Elucidating mechanisms of transcriptional regulation at the genome-scale

Federowicz, Stephen A.

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Elucidating mechanisms of transcriptional regulation at the genome-scale

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

in

Bioinformatics and Systems Biology

by

Stephen A. Federowicz

Committee in charge:

Professor Bernhard Palsson, Chair
Professor Wei Wang, Co-Chair
Professor Terence Hwa
Professor Milton Saier
Professor Gurol Suel

2014

The dissertation of Stephen A. Federowicz is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2014

DEDICATION

To my mom for your unconditional love and support. To my dad for always being there for me and for always giving me good advice, even when I was too stubborn to take it.

To my two sisters who always looked out for me and provided inspiration to look up to.

To my girlfriend Autumn who helped me through some of the darkest times.

To all of my amazing friends, and all those who have supported me along the way. We truly only exist in relation to others.

EPIGRAPH

Bhiksus, you should know that all of the teachings I give to you are a raft.

—The Diamond Sutra

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

This list should almost certainly be longer than it is going to be but here it goes.

The first two people I need to specifically acknowledge and thank are Joshua Lerman and Nathan Lewis. I do not feel I am exaggerating when I say that I may not have made it through grad school without both of you.

I also want to thank Haythem Latif for always hanging in there and being an amazing co-author.

Ali Ebrahim for putting up with my tabs instead of spaces. Karsten Zengler for talking with me about redox and bagels. Edward O'Brien for talking models and code. Harish Nagarajan for showing me the ropes. Aarash Bordbar for modeling expertise.Daniel Zelinski for stimulating discussions. Zak King for enthusiastic code talks.

I also thank Bernhard Palsson for his support and vision.

Chapter 2 in full is a reprint of a published manuscript: Cho, B.K., Federowicz, S.A., Embree, M., Park, Y.S., Kim, D., Palsson B.. The PuR regulon in Escherichia coli K12-MG1655. Nucleic Acids Research. 2011 Aug;39(15):6456-64. doi: 10.1093/nar/gkr307. The dissertation author was the second author of this paper, responsible for the computational analysis. The other authors were Byung-Kwan Cho, Mallory Embree, Young-Seob Park, Donghyuk Kim, and, Bernhard Ø. Palsson.

Chapter 3 in full is a reprint of a published manuscript: Cho, B.K., Federowicz, S.A., Park, Y.S., Zengler, K., Palsson B.. Deciphering the transcriptional regulatory logic of amino acid metabolism. Nature Chemical Biology. 2011 Nov 13;8(1):65-71. doi: 10.1038/nchembio.710. The dissertation author was the second author of this paper, responsible for the computational analysis. The other authors were Byung-Kwan Cho,

Young-Seob Park, Karsten Zengler, and, Bernhard Ø. Palsson.

Chapter 4 in full is a reprint of a submitted manuscript: Federowicz, S.A.*, King, Z.A.*, O'Brien, E.J.*, Ebrahim, A. *, Lerman, J.A.*, Lu, J. *, Sonnenschein, N., Latif, H., Lewis, N.E., Palsson, B.. The OME Framework for Genome-Scale Systems Biology. Submitted. * Indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Zak King (equal contributor), Edward O'Brien (equal contributor), Ali Ebrahim (equal contributor), Joshua Lerman (equal contributor), Justin Lu (equal contributor), Niko Sonnenschein, Haythem Latif, Nathan Lewis and, Bernhard Ø. Palsson.

Chapter 5 in full is a reprint of a manuscript in preparation to be submitted: Latif, H.*, Federowicz, S.A.*, Tarasova, J., Szubin, R., Carreri, J.U., Ebrahim, A., Zengler, K.A., Palsson, B.. Mechanistic and systems level analysis of canonical transcriptional activation in microbes using ChIP-exo. In Preparation. * Indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Haythem Latif (equal contributor), Janna Tarasova, Richard Szubin, Jose Carreri, Ali Ebrahim, Karsten Zengler and, Bernhard Ø. Palsson.

