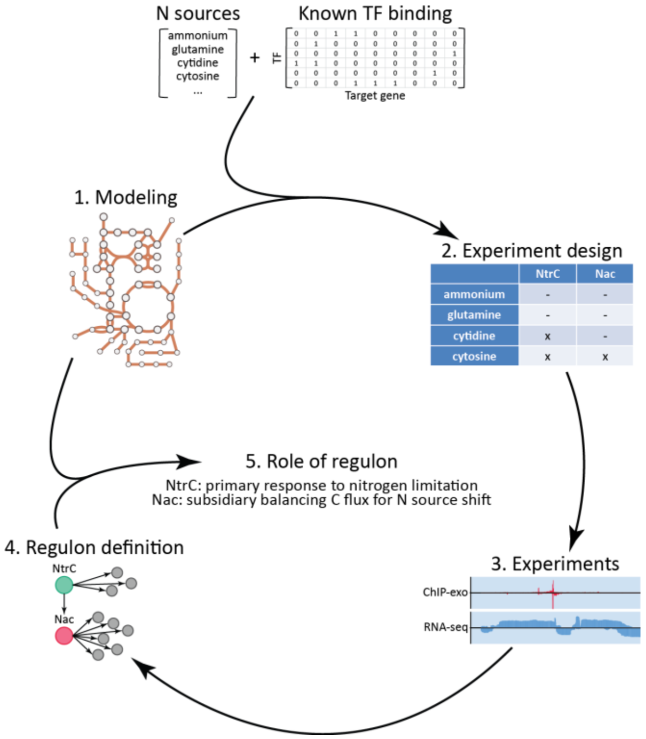
Development of cellular network models including recently developed models of metabolic network[155] and of metabolism and expression[141] enabled exploration in unexperimented solution spaces of *E. coli* in virtually every imaginable conditions. Comparison of different network states generated under different conditions can shed lights on differential cellular response to the environmental change rendered in parameterization of simulation. For instance, transactional change predicted with ME model[141] can present a cellular response in the transcription level to the environmental signal, thus combining simulations from that model with known, but limited, regulon information from the public database can result in prediction of conditions where a TF could be activated, but never experimentally validated.

In order to exploit this possibility, two major TFs in response to nitrogen limitation, NtrC and Nac[156-158], were chosen for the model-driven experimental design, because nitrogen metabolism is one of key parts in *E. coli* metabolism, however *in vivo* binding of those TFs has not been investigated at the genome-scale. There are similar studies for carbon metabolism, for aerobic/anaerobic metabolism[19-21], and for other TFs sensing nitrogen containing molecules, such as Lrp[16], ArgR/TrpR[17], PurR[18]. Thus, systems approach of model-based decision of experimental condition was applied for two major nitrogen-responsive TFs, NtrC and Nac. Cutting-edge ChIP-exo (chromatin immunoprecipitation with exonuclease treatment)[159] and RNA-seq was performed to reconstruct NtrC and Nac regulons, and further analysis was conducted to elucidate distinct roles of those regulons.

Model-driven prediction of activation conditions for TFs

Using a model of metabolism and expression (ME model) in *E. coli*[141], growth on glucose and different viable nitrogen sources in the model was simulated. Then predicted expression of genes in the model for each nitrogen source was compared with predicted expression for growth on ammonia to find a set of predicted differentially expressed genes. Previously annotated regulons for NtrC and Nac from the public database[160] were used to calculate regulons enriched in the differential set of genes for each alternative nitrogen source (Figure 29).

For NtrC, cytidine and cytosine were predicted to be nitrogen sources that would change the expression of genes that NtrC regulons were statistically enriched in. Thus NtrC was predicted to be activated under those conditions. For Nac, cytosine, but not cytidine, was predicted to make Nac activated. In addition to cytidine and cytosine, ammonia was tested for a negative control which has been known not to activate those TFs, and glutamine was also included as a positive control that was considered to activate them[156] although the model predicted otherwise.



	NtrC	Nac
ammonium	-	-
glutamine	-	-
cytidine	x	-
cytosine	x	x

NtrC: primary response to nitrogen limitation
Nac: subsidiary balancing C flux for N source shift

Figure 29. Flowchart of model-driven experimental design and following near 1-bp resolution experiments to reconstruct NtrC and Nac regulons, and further analysis to elucidate functional roles of those regulons.

On ammonia, and 3 other alternative nitrogen sources, glutamine, cytidine, and cytosine, a series of experiments including expression profiling and ChIP experiments were performed to confirm activation of TFs under the given conditions, and to experimentally measure expression change and *in vivo* TF bindings onto the *E. coli* genome. From the experimental measurements, expanded definition of NtrC and Nac regulons was defined, which were used with model simulation to identify roles of these regulons in responding to nitrogen limitation.

Experimental confirmation of predicted conditions

Activation of a transcription factor means a series of multiple events, including transcriptional activation of a gene encoding the transcription factor, increasing translation from the transcript resulting in more protein, and regulation of target genes by binding onto genomic locations. First, to see if there is up-regulation of *ntrC* and *nac* transcription, RNA-seq was performed with E. coli K-12 MG1655 cells grown on ammonium, glutamine, cytidine and cytosine up to mid-log phase. *amtB*, which encodes an ammonia transporter, was reported to be up-regulated when glutamine was supplemented as a sole nitrogen source[156], thus transcription level of *amtB*, *ntrC*, *nac* and 3 subunits of RNA polymerase (RNAP), *rpoN*, *rpoD*, and *rpoB*, was compared between ammonium and 3 other alternative nitrogen sources (Figure 30). In all alternative nitrogen sources, transcription of *amtB*, *ntrC* and *nac* was significantly up-regulated, while no significant change in transcription level of RNAP subunits was observed. In expansion of comparison, transcription level of 4,595 annotated genes in E. coli genome was analyzed to see how broad the response to nitrogen limiting condition is in terms of transcriptional change (Figure 31A). Using alternative nitrogen sources, glutamine, cytidine and cytosine, resulted in 667, 390, and 690 differentially expressed genes respectively, and expression of 1046 genes (22.8%) was changed under at least one condition, leaving 3548 (77.2%) not changed in all conditions. This number is much bigger than the previous report[161], where about 100 genes were claimed to be of the nitrogen regulated (Ntr) response.

To make sure that up-regulation in transcription of *ntrC* and *nac* resulted in increasing amount of protein, western blotting was performed to investigate protein level of NtrC, Nac, RpoN and RpoD (Figure 31B). In agreement with expression profiling with RNA-seq, protein expression of NtrC and Nac increased in glutamine, cytidine, and cytosine, although the amount of Nac on cytidine was much less than on glutamine and cytosine. This could explain why the number of differentially expressed genes on cytidine was the least among three conditions.

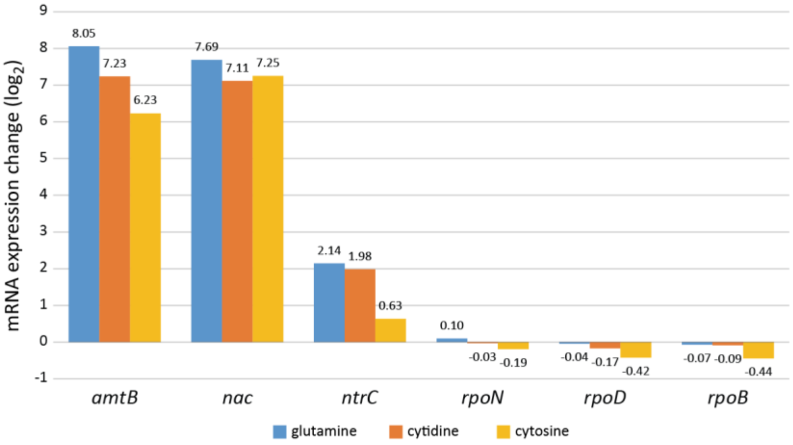


Figure 30. Expression changes of key enzymes under alternative nitrogen sources.

The final component of transcription factor activation is its binding onto genomic locations, regulating expression of target genes. Experimental measurement of TF binding onto genome was accomplished by performing recently developed ChIP-exo (Chromatin immunoprecipitation with exonuclease treatment) by adopting the original protocol[159], but modifying it for bacterial use. In total, 19, 249, 153 and 2171 binding sites across conditions were identified for NtrC, Nac, RpoN, and RpoD respectively (Figure 31C). Number of binding sites for two σ-factors, RpoN and RpoD, did not change much, whereas binding sites of NtrC and Nac increased on all 3 alternative nitrogen sources. For instance, bindings of NtrC increased from 5 to 19, and Nac bindings increased from 15 to > 240.

In summary, transcriptional expression of key components in response to nitrogen limiting conditions was up-regulated, which was followed by increase in protein amounts of those TFs. Direct measurement of TF bindings *in vivo* in a genome-scale manner showed increasing bindings of those two TFs onto genome. Thus, alternative nitrogen sources, glutamine, cytidine and cytosine, rendered nitrogen limiting conditions, and activated NtrC and Nac in transcriptional, translational, and binding activity levels.

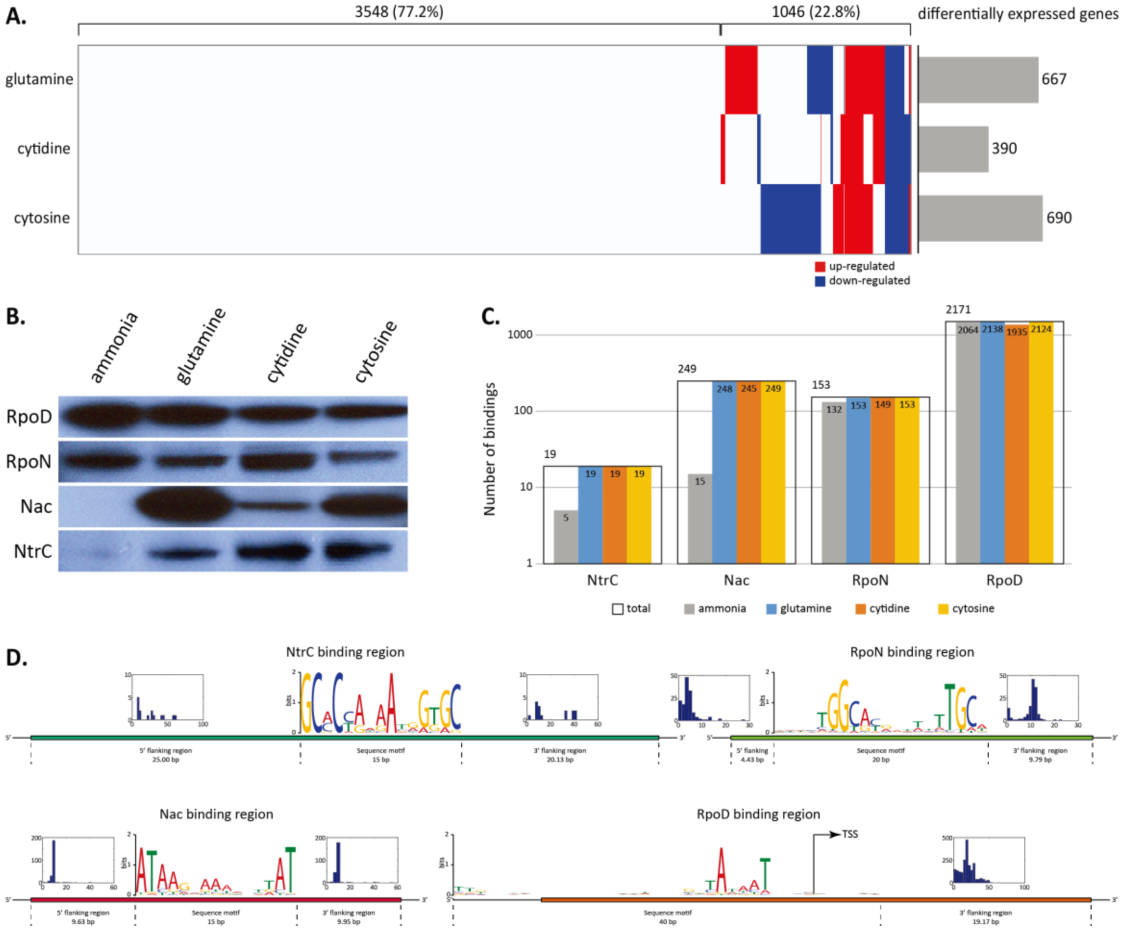


Figure 31. Experimental confirmation of activation conditions for NtrC and Nac.

Advantages of ChIP-exo over previous methods

In brief, ChIP-exo applies a 5’-3’ strand-specific exonuclease to a chromatin immunoprecipitated sample. Deep sequencing of an exonuclease-treated ChIP sample enables detection of exonuclease stop sites with near 1-bp resolution[159]. ChIP-exo method was claimed to have a better resolution and sensitivity, so that detection of much narrow peaks and peaks with weak bindings could be identified, which was not possible with ChIP-chip or ChIP-seq. To illustrate the resolution improvement that ChIP-exo method presents, ChIP experiment results for RpoD from the same condition with 3 different methods, ChIP-chip[3], ChIP-seq, and ChIP-exo, were compared to show binding signals upstream of *rpsU-dnaG-rpoD* and *ileX* operons. Bindings detected with three methods aligned well near the center (Figure 32). However, ChIP-exo definitely presented the best resolution, although ChIP-seq showed a better resolution over ChIP-chip.

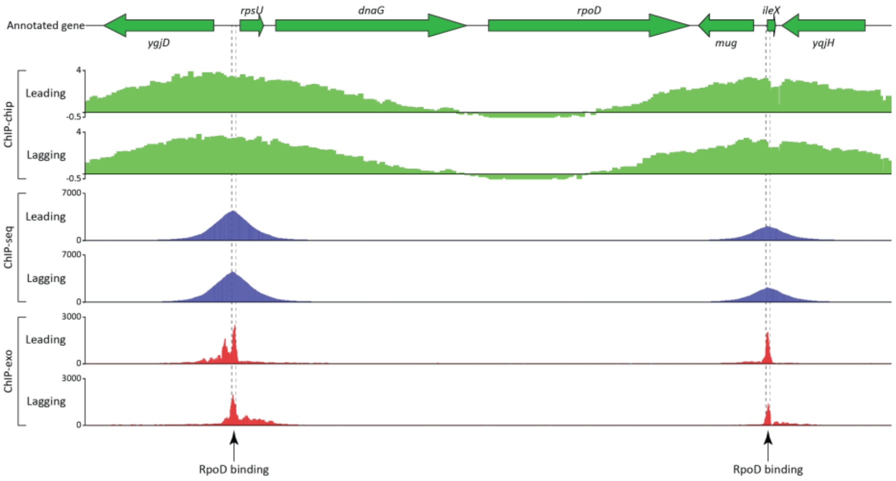


Figure 32. Comparison of 3 ChIP methods: ChIP-chip, ChIP-seq, and ChIP-exo for RpoD.