Chapter 6 in full is a reprint of a published manuscript: Federowicz, S.A.*, Kim, D.K., Ebrahim, A, Lerman, J.A., Nagarajan, H.N., Cho, B.K., Zengler, K.A., Palsson, B.. Determining the Control Circuitry of Redox Metabolism at the Genome-Scale. Plos Genetics. 2014 Apr;10(4): e1004264 doi: 10.1371/journal.pgen.1004264. The dissertation author was the primary author of this paper responsible for the research. The other authors were Donghyuk Kim, Ali Ebrahim, Joshua Lerman, Harish Nagarajan, Byung-Kwan Cho, Karsten Zengler, and, Bernhard Ø. Palsson.

VITA

2009	B. S. in Bioinformatics *cum laude*, University of California, Santa Cruz
2014	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Latif, H.*, **Federowicz, S.A.***, Tarasova, Y., Szubin, R., Carreri, J.U., Ebrahim, A., Zengler, K.A., Palsson, B.. Mechanistic and systems level analysis of canonical transcriptional activation in microbes using ChIP-exo. In Preparation.

Federowicz, S.A.*, King, Z.A.*, O'Brien, E.J.*, Ebrahim, A. *, Lerman, J.A.*, Lu, J. *, Sonnenschein, N., Latif, H., Lewis, N.E., Palsson, B.. The OME Framework for Genome-Scale Systems Biology. Submitted.

Federowicz, S.A.*, Kim, D.K., Ebrahim, A, Lerman, J.A., Nagarajan, H.N., Cho, B.K., Zengler, K.A., Palsson, B.. Determining the Control Circuitry of Redox Metabolism at the Genome-Scale. Plos Genetics. 2014 Apr;10(4): e1004264 doi: 10.1371/journal.pgen.1004264

Cho, B.K., **Federowicz, S.A.**, Park, Y.S., Zengler, K., Palsson B.. Deciphering the transcriptional regulatory logic of amino acid metabolism. Nature Chemical Biology. 2011 Nov 13;8(1):65-71. doi: 10.1038/nchembio.710.

Cho, B.K., **Federowicz, S.A.**, Embree, M., Park, Y.S., Kim, D., Palsson B.. The PuR regulon in Escherichia coli K12-MG1655. Nucleic Acids Research. 2011 Aug;39(15):6456-64. doi: 10.1093/nar/gkr307.

ABSTRACT OF THE DISSERTATION

Elucidating mechanisms of transcriptional regulation at the genome-scale

by

Stephen A. Federowicz

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Bernhard Palsson, Chair
Professor Wei Wang, Co-Chair

Throughout the course of evolution, almost all organisms have generated complex, hierarchical, and robust regulatory systems. One major component of these biological regulatory systems is the transcriptional activation or repression of gene expression. This regulation is carried out simultaneously across a genome by thousands of biological components at thousands of individual promoters. The sum total of all of the regulatory events and their interconnections or overlaps is commonly referred to as the transcriptional regulatory network. The focus of this thesis is to determine the mechanisms or guiding principles behind these transcriptional regulatory networks and to provide a basis upon which predictive mathematical models of these networks can be built. In the first section, a reconstruction of the full transcriptional regulatory network for a model organism is presented along with the OME software framework developed to handle the full complexity of genome-scale datasets and models. In the second section, the mechanisms of individual regulatory events are elucidated in a massively parallel fashion using ChIP-exonuclease and the OME framework. This leads to fundamental insights into the nature of transcriptional initiation complexes for canonical regulators. Finally, in the third section, an effort is undertaken to determine systems level mechanisms which dictate the coordinate regulation of hundreds of simultaneous regulatory events in response to major physiological and metabolic perturbations. Here we show that the two principal dimensions of a metabolic system, growth and the production of energy, drive not only the organization of the metabolic network, but also the organization of the transcriptional regulatory network.

Chapter 1

Introduction

1.1 Overview

Biological systems, relative to our current understanding, are inherently complex. Even in the most well studied model organisms, such as E. coli and S. cerevisiae, many basic questions are unanswered. This lack of understanding is at both the level of individual cellular components and at the level of overall cellular systems which emerge from the interaction and networks of individual components. At the level of individual components, unknowns include (to name a few) the allosteric regulatory sites of many enzymes[1], the precise functioning of RNA polymerase[2], and the structure-function relationship between small molecules and many proteins[3]. At the systems level only basic understandings of pathway organization[4], metabolic fluxes[5], and some regulatory features [6, 7] exist.

One success at the systems level has come in the field of constraint-based systems biology for metabolic networks[8, 9, 10]. By reconstructing the precise stoichiometry of every reaction in a cell and encoding it into a system of linear equations, the fluxes of all reactions in a cell can be simulated through optimization techniques[11]. Recently, these basic techniques have been extended to include not just metabolic reactions, but also the enzymes which catalyze each reaction[12]. These models, called ME models (Metabolism, Expression) have proved to be incredibly powerful by accounting for the cost of metabolic reactions (via protein biosynthesis through transcription and translation reactions of each enzyme gene product) and subsequently improving the predictive power of genome-scale models.

One long sought addition to metabolic network modeling has been the network of transcriptional regulation [13, 14, 15, 16, 17, 18, 19] composed of the set of regulatory proteins which sense internal or external signals and subsequently regulate the expression levels of genes related to each signal. For three primary reasons, the modeling of this transcriptional regulatory network has been difficult and remained infeasible outside of statistical or non-mechanistic modeling approaches. The first main reason is a lack of a comprehensive regulatory network reconstruction. The second is a lack of basic knowledge about the process of regulation at individual promoters. The third is a lack of systems level objectives and constraints which would enable a mathematical formulation of a regulatory network at the genome-scale.

This thesis aims to address each of these three core issues. We first present two characteristic works, Chapters 2 and 3, which demonstrate the process of reconstructing a regulatory network and integrating it with an understanding of metabolic networks. We then present a formalization of this process with the OME (operational, metabolic, expression) software framework that enables the reconstruction of full regulatory networks to addres the first key issue. We next significantly advance the understanding of canonical transcriptional acivation in Chapter 6 via a detailed analysis of ChIP-exonuclease and RNA-seq data in E. coli. Finally, we take a few steps forward in our understanding of the systems level features of regulatory systems in Chapter 7 by showing that the dimensionality of metabolic networks is mirrored in the dimensionality of the transcriptional regulatory network.