Binding peaks detected for NtrC, Nac, RpoN, and RpoD, with ChIP-exo had average widths of 57.2, 34.6, 33.7, and 50.1 bp with low variation in binding widths (Figure 33). Another advantage in ChIP-exo method is strong bindings are most reflected in peak height (number of reads), and less associated with broadness of peaks, while ChIP-chip and ChIP-seq methods have a tendency of showing broader peaks for stronger bindings. This property helps with locating where the genomic sequences that TF recognizes and binds on lie with improved resolution (Figure 32).

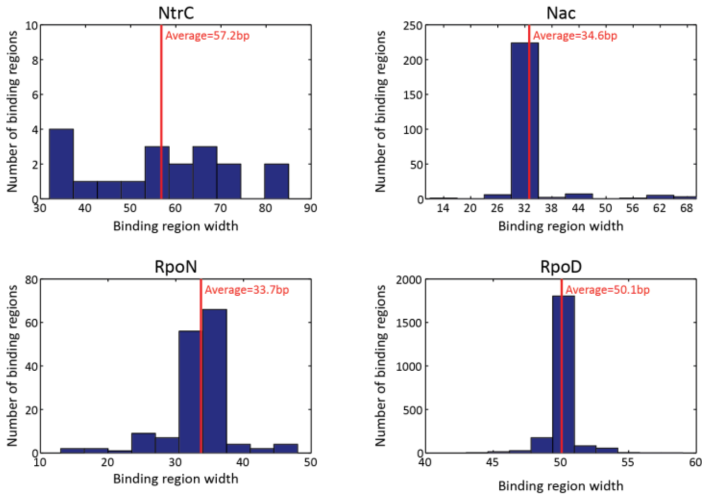


Figure 33. Distribution of binding region widths of NtrC, Nac, RpoN and RpoD.

Sharpness of ChIP-exo binding peaks makes it possible, for the first time, to detect bindings upstream of sRNAs. In *E. coli* K-12 MG1655, there are about 81 annotated sRNAs. Widths of them range from 53 to 436 bp with the average of 137.1 bp. Binding peaks from two previous methods are much wider than the sRNA length, and in many cases sRNAs are located in the vicinity of neighboring genes. Thus it was technically difficult to distinguish bindings for sRNAs. With an improved resolution, ChIP-exo overcomes this issue, enabling direct measurement of protein binding upstream of sRNAs. For instance, *spf* is 109 bp long, and binding of Nac and RpoD was identified by ChIP-exo method (Figure 34). Binding peak of RpoD was found near the 5’ end of *spf* gene, and was clearly separated from Nac binding peak. These observations were further supported with sequence motif analysis (Figure 34).

Thus, ChIP-exo evidently shows advantages over two long-established ChIP methods. Outpacing resolution and sensitivity of this method contributed to more accurate annotation of binding regions of TFs and σ-factors in this study, further to spatial pattern between TF and associated σ-factors, which is discussed in detail later.

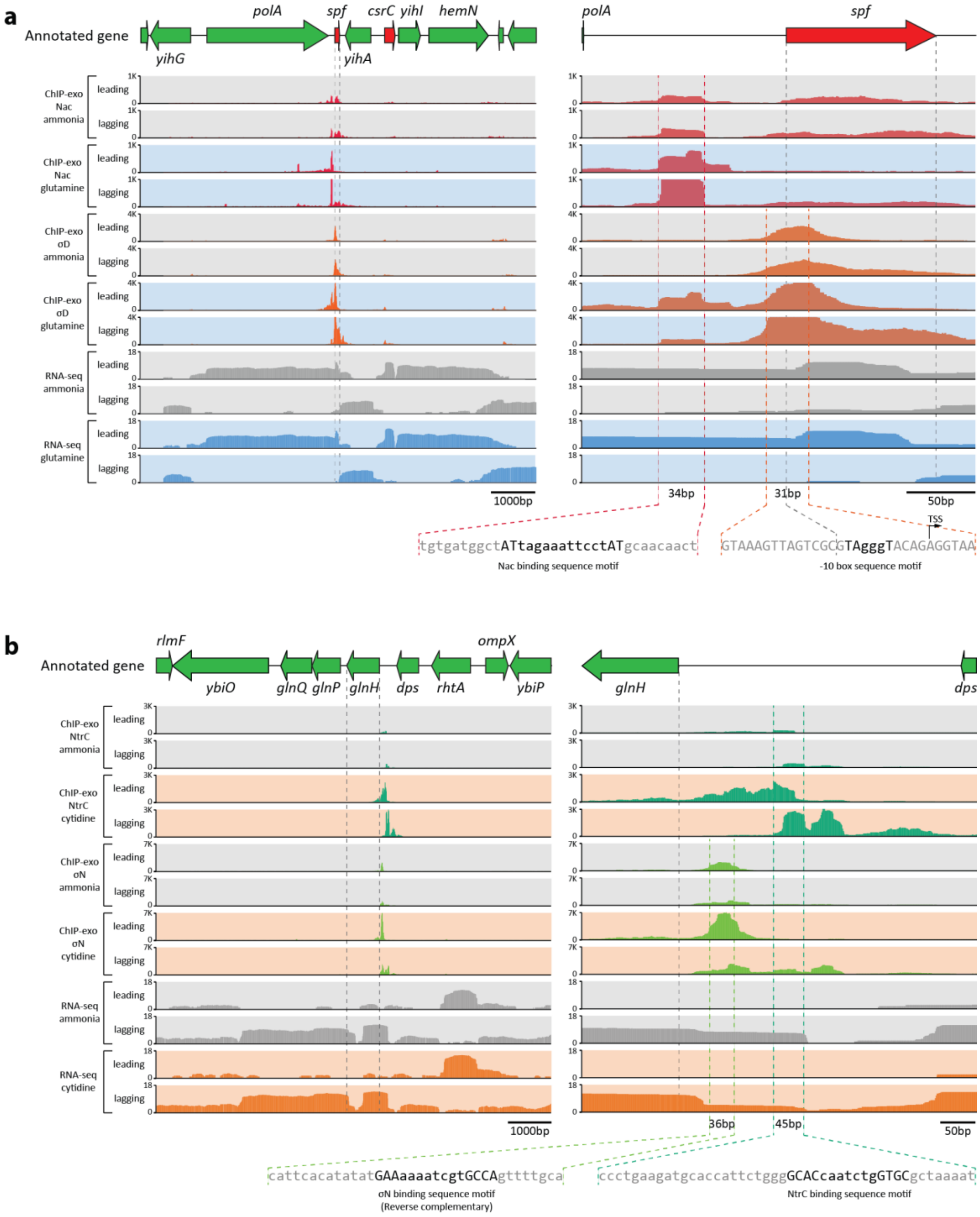


Figure 34. Examples of bindings for NtrC with RpoN and Nac with RpoD.

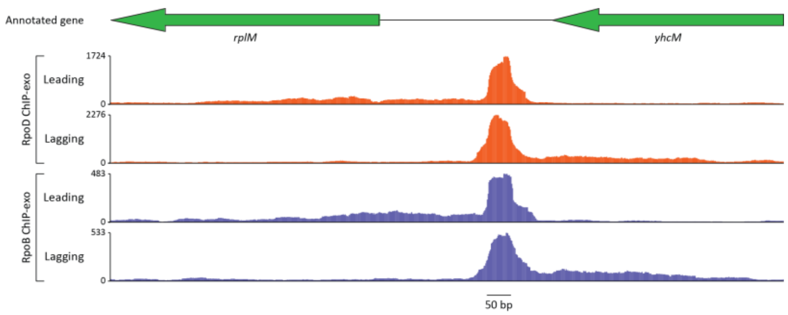


Figure 35. ChIP-exo binding of RpoD aligned with binding of RpoB.

Sequence motif analysis of binding regions

Since binding peaks detected with ChIP-exo method were so narrow, it became of interest if a sequence motif is found from those peak regions, and where the sequence motif lies inside the regions. To address these questions, MEME software[162] was used to retrieve sequence motifs. The sequence motifs from NtrC, Nac, RpoN, and RpoD binding sites were GCaCcaaaAtgGtGC, tGGcacgattttTGCa, ATAagnaaaanttAT, and ttgaca-15bp-gntAtaaT (lower-case characters indicate an information content <1 bit). These motifs were identical to the known motifs[3, 160, 163, 164]. Except for RpoD motif, sequence motifs located near the center of and inside binding regions (Figure 31D). For RpoD, only -10 box, gntAtaaT, was found inside the binding regions. This observation conflicts with the knowledge of RpoD, because RpoD is known to specifically recognize -10 and -35 boxes sequences, which are expected to be protected from exonuclease activity. RpoD binding peaks align well with RpoB binding peaks (Figure 35) and transcription start sites (TSSs) were dominantly located at the center of binding regions; ChIP may be capturing RpoB that are associated with RpoD.

In short, sequence motifs, which TFs presumably bind onto, located near the middle of binding regions and were identical to previously reported ones. Thus comparison of bindings of the same TF under different conditions is capable of detecting binding activation upstream of coding genes and sRNAs, further hinting functions of NtrC and Nac.

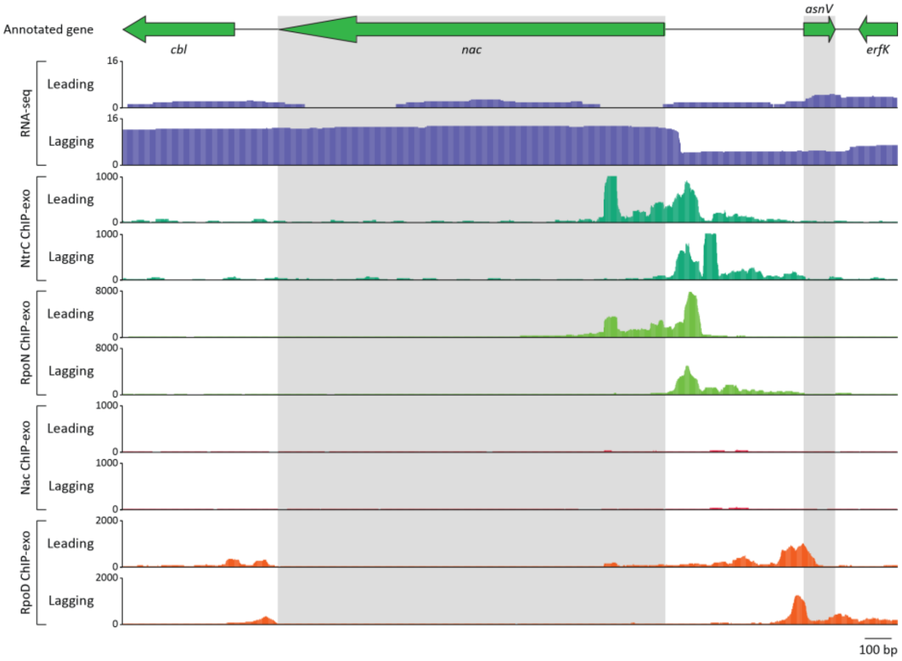


Figure 36. No binding of Nac upstream of nac identified by ChIP-exo dataset.

Confirmation of binding sites with comparison to known sites

From 19 and 249 total binding sites for NtrC and Nac, 16 and 247 binding sites were found upstream of genes, hence called as regulatory binding sites. These binding sites were compared to 4 and 3 known binding sites for NtrC and Nac. For NtrC, all of known sites upstream of *glnL*, *glnA*, *glnH*, and *astC*[165-168] were detected in the dataset of this study. Similarly, 2 of 3 known Nac binding sites, upstream of *codB*, and *ydcS*[169, 170], were detected, however Nac binding upstream of *nac*[160] was not. Nac binding for *nac* was claimed based on evidences from *K. aerogenes* and sequence alignment *nac* promoter regions between *E. coli* and *K. aerogenes*[158]. However, no binding was detected from ChIP-exo dataset (Figure 36), and *nac* does not have RpoD promoter, thus it may be more likely that Nac does not regulate *nac* in *E. coli*.

In sum, NtrC and Nac binding sites identified in this study cover all known binding sites, with expanding the current knowledge by 375% and 8,200%, respectively (Figure 37A).