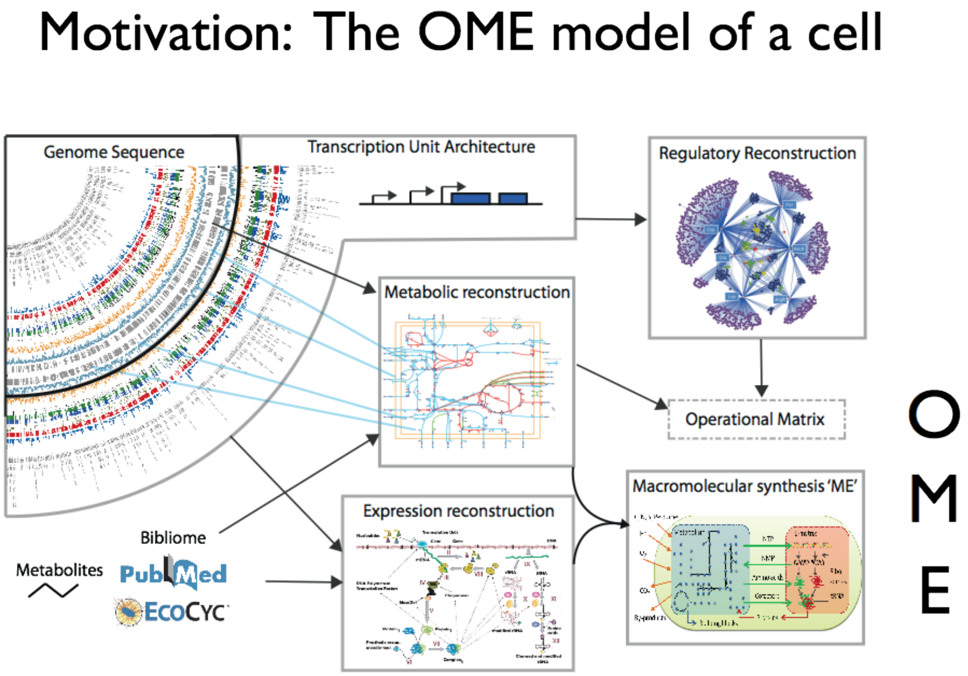

Motivation: The OME model of a cell
Genome Sequence
Transcription Unit Architecture
Regulatory Reconstruction
Metabolic reconstruction
Operational Matrix
Macromolecular synthesis 'ME'
Expression reconstruction
Bibliome
Metabolites
PubMed
EcoCyc
O
M
E

Figure 1.1: **Defining the OME model of a cell.** From left to right, the OME model of a cell starts with a genome sequence and corresponding genomic annotation. Genomic datasets measuring the transcriptome, transcription start sites, and proteomics are then mapped onto the genome to generate a full genomic metastructure. From this information a metabolic network reconstruction along with a full expression machinery reconstruction is developed. Simultaneously a reconstruction of the transcriptional regulatory network is developed. A synthesis of the O (operational for transcriptional regulatory network), M (metabolic network reconstruction), and E (expression machinery reconstruction) generates the full OME model of a cell. Representing a fully mechanistic whole cell model.

1.2 Global regulators in E. coli

Transcription factors represent one of the primary means by which an organism can sense and respond to stimuli. This is done through the detection of intra or extra cellular small molecules and metabolites which often directly bind a transcription factor or propagate a signal through two-component systems. The outcome is the direct and indirect regulation of a large number or genes which in the case of global transcription factors often represent a wide diversity of function. The patterns of regulation often fall into patterns of connected feed-back and feedforward loops in which the levels of a certain metabolite, element, or molecule will be kept in balance by the activation or repression of genes which transport, biosynthesize, or utilize the given molecule. This amounts to charting out the regulatory response in order to define what types of cellular pathways are regulated.

We experimentally determined the network level mechanisms of transcriptional response in E. coli by integrating genome wide binding analysis of five global transcription factors with associated expression profiling and transcription start site information. ChIP-chip experiments for ArcA, Fnr, FruR, Crp, Lrp, and ArgR revealed 143, 100, 45, 247, 144, and 61 unique binding regions, respectively. Binding peaks were mapped to gene expression data generated between a wild type strain and a gene deletion strain for each of the transcription factors to discern whether a binding event resulted in gene activation or repression. This allowed us to construct a functional network of transcriptional regulation that elucidates the response to major environmental stimuli. This includes the response to oxygen availability mediated by ArcA and Fnr, shifting carbon source mediated by FruR and Crp, and primary nitrogen source mediated by Lrp and ArgR. Subsequently, we were able to elucidate core connected feed back loops through which regulatory information flows and gene expression is systematically controlled.

ArcA or anoxic redox control is intimately linked to the anaerobic response and has been more recently shown to finely regulate micro-aerobic states[20]. Fnr named for the fumarate and nitrite reduction mutant phenotype that marked its initial discovery is also critical for the anaerobic response as it regulates the vast majority of transcripts encoding enzymes for usage of nitrogen sources as a secondary electron acceptor. FruR initially named as a fructose repressor is also known as Cra, or catabolite repressor activator. FruR was discovered as a cyclic AMP(cAMP) independent component of catabolite repression. Crp or cAMP receptor protein is one of the most heavily studied transcription factors in E. Coli which was initially found as the primary mode of catabolite repression. In general Crp activates a large number of operons which encode pathways for secondary usage of carbon compounds. While the cell is growing on a primary carbon source it produces enough cya (adenylate cyclase) so that as soon the TCA cycles starts to slow down due to a shift towards famine condition the free floating AMP can be turned into cAMP and thus Crp can turn on the secondary catabolic systems.

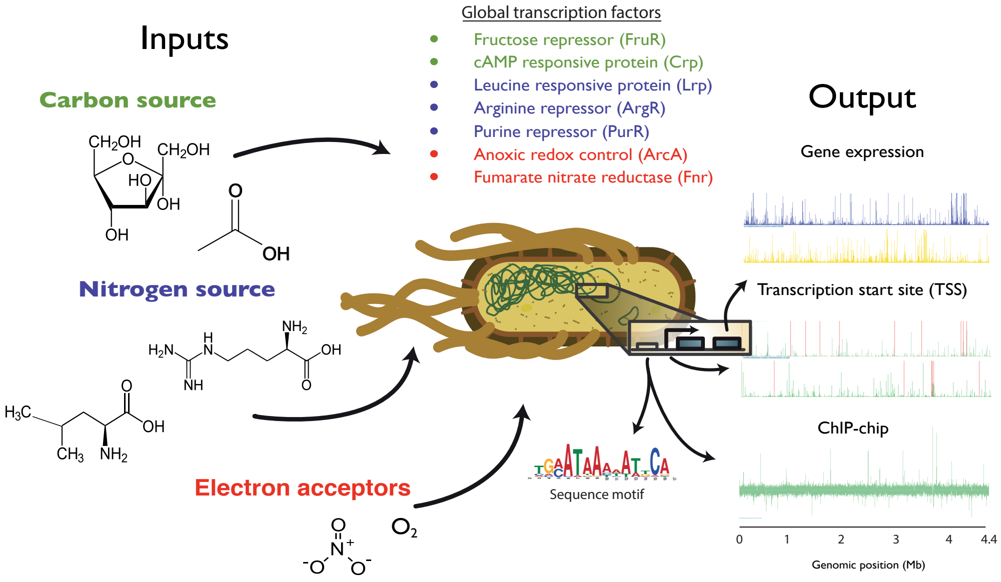

Inputs
Carbon source
Nitrogen source
Electron acceptors
O_2
Global transcription factors
Fructose repressor (FruR)
cAMP responsive protein (Crp)
Leucine responsive protein (Lrp)
Arginine repressor (ArgR)
Purine repressor (PurR)
Anoxic redox control (ArcA)
Fumarate nitrate reductase (Fnr)
Output
Gene expression
Transcription start site (TSS)
ChIP-chip
Sequence motif
0
1
2
3
4
4.4
Genomic position (Mb)

Figure 1.2: **Overview of experiments for regulatory network reconstruction.** The overall experimental approach was to elucidate the mechanisms of transcriptional regulation in response to three core physiological stimuli. This included carbon source, nitrogen source, and oxygen availibility. These key environmental parameters are sensed and mitigated by each of the transcription factors (TFs) under consideration. Cra and Crp are two classically studied TFs which primarily regulate central carbon metabolism and ensure optimality. ArcA and Fnr are also well studied TFs with central roles in anaerobic metabolism. ArgR and Lrp are well known for regulation of amino acid biosynthetic pathways. However, each TF also regulates large numbers of targets which are seeminly unrelated to its primary roles hence giving them the title of global regulators.

1.3 Analogies for systems biology

One common analogy for constraint-based metabolic modeling is that of a network of pipes, roads, or wires. To take the example of pipes, one can imagine the network of water pipes which flow throughout a city and deliver water from a source in the mountains to the sinks of each home. This network, even for a small city, will contain thousands of pipes of various sizes and types that are all configured to carry out the task of efficiently delivering water when needed. In a cell, this is analgous to the network of metabolic reactions which exists amoung thousands of small molecule metabolites and thousands of reactions which can transform one metabolite into another. In a cell the inputs are nutrient sources (food) which first flow through the catabolic reactions (pipes) of a cell before branching off into anabolic reactions (pipes) that result in growth, or into reactions which ultimately produce energy.