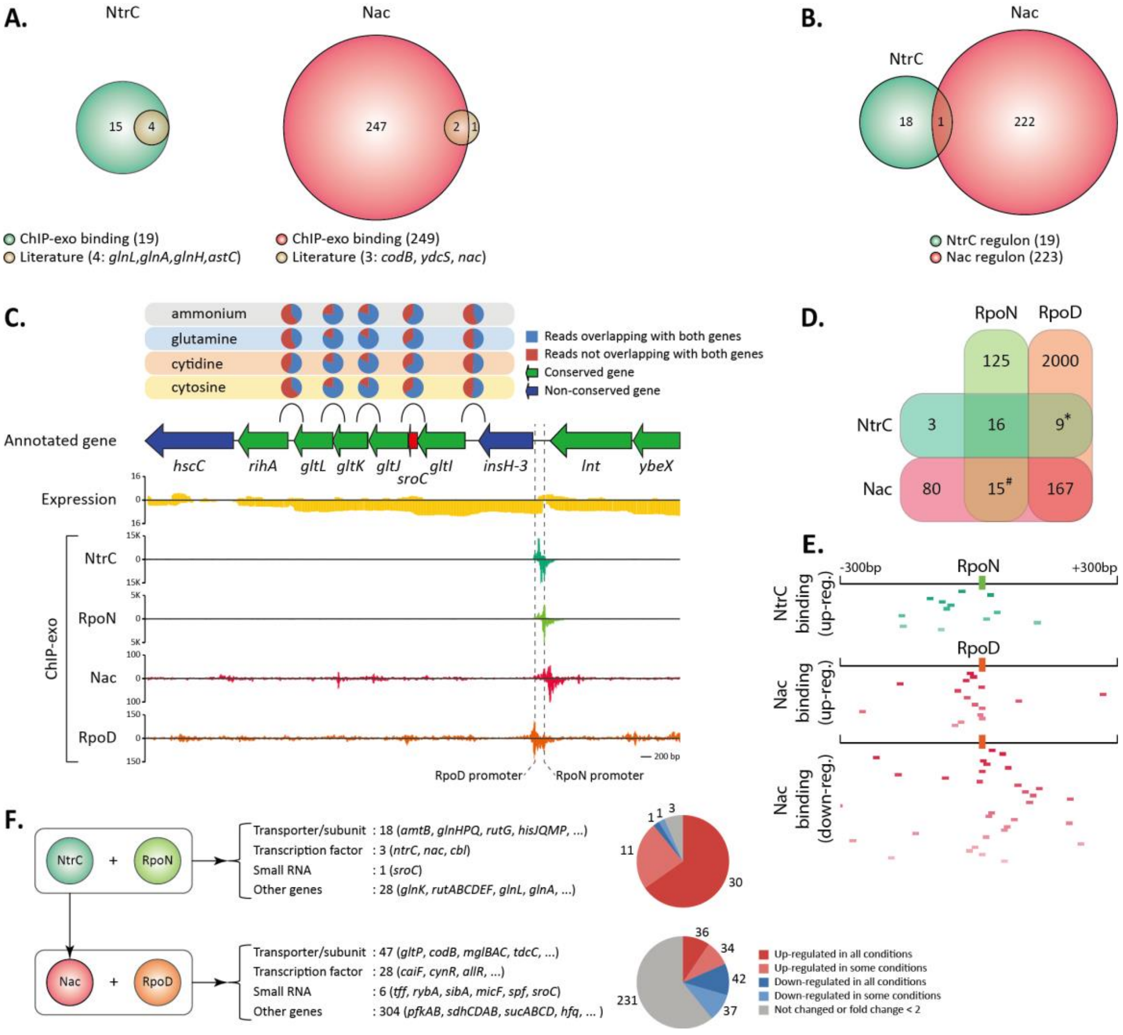


Figure 37. Reconstruction of NtrC and Nac reulgons.

Reconstruction of NtrC and Nac regulons

Although post-translational regulation on glutamine synthetase (*glnA*) by GlnD, GlnK, GlnB have been extensively studied[171-178], limited information of *in vivo* binding of those TFs has been capping knowledge on regulation of nitrogen metabolism in transcriptional level. To shed light on transcription regulation by NtrC and Nac, regulons of them were reconstructed by associating TF bindings with transcription units[3, 11]. From 3181 TUs covering 4485 genes (97.6% of annotated genes), 19 TUs were associated with NtrC and 223 TUs were with Nac.

Another interesting aspect from regulon reconstruction is that two regulons barely overlap with each other, except for 1 TU which is *insH-3* (Figure 37B). In the latest definitions of TUs[3, 11, 160], transposon-related *insH-3* was annotated to make one TU on its own. However, RNA-seq profiling with paired-end reads suggested there might be a longer transcript starting from *insH-3*, possibly including *gltI-sroC-gltJKL* (Figure 37C). Confirming this possibility, 49.3% of sequence reads covering intergenic region between *insH-3* and *gltI* overlapped both with insH-3 and *gltI*, indicating cotranscription of those two genes. Similar approach was applied to the downstream 6 genes, and there seems the longer TU has at least 6 genes from *insH-3* to *gltL*. Thus, two promoters with RpoD and RpoN upstream of *insH-3* contribute to transcription of glutamate/aspartate ABC transporter (*gltIJKL*), and binding of NtrC and Nac are associated with those σ-factors (Figure 37C). Transcriptomic expression of the longer TU was up-regulated under nitrogen-limiting conditions by NtrC and/or Nac, resulting in increased glutamate/aspartate transporters as a scavenging mechanism.

Transcriptomic comparison to *K. pneumoniae* genome[7] gives more insights on TU organization and conservation (Figure 38). In *K. pneumoniae*, *gltI-sroC-gltJKL* and two upstream coding genes, *lnt* and *ybeX*, are all conserved, and transcription starts upstream of *gltI*. However, *insH-3* is not found in *K. pneumoniae* genome. Thus, TU of *gltI-sroC-gltJKL* is conserved, but *insH-3* got into 5' UTR of this TU in *E. coli* K-12 MG1655.

Thus, reconstruction of regulons presented scant overlapping between two nitrogen-limiting responsive regulons, cluing distinct roles in response to nitrogen limiting condition and in nitrogen metabolism.

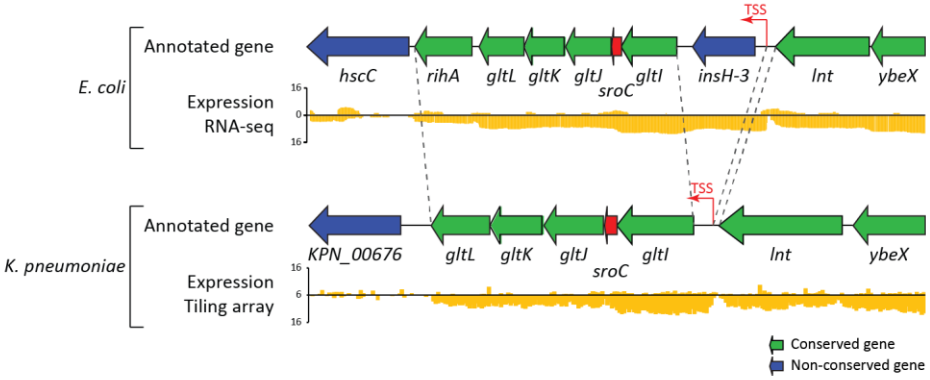


Figure 38. Comparison of TU organization containing *gltIJKL* operon between *E. coli* and *K. pneumoniae*.

Association of TF with σ-factor

Definition of two TF regulons raised a following question which σ-factor associates with each TUs in regulons. NtrC belongs to the RpoN-dependent activator family[179] and interacts with RpoN through adjuvant DNA-binding protein[180], thus its binding was expected to be accompanied with RpoN binding. Nac is postulated to serve as an adapter between NtrC and final RpoD-dependent promoters[156], so its binding would appear in the vicinity of RpoD binding regions. This observation has not been confirmed with *in vivo* measurement of TF and σ-factor binding, thus it is still an open question how many RpoN-dependent or RpoD-dependent promoters are associated with NtrC or Nac, or if there is any directionality of TF binding to σ-factor binding and how far their bindings align.

From ChIP-exo dataset and calculation of closed-located bindings from the dataset, NtrC binding was associated with RpoN binding, and Nac with RpoD (Figure 37D). Out of 19 NtrC binding sites, 16 were found with RpoN binding near them upstream of the same gene, 9 of which were from complicated promoters with RpoD and RpoN promoters. For instance, *glnA* has a distal RpoD-dependent promoter (glnAp1) and a proximal RpoN-dependent promoter (glnAp2)[160], which were captured from ChIP dataset (Figure 39). Upstream regulatory region of *insH-3-gltI-sroC-gltJKL* makes another example (Figure 37C). NtrC binding was found upstream of RpoN-dependent promoter, whereas no Nac binding was observed near RpoD-dependent promoter. This also exemplifies an extensive overlapping between RpoD regulon and RpoN regulon [3]. The majority of Nac bindings (167, 67.1%) adjoined RpoD bindings, while 15 bindings were found in promoters having both of RpoD and RpoN bindings (Figure 37C is one example). NtrC works dominantly as a transcription activator on RpoN-dependent promoter by binding upstream of promoter in many cases (Figure 37E). Nac works as a dual regulator on RpoD-dependent promoter, and it more binds upstream of promoter when up-regulating the downstream gene, whereas it binds downstream of promoter when down-regulating (Figure 37E).

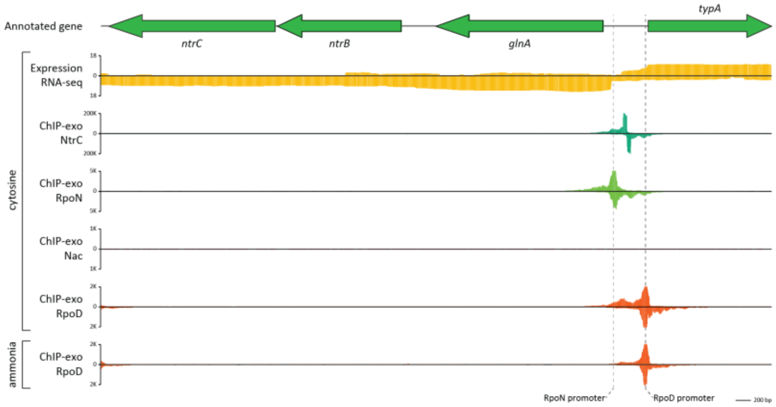


Figure 39. Complicated promoter structure upstream of *glnA*, and binding patterns of NtrC, Nac, RpoD and RpoN in that region.

Thus, NtrC binds in the vicinity where RpoN binds, while Nac does near RpoD. There are multiple promoter regions that have RpoN and RpoD-dependent promoters, however NtrC and Nac bind separately except for 1 region, indicating distinct roles of two TFs.

Contrasting roles of NtrC and Nac regulons

In addition to a bare overlap between NtrC and Nac regulons, functional analysis of regulons sheds more light on distinct roles of NtrC and Nac in response to nitrogen limitation. NtrC up-regulated 41 genes (30 genes under all alternative nitrogen sources, and 11 genes in some conditions), and down-regulated 2 genes (*yeaE* in all conditions, and *mipA* only in cytosine), leaving 3 genes not changed or changed less than 2 folds (Figure 37F). NtrC regulon contains 18 transporters or their subunits, 3 TFs, 1 sRNA and 28 other enzymes. Transporters are for mostly nitrogen sources including ammonia (*amtB*), glutamine (*glnHPQ*), glutamate (*gltIJKL*), histidine (*hisJ*), lysine/arginine (*hisQMP*, *argT*), xanthine/uracil (*rutG*), and others are less characterized (*yhdWXYZ*). NtrC up-regulates regulatory proteins too, including 3 TFs, *ntrC* itself, *nac*, and *cbl* and 2 post-translational regulatory proteins (*glnK*, and *glnL*). The role of *cbl* in nitrogen response still needs more investigation; however it is obvious that NtrC regulates major regulatory enzymes, responding to nitrogen limitation. Metabolic enzymes that NtrC regulates catalyze reactions for nitrogen-containing molecules including glutamine (*glnA*), pyrimidine (*rutABCDEF*), arginine (*astCADBE*), and D-alanyl-D-alanine (*ddpXABCDF*).

While NtrC regulates a smaller set of genes and mostly activates the expression of target genes, Nac regulates a larger group of genes and works as a dual regulator by up-regulating 70 genes and down-regulating 79 genes (Figure 37F). Another difference is NtrC regulates nitrogen-related regulatory proteins, transporters, and metabolic enzymes, Nac regulon covers beyond nitrogen-related enzymes. For instance, Nac binds upstream of *gltP* (glutamate/aspartate transporter), *codB* (cytosine transporter) however it also binds upstream of carbon source transporters such as *mglBAC* (galactose ABC transporter). Nac regulon includes a number of mostly locally acting TFs, some of which are known to be related carbon metabolism or in both carbon/nitrogen metabolism, such as *cynR* (cyanate binding transcriptional activator) *csiR* (carbon starvation induced regulator), *sfsB* (maltose metabolism related regulator), *gutM* (glucitol regulator), *ebgR* (evolved β-galactosidase repressor), *tdcA* (threonine and serine transcriptional regulator), *deoR* (deoxyribose regulator), *allR* (allantoin repressor), *caiF* (carnitine regulator), *lrp* (leucine-responsive regulatory protein), *lysR* (lysine regulator), *feaR* (phenylethylamine regulator), *xapR* (xanthosine/deoxyinosine transcriptional regulator), *asnC* (asparagine regulator), and *metR* (methionine biosynthesis related regulator). Moreover, interestingly Nac regulates some of key genes in glycolysis and TCA (Tricarboxylic acid) cycle; phosphofructokinase (*pfkA*, and *pfkB*), citrate synthase (*gltA*), succinate dehydrogenase (*sdhCDAB*), 2-oxoglutarate dehydrogenase (*sucAB*), and succinyl-CoA synthetase (*sucCD*). COG analysis of NtrC and Nac regulons showed genes for amino acid metabolism and signal transduction are more enriched in NtrC regulon, while Nac regulon has genes functionally enriched in energy production and amino acid metabolism (Figure 40).

In summary, NtrC with RpoN regulates TFs, transporters, and metabolic enzymes for responding to nitrogen-limiting condition. However, Nac in a company with RpoD is responsible for regulating a broader set of genes beyond nitrogen-related. Thus, NtrC and Nac have contrasting roles in response to nitrogen-limitation; however the role of Nac regulon in the response seems less obvious than NtrC regulon and the linkage between NtrC regulon and Nac regulon still needs more elaboration.

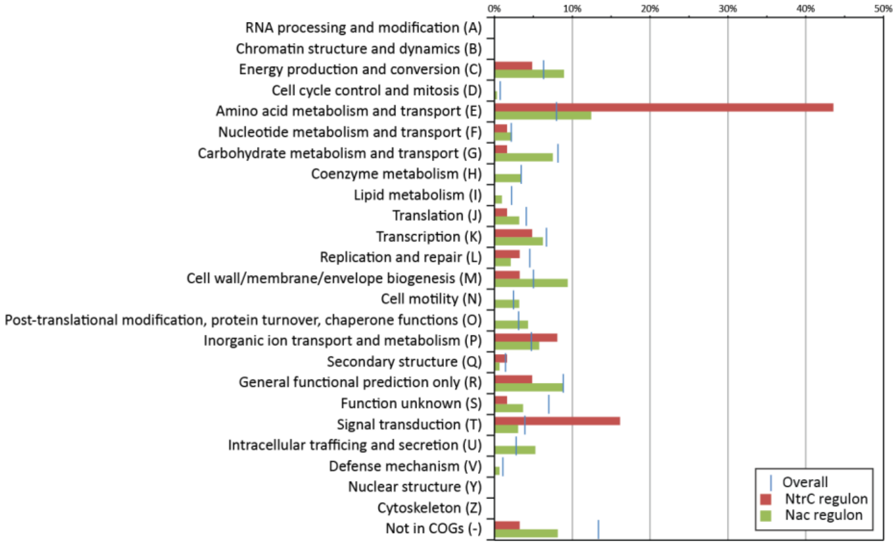
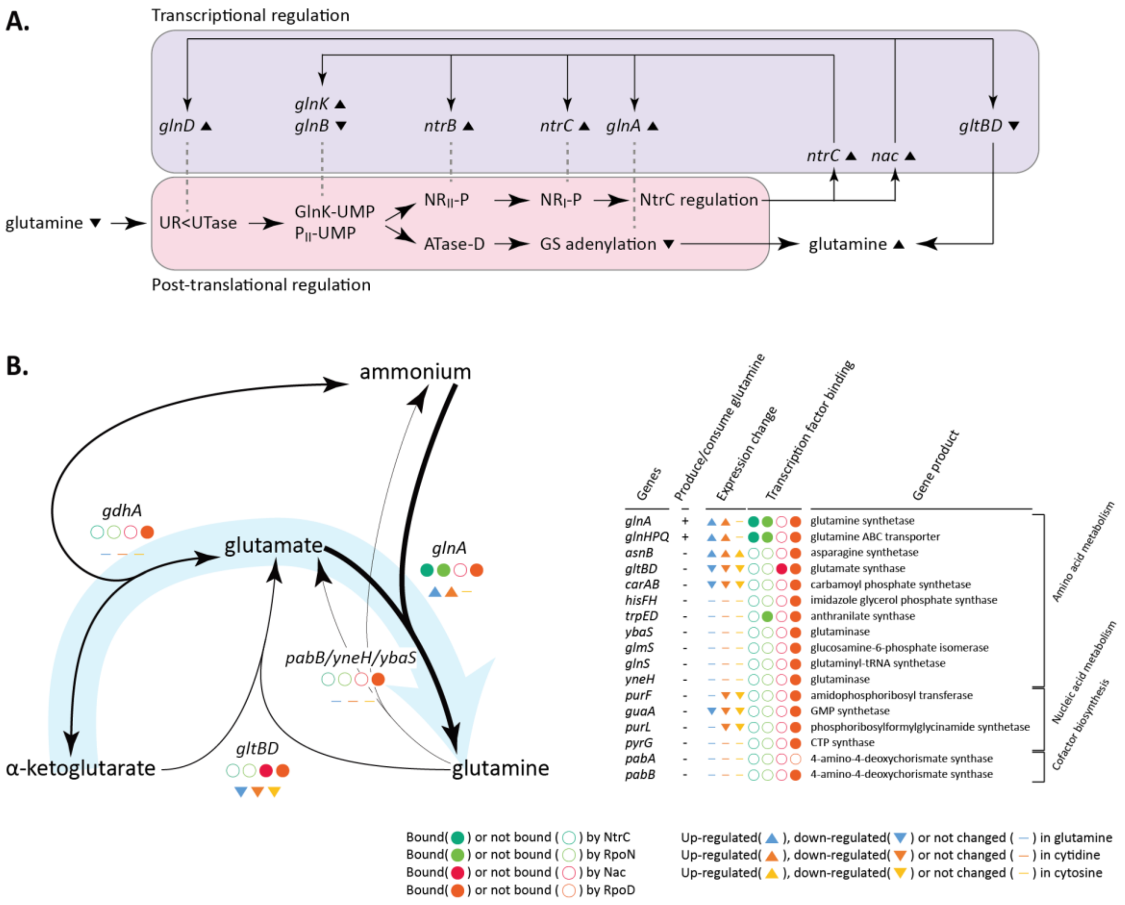


Figure 40. Functional analysis of NtrC and Nac regulons.

Primary response by NtrC regulon

Glutamine is the central molecule with which *E. coli* cells sense nitrogen-limiting condition[181], and the mechanism of Ntr regulatory cascade from sensing the low level of glutamine to activation of NR$_I$ with phosphorylation is relay of post-translational regulation (Figure 41A) and has been extensively studied[161]. In brief, low glutamine stimulates UTase (uridylyl-transferase) activity of GlnD, which is a single peptide with UTase and UR (uridylyl-removing) activities, by binding to a single site on the enzyme[182]. UTase activity of GlnD uridylylates two functionally redundant two proteins, P$_{II}$ and GlnK. Uridylylated P$_{II}$ and GlnK fail to interact with NR$_{II}$, which results in net phosphorylation of NR$_I$[183]. Phosphorylated NR$_I$ is an active form, activating transcription of NtrC regulon genes. Uridylylation of P$_{II}$ and possibly GlnK stimulates the deadenylylating activity of ATase (glutamine synthetase adenylyltransferase/deadenylase, *glnE*), and GlnE deadenylates glutamine synthetase (*glnA*) making it an active form[184]. As a result of this cascade, the internal level of glutamine can go up.


A.
Transcriptional regulation
glnD ▲
glnK ▲
glnB ▼
ntrB ▲
ntrC ▲
glnA ▲
ntrC ▲
nac ▲
gltBD ▼
glutamine ▼
UR<UTase
GlnK-UMP
P_{II}-UMP
NR_{II}-P
NR_{I}-P
NtrC regulation
ATase-D
GS adenylation ▼
glutamine ▲
Post-translational regulation
B.
ammonium
gdhA
glutamate
glnA
pabB/yneH/ybaS
α-ketoglutarate
gltBD
glutamine
Bound(●) or not bound (○) by NtrC
Bound(●) or not bound (○) by RpoN
Bound(●) or not bound (○) by Nac
Bound(●) or not bound (○) by RpoD
Up-regulated(▲), down-regulated(▼) or not changed (—) in glutamine
Up-regulated(▲), down-regulated(▼) or not changed (—) in cytidine
Up-regulated(▲), down-regulated(▼) or not changed (—) in cytosine

Genes	Produce/consume glutamine	Expression change			Transcription factor binding				Gene product	
glnA	+	▲	▲	—	●	●	○	●	glutamine synthetase	Amino acid metabolism
glnHPQ	+	▲	▲	—	●	●	○	●	glutamine ABC transporter	
asnB	-	▲	▲	▲	○	○	○	●	asparagine synthetase	
gltBD	-	▼	▼	▼	○	○	●	●	glutamate synthase	
carAB	-	▼	▼	▼	○	○	○	●	carbamoyl phosphate synthetase	
hisFH	-	—	—	—	○	○	○	●	imidazole glycerol phosphate synthase	
trpED	-	—	—	—	○	●	○	●	anthranilate synthase	
ybaS	-	—	—	—	○	○	○	●	glutaminase	
glmS	-	—	—	—	○	○	○	●	glucosamine-6-phosphate isomerase	
glnS	-	—	—	—	○	○	○	●	glutaminyl-tRNA synthetase	
yneH	-	—	—	—	○	○	○	●	glutaminase	
purF	-	—	▼	▼	○	○	○	●	amidophosphoribosyl transferase	Nucleic acid metabolism
guaA	-	▼	▼	▼	○	○	○	●	GMP synthetase	
purL	-	—	▼	▼	○	○	○	●	phosphoribosylformylglycinamide synthetase	
pyrG	-	—	—	—	○	○	○	●	CTP synthase	
pabA	-	—	—	—	○	○	○	○	4-amino-4-deoxychorismate synthase	Cofactor biosynthesis
pabB	-	—	—	—	○	○	○	●	4-amino-4-deoxychorismate synthase	

Figure 41. Transcriptional regulation of Ntr regulatory cascade by NtrC.

Unlike post-translational regulation of Ntr regulatory cascade, transcriptional level of this regulation has been less studied. In this regulatory cascade consisted of 10 genes, 8 genes are regulated either by NtrC or Nac. 5 genes, *ntrC*, *nac*, *ntrB*, *glnA* and *glnK*, are up-regulated by NtrC, 1 gene, *glnD*, was up-regulated by Nac, and 2 genes, *gltB* and *gltD*, were down-regulated by Nac, while 2 genes, *glnB* and *glnE*, are not regulated by them (Figure 41A, Figure 42). Thus, NtrC and Nac regulate the majority of regulatory components in this cascade, making multiple positive-forward loops. These loops make a complicated network with well-characterized network motifs including coherent type 1 feed-forward loop (C1-FFL) and positive auto-regulation (PAR). C1-FFL is a frequent motif found in *E. coli*[185] and has a function of a sign-sensitive delay element and a persistence detector[186], and PAR shows the slower response time than simple regulation and may lead to a bimodal distribution of protein level[187]. These properties may contribute to filtering out short signals of nitrogen limitation, rendering a response with a short delay for persistent signals, and shutting off the output fast when nitrogen-limiting condition is relieved.

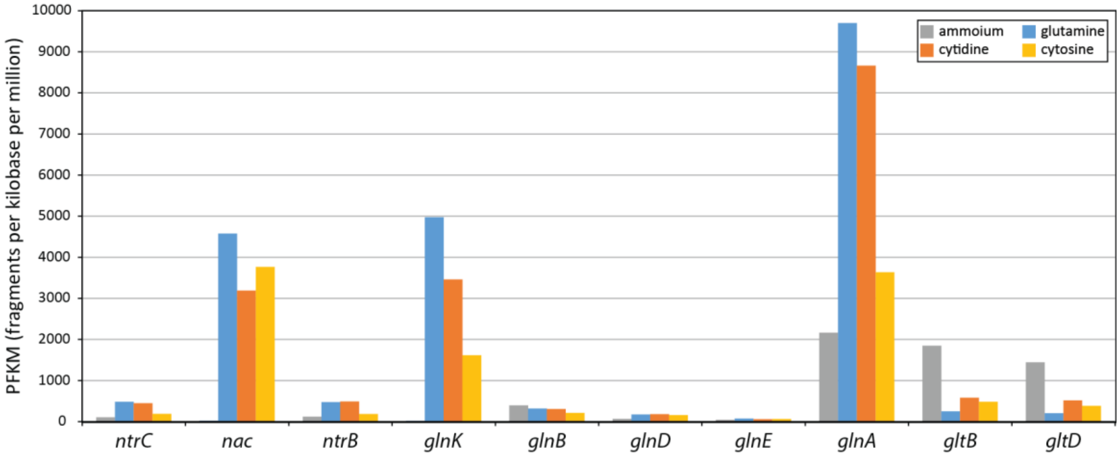


Figure 42. Expression change of genes in Ntr regulatory cascade under alternative nitrogen sources.

In *E. coli*, cytoplasmic glutamine is either synthesized from glutamate by glutamate synthetase (*glnA*) or is transported by glutamine ABC transporter (*glnHPQ*) (Figure 41B). Both operons were up-regulated by NtrC and its associated RpoN-dependent promoters. Other than asparagine synthetase B (*asnB*), all genes that consume glutamine was not changed (fold change < 2) or down-regulated, indicating that *E. coli* cells change the abundance of metabolic machineries towards increasing intracellular glutamine level. In *E. coli*, glutamate can be built up from α-ketoglutarate in two reactions. One is by glutamate dehydrogenase, which is encoded by *gdhA* and the gene has an RpoD-dependent promoter for constitutive expression and expression of *gdhA* did not change significantly upon alternative nitrogen sources. The other is by glutamate synthase, which is encoded by *gltBD*, and the operon also has an RpoD-dependent promoter, however Nac negatively regulates expression of *gltBD* upon alternative nitrogen sources.

Thus, securing enough glutamine pool under nitrogen limitation necessitates expression changes of regulatory and metabolic enzymes in Ntr regulatory cascade, and NtrC plays a central role in this complicated regulation. In addition to transcriptional regulation in the cascade, NtrC also activates transporters for favorable nitrogen sources as a scavenging mechanism, and induces expression of the other key TF, *nac*. Moreover, it becomes of interest how *E. coli* cells manage production and consumption of α-ketoglutarate under nitrogen limitation, because production of intracellular glutamine requires expense of α-ketoglutarate, which is one of key molecules in carbon metabolism.

Carbon flux rebalancing by Nac regulon

Since α-ketoglutarate is one of substrates in TCA cycle, glycolysis and TCA cycle pathways were analyzed in term of TF and σ-factor bindings and expression change (Figure 43A). Genes in those pathways were transcribed from RpoD-dependent promoters, and surprisingly 11 genes of them were regulated by Nac. Nac repressed expression of genes in TCA cycle, *pck*, *sucAB*, *sucCD*, and *sdhBADC*. Nac also repressed glycolysis genes, *pfkA*, *pfkB*, *fbaA*, and *ppc*, but the expression fold change was less than 2. Nac did not bind upstream of genes in PPP (pentose phosphate pathway), except for ones that work in glycolysis and PPP at the same time.