Chapter 2

The PurR regulon in Escherichia coli K-12 MG1655

2.1 Abstract

The PurR transcription factor plays a critical role in transcriptional regulation of purine metabolism in enterobacteria. Here, we elucidate the role of PurR under exogenous adenine stimulation at the genome-scale using high-resolution chromatin immunoprecipitation (ChIP)chip and gene expression data obtained under in vivo conditions. Analysis of microarray data revealed that adenine stimulation led to changes in transcript level of about 10% of Escherichia coli genes, including the purine biosyn- thesis pathway. The E. coli strain lacking the purR gene showed that a total of 56 genes are affected by the deletion. From the ChIPchip analysis, we determined that over 73% of genes directly regulated by PurR were enriched in the biosynthesis, utilization and transport of purine and pyrimidine nucleotides, and 20% of them were functionally unknown. Compared to the functional diversity of the regulon of the other general transcription factors in E. coli, the functions and size of the PurR regulon are limited.

2.2 Introduction

Metabolism enables a cell to assimilate exogenous nutrients both for energy generation and for macromolecular synthesis. A set of metabolic pathways directs the biosynthesis and utilization of nucleotides that is critical for virtually every aspect of cellular life. Purine and pyrimidine nucleotides constitute a part of the nucleic acids, cofactors in enzymatic reactions, intracellular and extracellular signals, phosphate donors and the major carriers of cellular energy [21]. Since the biosynthesis and utilization of the nucleotides is demanding of cellular resources and plays a broad role in cellular processes, imbalances between the different nucleotide pools significantly perturb the normal cellular functions(2,3).

In general, metabolism is tightly controlled by feedback inhibition of enzyme activity by metabolites and by transcriptional regulation by DNA-binding proteins. The regulatory action of DNA-binding proteins is also modulated by the exogenous nutrients as stimuli. Therefore, there is great interest in not only elucidating the set of genes under regulation by the same stimuli (defined as a stimulon), but also identifying the collection of genes under regulation by the same regulatory protein (defined as a regulon). In the case of purine nucleotide metabolism in Escherichia coli, purine repressor (PurR) tightly regulates transcription of the enzymes involved in inosine 50-monophosphate (IMP) biosynthesis and the conversion of IMP to adenosine monophosphate (AMP) and guanosine monophosphate (GMP)(2). The regulatory action of PurR on target genes is modulated by the binding of the small effector molecules [hypoxanthine (Hx) or guanine] and in effect endows PurR with the ability to affect transcriptional regulation [22]. In other words, upon availability of purine nucleotides from the environment, the activity of PurR can be enhanced to repress the expression of target genes. However, little is known about in vivo PurR-binding events and their causal relationships with gene expression at the genome scale in the presence or absence of purine nucleotides. Such information is needed to reconstruct the PurR regulon and to understand purine metabolism.

The E. coli transcriptional regulatory network is believed to have a hierarchical topology with several global transcription factors (TFs) at the top-level [23, 24, 25]. The global TFs were specified by the multiple functional categories of the genes regulated. By contrast, specific TFs restrict their target genes to the same metabolic pathways or the same functional categories [24]. Previously, PurR was classified into the group of general TFs; however, due to the lack of information on target genes in its regulon, the understanding of the role of PurR is limited. In particular, it is unclear whether its effects on E. coli metabolism are direct or indirect. If the effects are indirect, it is also unclear whether the indirect effects are made through other TFs or other metabolites.

In this study, we obtain and integrate genome-scale data from chromatin immunoprecipitation (ChIP)chip and gene expression profiling to elucidate the regulatory role of PurR at a genome scale. First, using changes in transcript levels on a genome scale, we defined the adenine stimulon from comprehensively established sets of genes differentially expressed in response to exogenous adenine. Second, we used the purR deletion mutant to determine the PurR-dependent genes affected by the deletion mutant. Third, we set out to comprehensively establish the PurR-binding regions on the E. coli genome experimentally to further elucidate any DNA sequence motif correlated with the PurR regulatory action. Fourth, we determined the regulatory action of PurR based on the causal relationships between the association of PurR and changes in transcript levels. In the end, the reconstruction of the regulatory network of PurR allows us to understand the role of the PurR regulon as a part of the broader adenine stimulon. The results show that the role of PurR regulon is locally acting but its effect on the entire metabolism is critical in response to the exogenous purine stimulation.

2.3 Results

2.3.1 Determination of gene expression changes to exogenous adenine

To characterize the changes in gene expression with exposure to exogenous adenine, global transcriptome analyses were performed using DNA microarrays. Escherichia coli strain K-12 MG1655 (wild type) was grown in M9 medium supplemented by 100mg/ml adenine. Samples were removed from the culture with and without adenine stimulation and used for the extraction of total RNA. Analysis of the DNA microarray datademonstrated that exogenous adenine stimulation led to changes in transcript level of about 10% of E. coli genes (Supplementary Table S1). Those include the upregulation of 144 genes and downregulation of 255 genes with more than 2-fold expression change and P-value ¡0.05. As previously well described (2), all of the genes associated with the purine biosynthesis pathway were repressed by the addition of adenine. Among them, purD encoding phosphoribosylamineglycine ligase showed the highest repression factor (43.89-fold) (Supplementary Table S2). Interestingly, adenine addition led to the high downregulation of several transporter genes, including codB, xanP, yeeF and uraA encoding cytosine transporter (34.52-fold), xanthine NCS2 transporter (20.12-fold), amino acid APC transporter (18.61-fold) and uracil NCS2 transporter (17.64-fold).

A significant portion of the genes downregulated by adenine addition is associated with the pyrimidine and amino acid biosynthetic pathways. In particular, the downregulation of genes comprising the arginine biosynthetic pathway clearly demonstrates that the purine metabolism links to the in vivo level of arginine in order to regulate the biosynthesis of deoxyribonucleic and ribonucleic acid[26]. On the other hand, the highly induced genes by adenine stimulation were involved in other various cellular processes (Supplementary Table S2). Among them, ydhC encoding drug MFS transporter had the highest activation factor (193.55-fold). Of interest, two non-coding RNAs, gcvB and rybB, were induced by a factor of 34.37 and 5.92, respectively. Previous studies showed that GcvB enhances the ability of E. coli to survive low pH by upregulating the levels of alternate sigma factor, RpoS[27]. Another alternate sigma factor, RpoE-dependent RybB, regulates the synthesis of major porins in E. coli[28]. Among genes in purine salvage pathways, add encoding adenosine deaminase was induced by adenine as well (17.61-fold) (2). The genes in thiamine and biotin biosynthesis pathways were also induced by the exogenous adenine. These observations demonstrate that a large number of the downregulated genes related with purine and pyrimidine biosynthesis and transport, arginine biosynthesis and ATP synthesis coupled with proton transport form a major portion (64%) of the adenine stimulon.

2.3.2 PurR-dependent transcriptome response to the exogenous adenine

Next, we studied the global response of a purR deletion mutant to adenine stimulation to identify PurR-dependent genes (Supplementary Table S1). To address this issue, we isolated total RNA from the isogenic purR deletion mutant during exponential growth phase and hybridized the cDNA obtained from the total RNA onto Affymetrix microarrays. A comparison of the gene expression levels between cells grown in the presence and absence of the PurR protein in response to the exogenous adenine revealed that a total of 56 genes exhibit differential expression with more than 2-fold change and a false discovery rate (FDR) value ¡0.05 (P-value = 0.0056) from analysis of variance (ANOVA) analysis (Supplementary Table S3).