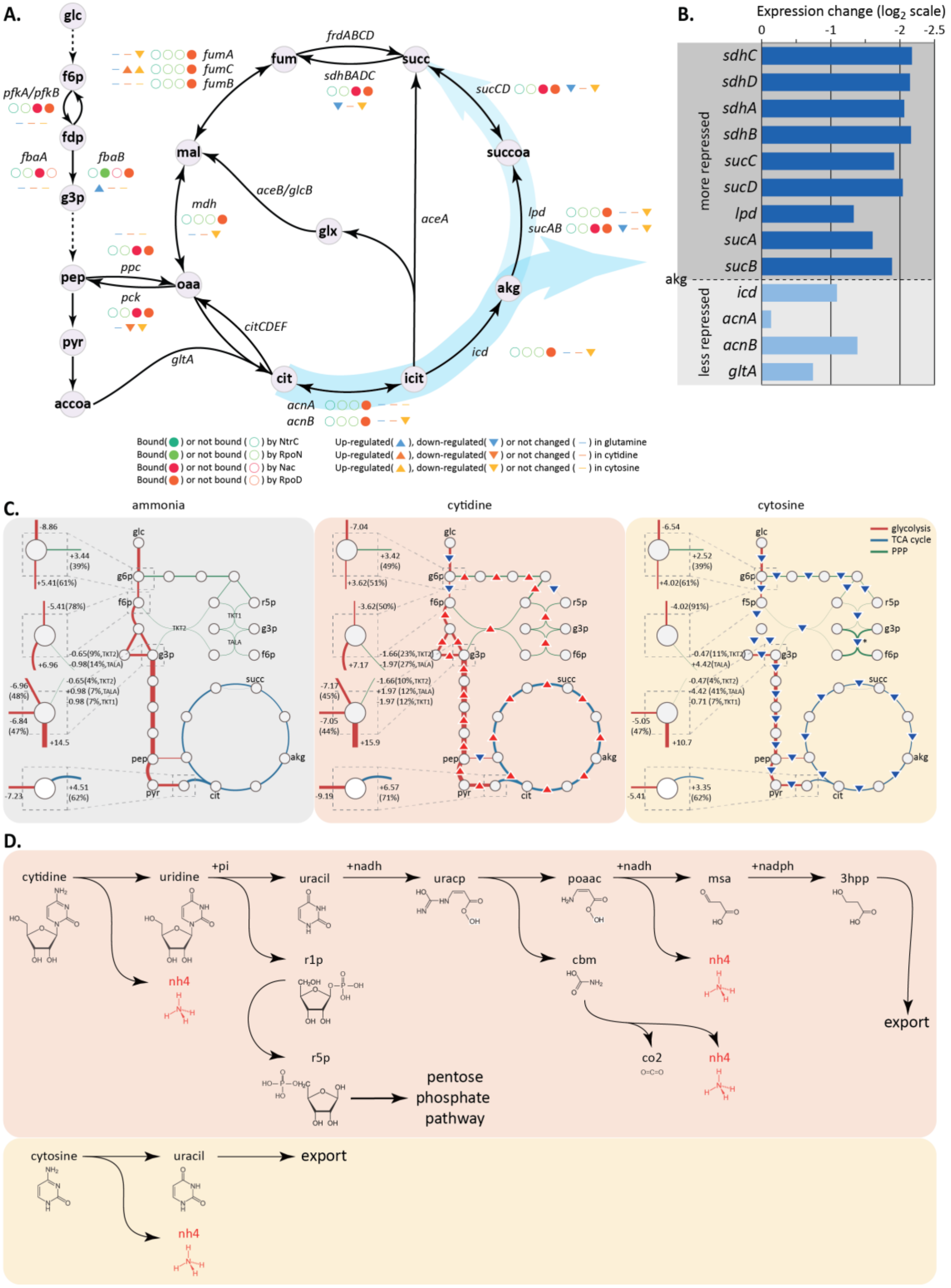


Figure 43. Carbon flux rebalancing by Nac in response to nitrogen source shift.

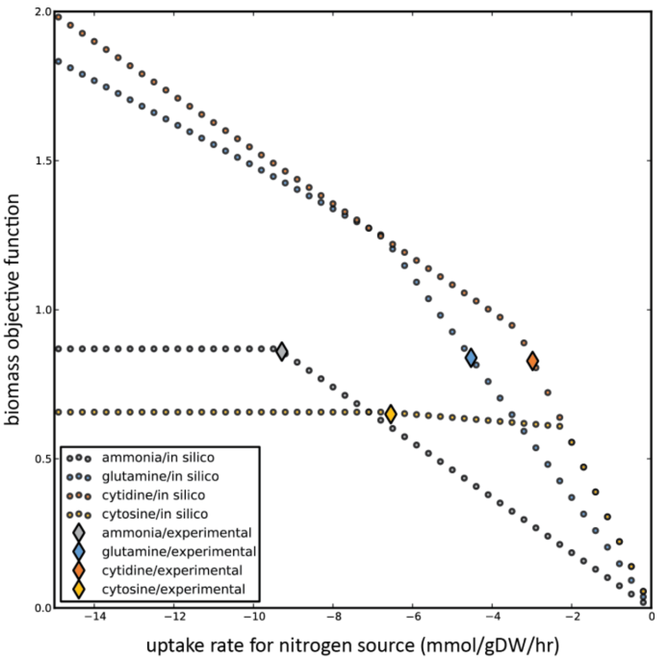


Figure 44. FBA analysis with *E. coli* metabolic model with varying nitrogen source uptake rates.

Among genes in TCA cycle, genes downstream of α-ketoglutarate were bound by Nac, and ones upstream of that molecule were not. It was postulated that genes encoding enzymes downstream of α-ketoglutarate would be more repressed for two reasons. First, Nac works as a repressor on enzymes in this pathway (Figure 43A). Second, cells are under nitrogen-limitting stress, which directly constrains growth *in silico* simulation (Figure 44), and α-ketoglutarate is a precursor for glutamine synthesis. As postulated, all downstream genes, *sdhBADC*, *sucCD*, *lpd*, and *sucAB*, were more down-regulated than upstream genes, *icd*, *acnAB*, and *gltA* when on cytosine (Figure 43B). Degree of repression on glutamine and cytidine was less than on cytidine; however two alternative nitrogen sources also showed the same pattern (Figure 45).

In sum, response to nitrogen limitation accompanies changes in gene expression of enzymes in TCA cycle to rebalance carbon flux towards facilitating glutamine pool maintenance. Nac plays a role in this process.

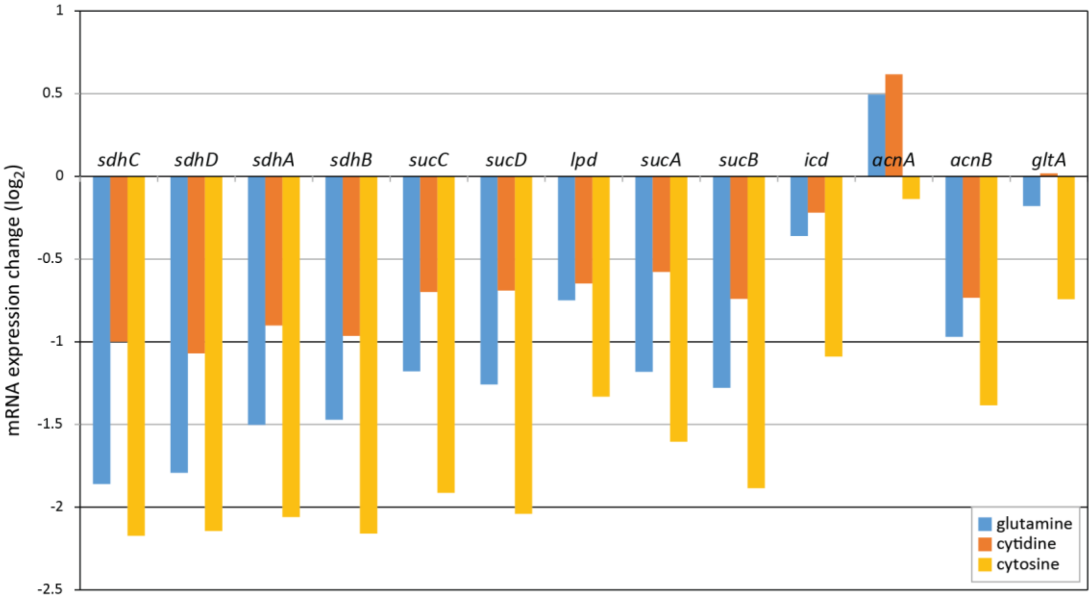


Figure 45. Expression change of gene in TCA cycle near α-ketoglutarate.

Less activation of Nac on cytidine

NtrC regulon has a primary responsibility of responding to nitrogen limiting environment, while Nac regulon has a subsidiary role of rebalancing carbon metabolism for nitrogen source shift. The remaining question is cytidine rendered nitrogen-limiting condition, mRNA expression of *nac* was up-regulated (Figure 30), and key enzymes in glutamine synthesis were up-regulated on cytidine (Figure 41B). However protein expression of it on cytidine was much less compared to on glutamine and cytosine (Figure 31B), the number of differentially expressed genes on cytidine was significantly less than on the other two alternative nitrogen sources (Figure 31A), and expression change of metabolic genes in TCA cycle on cytidine was the least among the three nitrogen sources (Figure 43B, Figure 45).

To explain this discrepancy, FBA (flux balance analysis) with *E. coli* metabolic model[155] and experimentally measured glucose uptake rate was performed to see internal flux states under different nitrogen sources (Figure 43C, Figure 44). On ammonia, glucose uptake rate was 8.86 mmol/gDW/hr, and 39% of g6p (glucose 6-phosphate) went into PPP, leaving 61% remaining in glycolysis pathway. Flux distribution through g6p on cytosine is same as ammonia, but with lower glucose uptake rate of 6.54 mmol/gDW/hr. On cytidine, 7.04 mmol/gDW/hr flux into g6p was split differently, more flux towards PPP (49%) and less flux to downstream glycolysis (51%). Interestingly, however, most fluxes through PPP increased, and as a result, flux from f5p (fructose 5 phosphate) and all fluxes downstream increased. These raised flux in glycolysis pushed more flux through TCA cycle, too (Figure 43C). Thus, compared to on ammonia, there is less or no need to repress activities through TCA cycle, which in turn requires less repression of genes of enzymes. This means there is less need for Nac to repress genes, which potentially explains why less activation of Nac was observed on cytidine.

Further analysis gives insights into how increased flux through PPP was possible on cytidine (Figure 43D). Flux from cytidine uptake went into cytidine deaminase (*cdd*) reaction to make uridine and ammonium. Uridine breaks into uracil and ribose 1-phosphate (r1p) by uridine phosphorylase (*udp*). This r1p is converted to r5p, which can go into PPP. To accommodate more r5p, there should be more xylulose 5-phosphate (xu5p), which explains more flux into PPP from g6p on cytidine. This analysis also enables better understanding on how differently uracil could be used when cytidine or cytosine is a sole nitrogen source (Figure 43D). Uracil still has 2 nitrogen molecules, however in order to harvest them, 2 molecules of NADH+ and 1 of NADPH+ are required. Flux analysis showed cells could fully assimilate nitrogen from cytidine by utilizing more energy, which could be a part of more activated glycolysis and TCA cycle on cytidine. However, on cytosine, the flux model predicted that cells would take one nitrogen molecule from cytosine and export uracil out of cell, which is less efficient but requires less energy.

Estimating fluxes through metabolic reaction with using *E. coli* model and FBA analysis under different nitrogen sources enabled explanation of seemingly contradictory observation on less activation of Nac. This was because ribose part of cytidine could be utilized to make more energy and result in more activated glycolysis and TCA cycle. Also this analysis gave insights into how cells could decide how many nitrogen molecules from uracil depending on the energy availability in the cell.

Acknowledgements

Chapter 6, in total, is a reprint of the material as it appears in Kim, D., Ebrahim, A., Seo, S.W., Bordbar, A., Palsson, B.Ø. Elucidating transcriptional regulation of nitrogen metabolism with systems approaches. *In preparation.* I was the primary author, while the co-authors participated in the research that served as the basis for this study.

Chapter 7: Towards a detailed understanding of bacterial transcription regulation with systems approaches

At the core of phenotype-genotype relationship, there are multiple levels of gene expression and gene regulation, among which arguably transcription regulation, more specifically regulation of transcription initiation, is the most important and efficient regulatory point. In this dissertation, multiple regulatory components including TSS, σ-factors and associated promoter regions, and TFs are assessed at the genome-scale.

Increasing needs for bioinformatic tools.

As it becomes easier and cheaper to generate sequencing-based data, data analysis, not data generation, becomes the rate-limiting step in genomics and transcriptomics studies[28]. Thus, there has been an increasing need for data visualization tools that facilitate analysis by enabling researchers to explore, interpret, and manipulate the multiple –omics datasets, and in cases to perform computations based on the datasets[28]. There are a number of visualization and analysis software, however most of them are developed on Java and often fail to accommodate multiple –omics datasets at the same time or have some performance issues. MetaScope was implemented with C# and overcomes these problems by allowing manipulation of large datasets over giga bytes, thus it has been extensively used in all studies in the dissertation, and other studies in the research group. However, MetaScope has limited capability for on-the-fly computation, leaving rooms for upgrades for implementing a mechanism to allow user-defined functions or procedures by inter-operability to python or R programming languages.

Comparative systems biology on regulatory elements by genome-wide TSS profiling

TSS profiling of two enterobacteria, *E. coli* and *K. pneumoniae*, expanded the understanding on properties of bacterial regulatory elements. First, genome-wide identification of TSSs with modified 5' RACE with deep-sequencing or TSS-seq provided comprehensive sets of TSSs in those species. Analysis on those TSS datasets showed *E. coli* and *K. pneumoniae* share common regulatory elements including usage of multiple TSSs, 5' UTR length, Shine-Dalgarno sequence motif, promoter sequence motif, and so on. However, comparison of TSSs and promoters defined by those TSSs upstream of orthologous genes presented different organization of regulatory upstream regions, indicating different gene regulation between two closely related organisms. Further analysis on small RNAs in a comparative perspective showed that there is a tendency that more important regulatory elements such as binding sites in small RNAs are more conserved than the other regions. This observation seems not limited to only small RNAs, sequence conservation analysis of known TF binding sites in *E. coli* and mapping them onto *K. pneumoniae* genomic sequence dictated possibilities of conservation and disruption of TF binding sites for many TFs (Figure 46). Adding more experimental evidences onto the computational prediction is expected to give more insights on how microorganisms manage to adapt and evolve not just only by evolving gene contents, but also by modifying regulatory components.

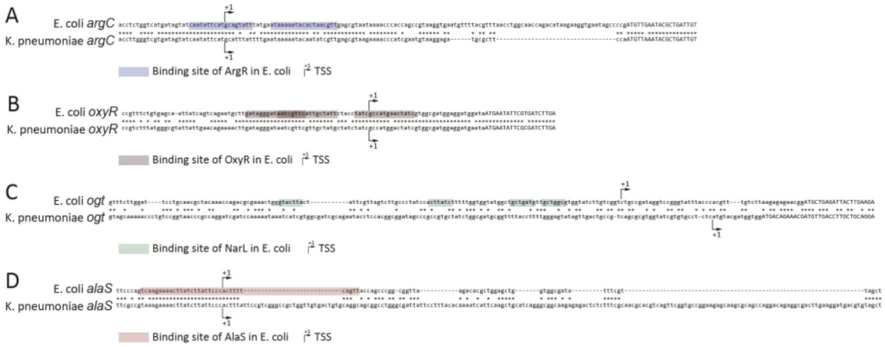


Figure 46. Conservation and disruption of TF binding sites between *E. coli* and *K. pneumoniae*.

Genome-scale reconstruction of σ-factor network in E. coli

Reconstruction of σ-factor network provides a valuable resource, from which new biological findings could be extracted. In order to reconstruct σ-factor network, genome-wide identification of RNAP and σ-factor bindings were obtained by performing ChIP-chip experiments under multiple conditions, resulting in enumeration of active promoter regions and σ-factor association with those promoter regions which is holoenzyme-binding region map. This holoenzyme map was integrated with strand-specific TSS information to build promoter map. In turn, this promoter map was combined with known TU information[11]. The reconstruction of σ-factor network was first used to show interaction between two major σ-factors: σ^{70} and σ^{38} that work in competitive mode to render negative regulation by σ^{38} under stationary phase in *E. coli*. In addition, σ-factor binding in *E. coli* was compared to *K. pneumoniae*, and it was found that TU organization and σ-factor regulation onto some TUs are different, indicating quite diverse regulatory mechanisms on the conserved gene contents (Figure 47), which expands the observation found in comparison of TSSs between those two microorganisms.

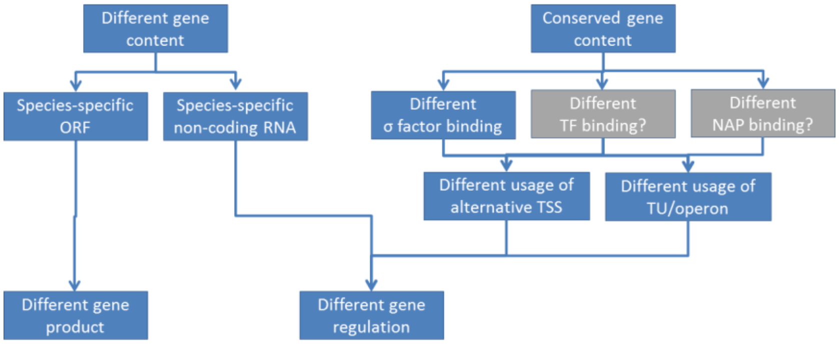

Different gene content
Conserved gene content
Species-specific ORF
Species-specific non-coding RNA
Different σ factor binding
Different TF binding?
Different NAP binding?
Different usage of alternative TSS
Different usage of TU/operon
Different gene product
Different gene regulation

Figure 47. Various mechanisms for different regulation on conserved gene contents.

Systems determination of Fur transcriptional regulatory network

Elucidation of Fur regulon exemplifies usage of near 1-bp resolution ChIP-exo protocol and reveals its play in iron regulation. In order to reconstruct Fur regulon, ChIP-exo experiments with and without iron, and RNA-seq experiments on WT and Δ*fur* with and without iron were performed. Based on these experimental measurements, Fur regulon was reconstructed and target genes were categorized into 4 groups by activation/repression and binding with iron. No case for *apo*-Fur repression (AR) was found, but 4 genes in one TU were categorized as *apo*-Fur activation (AA). More genes were to be in *holo*-Fur activation (HA) mode, while *holo*-Fur repression (HR) was the biggest group. From this reconstruction, a regulatory network motif with feedback loops was found for Fur regulating transportation and utilization of iron in a complicated fashion. Further functional analysis showed that Fur regulon is also involved in broader cellular processes beyond iron metabolism including DNA synthesis, nutrient search and energy metabolism.

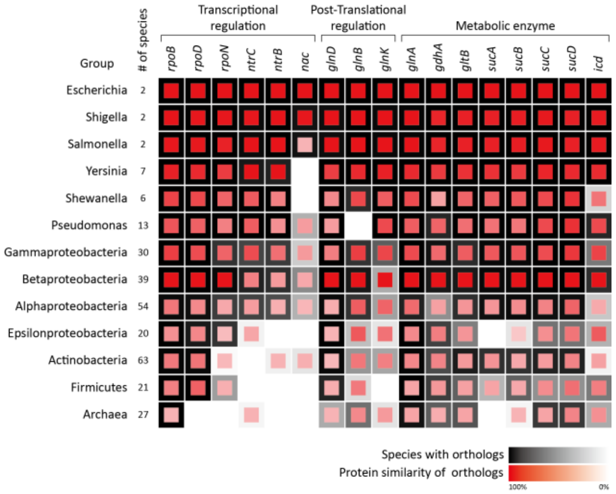


Figure 48. Conservation of regulatory components and metabolic enzymes in nitrogen metabolism.

Elucidating transcriptional regulation of nitrogen metabolism

Rendezvous of model-based computation and genome-scale experiments enables intelligent experimental design and interpretation of regulation in nitrogen metabolism. Unlike carbon and oxygen metabolism which have been investigated with systems approaches, nitrogen metabolism has not been studied at the same scale, except for more specific regulons. In order to investigate NtrC and Nac regulons at the systems level, model-driven prediction of activation conditions for NtrC and Nac was utilized, and experiments were designed for cytidine and cytosine from the prediction with ammonia as a negative control and glutamine as a positive control. ChIP-exo and RNA-seq experiments were performed, from which NtrC and Nac regulons were reconstructed. From the in-depth analysis with the reconstruction and predicted flux states calculated from M model concluded contrasting roles of NtrC and Nac regulons. In other words, NtrC regulon is responsible for primary response to nitrogen-limiting condition, while Nac regulon is in charge of rebalancing carbon metabolism to accommodate flux changes induced by nitrogen source shift. Conservation analysis of regulatory components and metabolic enzymes (Figure 48) and conservation of TFs in γ-proteobacteria (Figure 49) showed that metabolic enzymes are more conserved through all groups in microorganisms ranging from *Escherichia* to archaea, giving additional evidences to diver regulation on conserved gene contents.

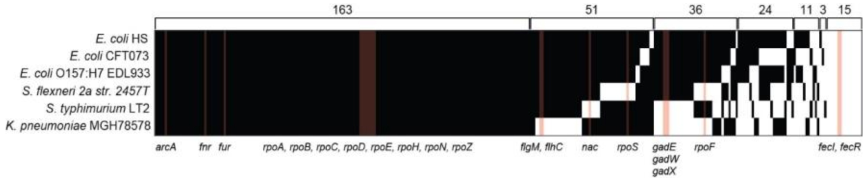


Figure 49. Conservation of TFs in γ-proteobacteria.

Conclusion

Thus, systems approaches enable a genome-scale assessment of regulatory components in multiple levels, and contribute to expansion of the current knowledge of bacterial transcription regulation and transcription initiation.

References

1. Decker, K.B. & Hinton, D.M. Transcription regulation at the core: similarities among bacterial, archaeal, and eukaryotic RNA polymerases. *Annu Rev Microbiol* **67**, 113-139 (2013).

2. Gruber, T.M. & Gross, C.A. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* **57**, 441-466 (2003).

3. Cho, B.K., Kim, D., Knight, E.M., Zengler, K. & Palsson, B.O. Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states. *BMC biology* **12**, 4 (2014).

4. Maciąg, A., Peano, C., Pietrelli, A., Egli, T., De Bellis, G. & Landini, P. In vitro transcription profiling of the sigmaS subunit of bacterial RNA polymerase: re-definition of the sigmaS regulon and identification of sigmaS-specific promoter sequence elements. *Nucleic Acids Res* **39**, 5338-5355 (2011).

5. Ades, S.E., Grigorova, I.L. & Gross, C.A. Regulation of the alternative sigma factor sigma(E) during initiation, adaptation, and shutoff of the extracytoplasmic heat shock response in Escherichia coli. *J Bacteriol* **185**, 2512-2519 (2003).

6. Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**, 327-335 (1996).

7. Kim, D., Hong, J.S., Qiu, Y., Nagarajan, H., Seo, J.H., Cho, B.K., Tsai, S.F. & Palsson, B.Ø. Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS Genet* **8**, e1002867 (2012).

8. McGrath, P.T., Lee, H., Zhang, L., Iniesta, A.A., Hottes, A.K., Tan, M.H., Hillson, N.J., Hu, P., Shapiro, L. & McAdams, H.H. High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol* **25**, 584-592 (2007).

9. 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, A.M., Collado-Vides, J. & Morett, E. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* **4**, e7526 (2009).

10. Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J.S., López-

Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A.M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E. & Collado-Vides, J. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* **39**, D98-105 (2011).

11. Cho, B.K., Zengler, K., Qiu, Y., Park, Y.S., Knight, E.M., Barrett, C.L., Gao, Y. & Palsson, B.Ø. The transcription unit architecture of the Escherichia coli genome. *Nat Biotechnol* **27**, 1043-1049 (2009).

12. Qiu, Y., Cho, B.K., Park, Y.S., Lovley, D., Palsson, B.Ø. & Zengler, K. Structural and operational complexity of the Geobacter sulfurreducens genome. *Genome research* **20**, 1304-1311 (2010).

13. Jäger, D., Sharma, C.M., Thomsen, J., Ehlers, C., Vogel, J. & Schmitz, R.A. Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* **106**, 21878-21882 (2009).

14. Buck, M.J. & Lieb, J.D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349-360 (2004).

15. Conrad, T.M., Frazier, M., Joyce, A.R., Cho, B.K., Knight, E.M., Lewis, N.E., Landick, R. & Palsson, B.Ø. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. *Proc Natl Acad Sci U S A* **107**, 20500-20505 (2010).

16. Cho, B.K., Barrett, C.L., Knight, E.M., Park, Y.S. & Palsson, B.O. Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19462-19467 (2008).

17. Cho, B.K., Federowicz, S., Park, Y.S., Zengler, K. & Palsson, B.O. Deciphering the transcriptional regulatory logic of amino acid metabolism. *Nature chemical biology* **8**, 65-71 (2012).

18. Cho, B.K., Federowicz, S.A., Embree, M., Park, Y.S., Kim, D. & Palsson, B.Ø. The PurR regulon in Escherichia coli K-12 MG1655. *Nucleic Acids Res* **39**, 6456-6464 (2011).

19. Cho, B.K., Knight, E.M. & Palsson, B.O. Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA. *Microbiology* **152**, 2207-2219 (2006).

20. Myers, K.S., Yan, H., Ong, I.M., Chung, D., Liang, K., Tran, F., Keleş, S., Landick, R. & Kiley, P.J. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet* **9**, e1003565 (2013).

21. Park, D.M., Akhtar, M.S., Ansari, A.Z., Landick, R. & Kiley, P.J. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. *PLoS Genet* **9**, e1003839 (2013).

22. Prieto, A.I., Kahramanoglou, C., Ali, R.M., Fraser, G.M., Seshasayee, A.S. & Luscombe, N.M. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res* **40**, 3524-3537 (2012).

23. Cho, B.K., Knight, E.M., Barrett, C.L. & Palsson, B.O. Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. *Genome research* **18**, 900-910 (2008).

24. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419 (2011).

25. Rhee, H.S. & Pugh, B.F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current protocols in molecular biology* **Chapter 21**, Unit 21 24 (2012).

26. Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P.F. & Vogel, J. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**, 250-255 (2010).

27. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenny, K., Sutton, G., FitzHugh, W., Fields, C., Gocyne, J.D., Scott, J., Shirley, R., Liu, L.I., Glodek, A., Kelley, J.M., Widman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. & Venter, J.C.. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).

28. Nielsen, C.B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. Visualizing genomes: techniques and challenges. *Nat Methods* **7**, S5-S15 (2010).

29. Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A. & Loraine, A.E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730-2731 (2009).

30. Engels, R., Yu, T., Burge, C., Mesirov, J.P., DeCaprio, D. & Galagan, J.E. Combo: a whole genome comparative browser. *Bioinformatics* **22**, 1782-1783 (2006).

31. Bare, J.C., Koide, T., Reiss, D.J., Tenenbaum, D. & Baliga, N.S. Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics* **11**, 382 (2010).

32. Ogawa, W., Li, D.W., Yu, P., Begum, A., Mizushima, T., Kuroda, T. & Tsuchiya, T. Multidrug resistance in Klebsiella pneumoniae MGH78578 and cloning of genes responsible for the resistance. *Biol Pharm Bull* **28**, 1505-1508 (2005).

33. McClelland, M., Florea, L., Sanderson, K., Clifton, S.W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R.K. & Miller, W. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **28**, 4974-4986 (2000).

34. Fouts, D.E., Tyler, H.L., DeBoy, R.T., Daugherty, S., Ren, Q., Badger, J.H., Durkin, A.S., Huot, H., Shrivastava, S., Kothari, S., Dodson, R.J., Mohamoud, Y., Khouri, H., Roesch, L.F., Krogfelt, K.A., Struve, C., Triplett, E.W. & Methé, B.A. Complete genome sequence of the N2-fixing broad host range endophyte Klebsiella pneumoniae 342 and virulence predictions verified in mice. *PLoS Genet* **4**, e1000141 (2008).

35. Dieterich, G., Karst, U., Fischer, E., Wehland, J. & Jansch, L. LEGER: knowledge database and visualization tool for comparative genomics of pathogenic and non-pathogenic Listeria species. *Nucleic Acids Res* **34**, D402-406 (2006).

36. Edwards, R.A., Olsen, G.J. & Maloy, S.R. Comparative genomics of closely related salmonellae. *Trends Microbiol* **10**, 94-99 (2002).

37. Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couvé, E., de Daruvar, A., Dehoux, P., Domann, E., Domínguez-Bernal, G., Duchaud, E., Durant, L., Dussurget, O., Entian, K.D., Fsihi, H., García-del Portillo, F., Garrido, P., Gautier, L., Goebel, W., Gómez-López, N., Hain, T., Hauf, J., Jackson, D., Jones, L.M., Kaerst, U., Kreft, J., Kuhn, M., Kunst, F., Kurapkat, G., Madueno, E., Maitournam, A., Vicente, J.M., Ng, E., Nedjari, H., Nordsiek, G., Novella, S., de Pablos, B., Pérez-Diaz, J.C., Purcell, R., Remmel, B., Rose, M., Schlueter, T., Simoes, N., Tierrez, A., Vázquez-Boland, J.A., Voss, H., Wehland, J. & Cossart, P. Comparative genomics of Listeria species. *Science* **294**, 849-852 (2001).

38. Hertz, G.Z. & Stormo, G.D. Escherichia coli promoter sequences: analysis and prediction. *Methods Enzymol* **273**, 30-42 (1996).

39. Harley, C.B. & Reynolds, R.P. Analysis of E. coli promoter sequences. *Nucleic Acids Res* **15**, 2343-2361 (1987).

40. Rajewsky, N., Socci, N.D., Zapotocky, M. & Siggia, E.D. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res* **12**, 298-308 (2002).

41. Dieterich, C., Wang, H., Rateitschak, K., Luz, H. & Vingron, M. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res* **31**, 55-57 (2003).

42. Gelfand, M.S., Koonin, E.V. & Mironov, A.A. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* **28**, 695-705 (2000).

43. Kaberdin, V.R. & Blasi, U. Translation initiation and the fate of bacterial mRNAs. *FEMS Microbiol Rev* **30**, 967-979 (2006).

44. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engström, P.G., Frith, M.C., Forrest, A.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A. & Hayashizaki, Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).

45. Raghavan, R., Groisman, E.A. & Ochman, H. Genome-wide detection of novel regulatory RNAs in E. coli. *Genome Res* **21**, 1487-1497 (2011).

46. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H. & Altuvia, S. Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol* **11**, 941-950 (2001).

47. Vogel, J., Bartels, V., Tang, T.H., Churakov, G., Slagter-Jäger, J.G., Hüttenhofer, A. & Wagner, E.G. RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* **31**, 6435-6443 (2003).

48. Papenfort, K., Said, N., Welsink, T., Lucchini, S., Hinton, J.C. & Vogel, J. Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA. *Mol Microbiol* **74**, 139-158 (2009).

49. Lu, X., Goodrich-Blair, H. & Tjaden, B. Assessing computational tools for the discovery of small RNA genes in bacteria. *RNA* **17**, 1635-1647 (2011).

50. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337 (2009).

51. Masukata, H. & Tomizawa, J. Control of primer formation for ColE1 plasmid replication: conformational change of the primer transcript. *Cell* **44**, 125-136 (1986).

52. Bailey, T.L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-29 (1995).

53. Burgess, R.R. & Anthony, L. How sigma docks to RNA polymerase and what sigma does. *Curr Opin Microbiol* **4**, 126-131 (2001).

54. Hawley, D.K. & McClure, W.R. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res* **11**, 2237-2255 (1983).

55. Zhang, Z. & Dietrich, F.S. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res* **33**, 2838-2851 (2005).

56. Crick, F.H. Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**, 548-555 (1966).

57. Romby, P., Vandenesch, F. & Wagner, E.G. The role of RNAs in the regulation of virulence-gene expression. *Curr Opin Microbiol* **9**, 229-236 (2006).

58. Park, S.Y., Cromie, M.J., Lee, E.J. & Groisman, E.A. A bacterial mRNA leader that employs different mechanisms to sense disparate intracellular signals. *Cell* **142**, 737-748 (2010).

59. Peer, A. & Margalit, H. Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions. *J Bacteriol* **193**, 1690-1701 (2011).

60. Gogol, E.B., Rhodius, V.A., Papenfort, K., Vogel, J. & Gross, C.A. Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon. *Proc Natl Acad Sci U S A* **108**, 12875-12880 (2011).

61. Møller, T., Franch, T., Højrup, P., Keene, D.R., Bächinger, H.P., Brennan, R.G. & Valentin-Hansen, P. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell* **9**, 23-30 (2002).

62. Valentin-Hansen, P., Eriksen, M. & Udesen, C. The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol* **51**, 1525-1533 (2004).

63. Chiang, M.K., Lu, M.C., Liu, L.C., Lin, C.T. & Lai, Y.C. Impact of Hfq on global gene expression and virulence in Klebsiella pneumoniae. *PLoS One* **6**, e22248 (2011).

64. Keseler, I.M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G. & Karp, P.D. EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* **37**, D464-470 (2009).

65. Soper, T., Mandin, P., Majdalani, N., Gottesman, S. & Woodson, S.A. Positive regulation by small RNAs and the role of Hfq. *Proc Natl Acad Sci U S A* **107**, 9602-9607 (2010).

66. Majdalani, N., Hernandez, D. & Gottesman, S. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol* **46**, 813-826 (2002).

67. Mandin, P. & Gottesman, S. Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *Embo J* **29**, 3094-3107 (2010).

68. Kawamoto, H., Koide, Y., Morita, T. & Aiba, H. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol* **61**, 1013-1022 (2006).

69. Horler, R.S. & Vanderpool, C.K. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res* **37**, 5465-5476 (2009).

70. Wu, C.C., Huang, Y.J., Fung, C.P. & Peng, H.L. Regulation of the Klebsiella pneumoniae Kpc fimbriae by the site-specific recombinase KpcI. *Microbiology* **156**, 1983-1992 (2010).

71. Wu, C.C., Lin, C.T., Cheng, W.Y., Huang, C.J., Wang, Z.C. & Peng, H.L. Fur-dependent MrkHI regulation of type 3 fimbriae in Klebsiella pneumoniae CG43. *Microbiology* **158**, 1045-1056 (2012).

72. Wilksch, J.J., Yang, J., Clements, A., Gabbe, J.L., Short, K.R., Cao, H., Cavaliere, R., James, C.E., Whitchurch, C.B., Schembri, M.A., Chuah, M.L., Liang, Z.X., Wijburg, O.L., Jenney, A.W., Lithgow, T. & Strugnell, R.A. MrkH, a novel c-di-GMP-dependent transcriptional activator, controls Klebsiella pneumoniae biofilm formation by regulating type 3 fimbriae expression. *PLoS Pathog* **7**, e1002204 (2011).

73. Rosenblum, R., Khan, E., Gonzalez, G., Hasan, R. & Schneiders, T. Genetic regulation of the ramA locus and its expression in clinical isolates of Klebsiella pneumoniae. *Int J Antimicrob Agents* **38**, 39-45 (2011).

74. Lin, C.T. & Peng, H.L. Regulation of the homologous two-component systems KvgAS and KvhAS in Klebsiella pneumoniae CG43. *J Biochem* **140**, 639-648 (2006).

75. Cao, V., Lambert, T. & Courvalin, P. ColE1-like plasmid pIP843 of Klebsiella pneumoniae encoding extended-spectrum beta-lactamase CTX-M-17. *Antimicrob Agents Chemother* **46**, 1212-1217 (2002).

76. Rice, L.B., Carias, L.L., Hujer, A.M., Bonafede, M., Hutton, R., Hoyen, C. & Bonomo, R.A. High-level expression of chromosomally encoded SHV-1 beta-lactamase and an outer membrane protein change confer resistance to ceftazidime and piperacillin-tazobactam in a clinical isolate of Klebsiella pneumoniae. *Antimicrob Agents Chemother* **44**, 362-367 (2000).

77. Achenbach, L.A. & Yang, W. The fur gene from Klebsiella pneumoniae: characterization, genomic organization and phylogenetic analysis. *Gene* **185**, 201-207 (1997).

78. de la Riva, L., Badia, J., Aguilar, J., Bender, R.A. & Baldoma, L. The hpx genetic system for hypoxanthine assimilation as a nitrogen source in Klebsiella pneumoniae: gene organization and transcriptional regulation. *J Bacteriol* **190**, 7892-7903 (2008).

79. Goss, T.J. The ArgP protein stimulates the Klebsiella pneumoniae gdhA promoter in a lysine-sensitive manner. *J Bacteriol* **190**, 4351-4359 (2008).

80. Liu, Q. & Bender, R.A. Complex regulation of urease formation from the two promoters of the ure operon of Klebsiella pneumoniae. *J Bacteriol* **189**, 7593-7599 (2007).

81. Rosario, C.J. & Bender, R.A. Importance of tetramer formation by the nitrogen assimilation control protein for strong repression of glutamate dehydrogenase formation in Klebsiella pneumoniae. *J Bacteriol* **187**, 8291-8299 (2005).

82. Grande, R.A., Valderrama, B. & Morett, E. Suppression analysis of positive control mutants of NifA reveals two overlapping promoters for Klebsiella pneumoniae rpoN. *J Mol Biol* **294**, 291-298 (1999).

83. Cheema, A.K., Choudhury, N.R. & Das, H.K. A- and T-tract-mediated intrinsic curvature in native DNA between the binding site of the upstream activator NtrC and the nifLA promoter of Klebsiella pneumoniae facilitates transcription. *J Bacteriol* **181**, 5296-5302 (1999).

84. Achenbach, L.A. & Genova, E.G. Transcriptional regulation of a second flavodoxin gene from Klebsiella pneumoniae. *Gene* **194**, 235-240 (1997).

85. Lin, J.T. & Stewart, V. Nitrate and nitrite-mediated transcription antitermination control of nasF (nitrate assimilation) operon expression in Klebsiella pheumoniae M5al. *J Mol Biol* **256**, 423-435 (1996).

86. Collins, C.M., Gutman, D.M. & Laman, H. Identification of a nitrogen-regulated promoter controlling expression of Klebsiella pneumoniae urease genes. *Mol Microbiol* **8**, 187-198 (1993).

87. Charlton, W., Cannon, W. & Buck, M. The Klebsiella pneumoniae nifJ promoter: analysis of promoter elements regulating activation by the NifA promoter. *Mol Microbiol* **7**, 1007-1021 (1993).

88. Buck, M. & Cannon, W. Mutations in the RNA polymerase recognition sequence of the Klebsiella pneumoniae nifH promoter permitting transcriptional activation in the absence of NifA binding to upstream activator sequences. *Nucleic Acids Res* **17**, 2597-2612 (1989).

89. Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P. & Hayashizaki, Y. Dynamic usage of transcription start sites within core promoters. *Genome Biol* **7**, R118 (2006).

90. Suzuki, Y., Yamashita, R., Nakai, K. & Sugano, S. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* **30**, 328-331 (2002).

91. David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W. & Steinmetz, L.M. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**, 5320-5325 (2006).

92. Cowing, D.W. & Gross, C.A. Interaction of Escherichia coli RNA polymerase holoenzyme containing sigma 32 with heat shock promoters. DNase I footprinting and methylation protection. *J Mol Biol* **210**, 513-520 (1989).

93. Lupski, J.R., Smiley, B.L. & Godson, G.N. Regulation of the rpsU-dnaG-rpoD macromolecular synthesis operon and the initiation of DNA replication in Escherichia coli K-12. *Mol Gen Genet* **189**, 48-57 (1983).

94. Burton, Z.F., Gross, C.A., Watanabe, K.K. & Burgess, R.R. The operon that encodes the sigma subunit of RNA polymerase also encodes ribosomal protein S21 and DNA primase in E. coli K12. *Cell* **32**, 335-349 (1983).

95. Yamamoto, K., Yata, K., Fujita, N. & Ishihama, A. Novel mode of transcription regulation by SdiA, an Escherichia coli homologue of the quorum-sensing regulator. *Mol Microbiol* **41**, 1187-1198 (2001).

96. Skippington, E. & Ragan, M.A. Evolutionary Dynamics of Small RNAs in 27 Escherichia coli and Shigella Genomes. *Genome Biol Evol* **4**, 330-345 (2012).

97. Caldara, M., Charlier, D. & Cunin, R. The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. *Microbiology* **152**, 3343-3354 (2006).

98. Charlier, D., Roovers, M., Van Vliet, F., Boyen, A., Cunin, R., Nakamura, Y., Glansdorff, N. & Piérard, A. Arginine regulon of Escherichia coli K-12. A study of repressor-operator interactions and of in vitro binding affinities versus in vivo repression. *J Mol Biol* **226**, 367-386 (1992).

99. Toledano, M.B., Kullik, I., Trinh, F., Baird, P.T., Schneider, T.D. & Storz, G. Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: a mechanism for differential promoter selection. *Cell* **78**, 897-909 (1994).

100. Squire, D.J., Xu, M., Cole, J.A., Busby, S.J. & Browning, D.F. Competition between NarL-dependent activation and Fis-dependent repression controls expression from the Escherichia coli yeaR and ogt promoters. *Biochem J* **420**, 249-257 (2009).

101. Putney, S.D. & Schimmel, P. An aminoacyl tRNA synthetase binds to a specific DNA sequence and regulates its gene transcription. *Nature* **291**, 632-635 (1981).

102. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. & Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).

103. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500 (2003).

104. Ishihama, A. Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol* **54**, 499-518 (2000).

105. Osterberg, S., del Peso-Santos, T. & Shingler, V. Regulation of alternative sigma factor use. *Annual review of microbiology* **65**, 37-55 (2011).

106. Sharma, U.K. & Chatterji, D. Transcriptional switching in Escherichia coli during stress and starvation by modulation of sigma activity. *FEMS Microbiol Rev* **34**, 646-657 (2010).

107. Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R. & Ishihama, A. Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *J Biol Chem* **280**, 1448-1456 (2005).

108. Herring, C.D., Raffaelle, M., Allen, T.E., Kanin, E.I., Landick, R., Ansari, A.Z. & Palsson, B.Ø. Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *Journal of bacteriology* **187**, 6166-6174 (2005).

109. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A.M., Treviño-Quintanilla, L. & Collado-Vides, J. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**, D120-124 (2008).

110. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).

111. Typas, A., Becker, G. & Hengge, R. The molecular basis of selective promoter activation by the sigmaS subunit of RNA polymerase. *Mol Microbiol* **63**, 1296-1306 (2007).

112. Typas, A. & Hengge, R. Role of the spacer between the -35 and -10 regions in sigmas promoter selectivity in Escherichia coli. *Mol Microbiol* **59**, 1037-1051 (2006).

113. Weber, H., Polen, T., Heuveling, J., Wendisch, V.F. & Hengge, R. 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 (2005).

114. Ishihama, A. Promoter selectivity of prokaryotic RNA polymerases. *Trends Genet* **4**, 282-286 (1988).

115. Loewen, P.C., Hu, B., Strutinsky, J. & Sparling, R. Regulation in the rpoS regulon of Escherichia coli. *Can J Microbiol* **44**, 707-717 (1998).

116. Farewell, A., Kvint, K. & Nystrom, T. Negative regulation by RpoS: a case of sigma factor competition. *Mol Microbiol* **29**, 1039-1051 (1998).

117. Jishage, M. & Ishihama, A. Regulation of RNA polymerase sigma subunit synthesis in Escherichia coli: intracellular levels of sigma 70 and sigma 38. *J Bacteriol* **177**, 6832-6835 (1995).

118. Seo, J.H., Hong, J.S., Kim, D., Cho, B.K., Huang, T.W., Tsai, S.F., Palsson, B.O. & Charusanti, P. Multiple-omic data analysis of Klebsiella pneumoniae MGH 78578 reveals its transcriptional architecture and regulatory features. *BMC genomics* **13**, 679 (2012).

119. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640-6645 (2000).

120. Powell, B.S., Court, D.L., Inada, T., Nakamura, Y., Michotey, V., Cui, X., Reizer, A., Saier, M.H. Jr & Reizer, J. Novel proteins of the phosphotransferase system encoded within the rpoN operon of Escherichia coli. Enzyme IIANtr affects growth on organic nitrogen and the conditional lethality of an erats mutant. *J Biol Chem* **270**, 4822-4839 (1995).

121. Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J.S., 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, A.M., Bonavides-Martínez, C., Balderas-Martínez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E. & Collado-Vides, J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* **41**, D203-213 (2013).

122. Carpenter, B.M., Whitmire, J.M. & Merrell, D.S. This is not your mother's repressor: the complex role of fur in pathogenesis. *Infection and immunity* **77**, 2590-2601 (2009).

123. Andrews, S.C., Robinson, A.K. & Rodriguez-Quinones, F. Bacterial iron homeostasis. *FEMS Microbiol Rev* **27**, 215-237 (2003).

124. Nandal, A., Huggins, C.C., Woodhall, M.R., McHugh, J., Rodríguez-Quiñones, F., Quail, M.A., Guest, J.R. & Andrews, S.C. Induction of the ferritin gene (ftnA) of Escherichia coli by Fe(2+)-Fur is mediated by reversal of H-NS silencing and is RyhB independent. *Molecular microbiology* **75**, 637-657 (2010).

125. Carpenter, B.M., Gilbreath, J.J., Pich, O.Q., McKelvey, A.M., Maynard, E.L., Li, Z.Z. & Merrell, D.S. Identification and Characterization of Novel Helicobacter pylori apo-Fur-Regulated Target Genes. *J Bacteriol* **195**, 5526-5539 (2013).

126. Butcher, J., Sarvan, S., Brunzelle, J.S., Couture, J.F. & Stintzi, A. Structure and regulon of Campylobacter jejuni ferric uptake regulator Fur define apo-Fur regulation. *Proc Natl Acad Sci U S A* **109**, 10047-10052 (2012).

127. Deng, X., Sun, F., Ji, Q., Liang, H., Missiakas, D., Lan, L. & He, C. Expression of multidrug resistance efflux pump gene norA is iron responsive in Staphylococcus aureus. *J Bacteriol* **194**, 1753-1762 (2012).

128. Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A.G., Mackie, A., Paulsen, I., Gunsalus, R.P. & Karp, P.D. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res* **39**, D583-590 (2011).

129. McHugh, J.P., Rodríguez-Quinoñes, F., Abdul-Tehrani, H., Svistunenko, D.A., Poole, R.K., Cooper, C.E. & Andrews, S.C. Global iron-dependent gene regulation in Escherichia coli. A new mechanism for iron homeostasis. *J Biol Chem* **278**, 29478-29486 (2003).

130. Masse, E. & Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A* **99**, 4620-4625 (2002).

131. Green, J. & Paget, M.S. Bacterial redox sensors. *Nat Rev Microbiol* **2**, 954-966 (2004).

132. Cao, J., Woodhall, M.R., Alvarez, J., Cartron, M.L. & Andrews, S.C. EfeUOB (YcdNOB) is a tripartite, acid-induced and CpxAR-regulated, low-pH Fe2+ transporter that is cryptic in Escherichia coli K-12 but functional in E. coli O157:H7. *Mol Microbiol* **65**, 857-875 (2007).

133. Lee, J., Page, R., García-Contreras, R., Palermino, J.M., Zhang, X.S., Doshi, O., Wood, T.K. & Peti, W. Structure and function of the Escherichia coli protein YmgB: a protein critical for biofilm formation and acid-resistance. *J Mol Biol* **373**, 11-26 (2007).

134. Belitsky, B.R. & Sonenshein, A.L. Genome-wide identification of Bacillus subtilis CodY-binding sites at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 7026-7031 (2013).

135. Baichoo, N. & Helmann, J.D. Recognition of DNA by Fur: a reinterpretation of the Fur box consensus sequence. *J Bacteriol* **184**, 5826-5832 (2002).

136. Semsey, S., Andersson, A.M., Krishna, S., Jensen, M.H., Massé, E. & Sneppen, K. Genetic regulation of fluxes: iron homeostasis of Escherichia coli. *Nucleic Acids Res* **34**, 4960-4967 (2006).

137. Desnoyers, G., Morissette, A., Prevost, K. & Masse, E. Small RNA-induced differential degradation of the polycistronic mRNA iscRSUA. *Embo J* **28**, 1551-1561 (2009).

138. Krishna, S., Semsey, S. & Sneppen, K. Combinatorics of feedback in cellular uptake and metabolism of small molecules. *Proc Natl Acad Sci U S A* **104**, 20815-20819 (2007).

139. Martin, J.E. & Imlay, J.A. The alternative aerobic ribonucleotide reductase of Escherichia coli, NrdEF, is a manganese-dependent enzyme that enables cell replication during periods of iron starvation. *Mol Microbiol* **80**, 319-334 (2011).

140. Gerstle, K., Klatschke, K., Hahn, U. & Piganeau, N. The small RNA RybA regulates key-genes in the biosynthesis of aromatic amino acids under peroxide stress in E. coli. *RNA biology* **9**, 458-468 (2012).

141. O'Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R. & Palsson, B.O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol* **9**, 693 (2013).

142. Friedman, D.B., Stauff, D.L., Pishchany, G., Whitwell, C.W., Torres, V.J. & Skaar, E.P. Staphylococcus aureus redirects central metabolism to increase iron availability. *PLoS pathogens* **2**, e87 (2006).

143. Park, T.H., Park, J.H., Kim, J.K., Seo, S.W., Rah, D.K. & Chang, C.H. Analysis of 15 Cases of Auricular Keloids Following Conchal Cartilage Grafts in an Asian Population. *Aesthetic plastic surgery* (2013).

144. Morgan, B. & Lahav, O. The effect of pH on the kinetics of spontaneous Fe(II) oxidation by O2 in aqueous solution--basic principles and a simple heuristic description. *Chemosphere* **68**, 2080-2084 (2007).

145. Wu, Y. & Outten, F.W. IscR controls iron-dependent biofilm formation in Escherichia coli by regulating type I fimbria expression. *J Bacteriol* **191**, 1248-1257 (2009).

146. Grainger, D.C., Hurd, D., Goldberg, M.D. & Busby, S.J. Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res* **34**, 4642-4652 (2006).

147. Shimada, T., Ishihama, A., Busby, S.J. & Grainger, D.C. The Escherichia coli RutR transcription factor binds at targets within genes as well as intergenic regions. *Nucleic acids research* **36**, 3950-3955 (2008).

148. Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. & Regev, A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-715 (2010).

149. McCloskey, D., Palsson, B.O. & Feist, A.M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* **9**, 661 (2013).

150. Cho, B.K., Knight, E.M. & Palsson, B.O. PCR-based tandem epitope tagging system for Escherichia coli genome engineering. *BioTechniques* **40**, 67-72 (2006).

151. Datta, S., Costantino, N. & Court, D.L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379**, 109-115 (2006).

152. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. & Noble, W.S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009).

153. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

154. Joyce, A.R. & Palsson, B.O. The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology* **7**, 198-210 (2006).

155. Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M. & Palsson, B.Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* **7**, 535 (2011).

156. Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A. & Kustu, S. Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation. *Proc Natl Acad Sci U S A* **97**, 14674-14679 (2000).

157. Camarena, L., Poggio, S., Garcia, N. & Osorio, A. Transcriptional repression of gdhA in Escherichia coli is mediated by the Nac protein. *FEMS microbiology letters* **167**, 51-56 (1998).

158. Muse, W.B. & Bender, R.A. The nac (nitrogen assimilation control) gene from Escherichia coli. *J Bacteriol* **180**, 1166-1173 (1998).

159. Rhee, H.S. & Pugh, B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295-301 (2012).

160. Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A.G., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I. & Karp, P.D. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* **41**, D605-612 (2013).

161. Reitzer, L. Nitrogen assimilation and global regulation in Escherichia coli. *Annu Rev Microbiol* **57**, 155-176 (2003).

162. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).

163. Novichkov, P.S., Kazakov, A.E., Ravcheev, D.A., Leyn, S.A., Kovaleva, G.Y., Sutormin, R.A., Kazanov, M.D., Riehl, W., Arkin, A.P., Dubchak, I. & Rodionov, D.A. RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC genomics* **14**, 745 (2013).

164. Pomposiello, P.J., Janes, B.K. & Bender, R.A. Two roles for the DNA recognition site of the Klebsiella aerogenes nitrogen assimilation control protein. *J Bacteriol* **180**, 578-585 (1998).

165. Claverie-Martin, F. & Magasanik, B. Role of integration host factor in the regulation of the glnHp2 promoter of Escherichia coli. *Proc Natl Acad Sci U S A* **88**, 1631-1635 (1991).

166. Ninfa, A.J., Reitzer, L.J. & Magasanik, B. Initiation of Transcription at the Bacterial Ginap2 Promoter by Purified Escherichia-Coli Components Is Facilitated by Enhancers. *Cell* **50**, 1039-1046 (1987).

167. Ueno-Nishio, S., Mango, S., Reitzer, L.J. & Magasanik, B. Identification and regulation of the glnL operator-promoter of the complex glnALG operon of Escherichia coli. *J Bacteriol* **160**, 379-384 (1984).

168. Atkinson, M.R., Blauwkamp, T.A., Bondarenko, V., Studitsky, V. & Ninfa, A.J. Activation of the glnA, glnK, and nac promoters as Escherichia coli undergoes the transition from nitrogen excess growth to nitrogen starvation. *J Bacteriol* **184**, 5358-5363 (2002).

169. Muse, W.B., Rosario, C.J. & Bender, R.A. Nitrogen regulation of the codBA (cytosine deaminase) operon from Escherichia coli by the nitrogen assimilation control protein, NAC. *J Bacteriol* **185**, 2920-2926 (2003).

170. Schneider, B.L., Hernandez, V.J. & Reitzer, L. Putrescine catabolism is a metabolic response to several stresses in Escherichia coli. *Mol Microbiol* **88**, 537-550 (2013).

171. Garcia, E. & Rhee, S.G. Cascade control of Escherichia coli glutamine synthetase. Purification and properties of PII uridylyltransferase and uridylyl-removing enzyme. *J Biol Chem* **258**, 2246-2253 (1983).

172. van Heeswijk, W.C., Hoving, S., Molenaar, D., Stegeman, B., Kahn, D. & Westerhoff, H.V. An alternative PII protein in the regulation of glutamine synthetase in Escherichia coli. *Mol Microbiol* **21**, 133-146 (1996).

173. Blauwkamp, T.A. & Ninfa, A.J. Physiological role of the GlnK signal transduction protein of Escherichia coli: survival of nitrogen starvation. *Mol Microbiol* **46**, 203-214 (2002).

174. Javelle, A., Severi, E., Thornton, J. & Merrick, M. Ammonium sensing in Escherichia coli. Role of the ammonium transporter AmtB and AmtB-GlnK complex formation. *J Biol Chem* **279**, 8530-8538 (2004).

175. Vasudevan, S.G., Gedye, C., Dixon, N.E., Cheah, E., Carr, P.D., Suffolk, P.M., Jeffrey, P.D. & Ollis, D.L. Escherichia coli PII protein: purification, crystallization and oligomeric structure. *FEBS Lett* **337**, 255-258 (1994).

176. Liu, J. & Magasanik, B. Activation of the dephosphorylation of nitrogen regulator I-phosphate of Escherichia coli. *J Bacteriol* **177**, 926-931 (1995).

177. Atkinson, M.R., Blauwkamp, T.A. & Ninfa, A.J. Context-dependent functions of the PII and GlnK signal transduction proteins in Escherichia coli. *J Bacteriol* **184**, 5364-5375 (2002).

178. van Heeswijk, W.C., Wen, D., Clancy, P., Jaggi, R., Ollis, D.L., Westerhoff, H.V. & Vasudevan, S.G. The Escherichia coli signal transducers PII (GlnB) and GlnK form heterotrimers in vivo: fine tuning the nitrogen signal cascade. *Proc Natl Acad Sci U S A* **97**, 3942-3947 (2000).

179. Studholme, D.J. & Buck, M. The biology of enhancer-dependent transcriptional regulation in bacteria: insights from genome sequences. *FEMS microbiology letters* **186**, 1-9 (2000).

180. Huo, Y.X., Nan, B.Y., You, C.H., Tian, Z.X., Kolb, A. & Wang, Y.P. FIS activates glnAp2 in Escherichia coli: role of a DNA bend centered at -55, upstream of the transcription start site. *FEMS microbiology letters* **257**, 99-105 (2006).

181. Ikeda, T.P., Shauger, A.E. & Kustu, S. Salmonella typhimurium apparently perceives external nitrogen limitation as internal glutamine limitation. *J Mol Biol* **259**, 589-607 (1996).

182. Jiang, P., Peliska, J.A. & Ninfa, A.J. Enzymological characterization of the signal-transducing uridylyltransferase/uridylyl-removing enzyme (EC 2.7.7.59) of Escherichia coli and its interaction with the PII protein. *Biochemistry* **37**, 12782-12794 (1998).

183. Ninfa, A.J. & Atkinson, M.R. PII signal transduction proteins. *Trends Microbiol* **8**, 172-179 (2000).

184. Forchhammer, K., Hedler, A., Strobel, H. & Weiss, V. Heterotrimerization of PII-like signalling proteins: implications for PII-mediated signal transduction systems. *Mol Microbiol* **33**, 338-349 (1999).

185. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* **100**, 11980-11985 (2003).

186. Mangan, S., Zaslaver, A. & Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* **334**, 197-204 (2003).

187. Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo J* **20**, 2528-2535 (2001).