Nineteen genes (34%) showed increased transcript levels in response to the exogenous adenine due to regulation by PurR (Supplementary Table S3). None of these 19 genes has been previously reported to be directly regulated by PurR. On the other hand, transcription of the 37 genes (66%) was repressed by PurR. Eleven genes of the IMP (inosine 5-monophosphate) biosynthetic pathway from PRPP (5-phosphoribosyl-1-pyrophosphate) clustered into this group. It has been previously determined that 10 of them were directly repressed by the PurR protein (1624). Eight genes in pyrimidine bio- synthesis and transport pathways were directly or indirectly regulated by the PurR protein, that include carA, carB, pyrB, pyrI, pyrC, pyrD, codA and codB. It has been experimentally determined that four of them (carA, carB, pyrD and pyrC) were directly repressed by the PurR [29, 30, 31, 32, 33]. Transcription of yieG (2.91-fold), xanP (30.82-fold), and uraA (2.04-fold) encoding adenine, xanthine and uracil transporter, respectively, were also affected by the purR deletion, indicating direct or indirect regulatory effect of the PurR protein on the adenine and uracil transport systems [34]. Transcriptional repression of genes in arginine biosynthesis pathway (argA, argB and argC) is potentially mediated by the PurR protein. Interestingly, acid stress response genes (hdeB, hdeA and hdeD) decreased expression in a purR deletion mutant. Consistent with this observation, the level of messenger RNA (mRNA) transcript of gadY, a regulatory small RNA that is highly upregulated by low pH [35], was affected by the purR deletion. A hallmark of the E. coli response to the exogenous adenine is the rapid and strong repression of a set of genes in purine biosynthesis pathway. The observed repression of all of these genes in the wild-type strain, but not the purR deletion mutant, provided an internal validation of the microarray experiment.

2.3.3 Genome-wide identification of PurR regulon

PurR-binding regions have been characterized by in vitro DNA-binding experiments and mutational analysis; however, direct analysis of in vivo PurR binding is not available. We thus employed the ChIP coupled with microarrays (ChIPchip) approach to determine the in vivo PurR-binding regions in E. coli cells under either the presence or

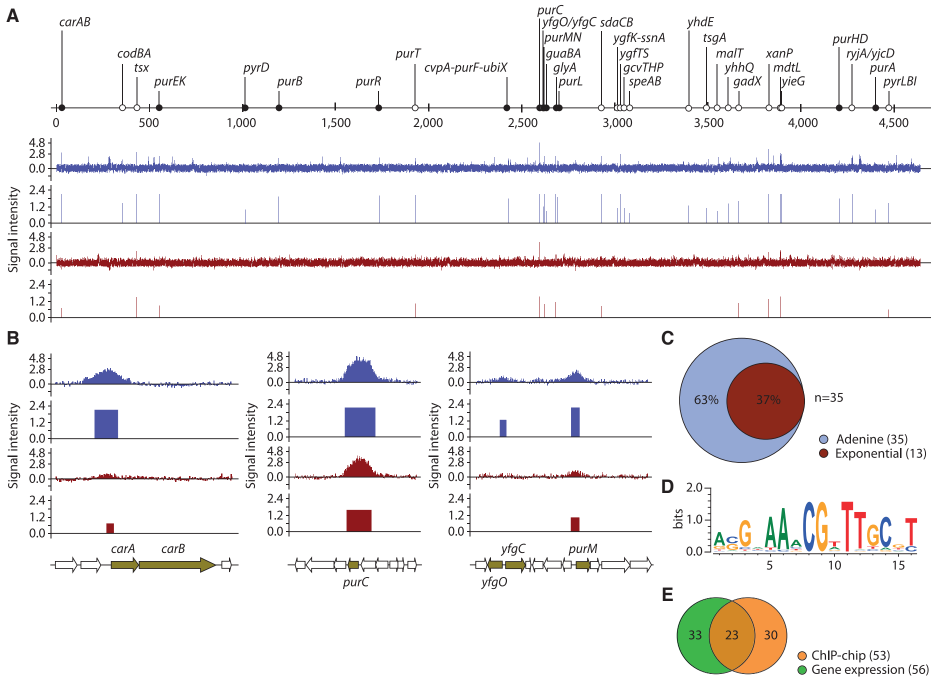


Figure 2.1: **Genome-wide distribution of PurR-binding regions.** (A) An overview of PurR-binding profiles across the E. coli genome at exponential growth phase in the presence (blue) or absence (red) of exogenous adenine. Black and white dots indicate previously known and newly found PurR-binding regions, respectively. (B) Determination of genuine PurR-binding regions on the selected regions. Promoter regions of carAB, purC, yfgO/yfgC and purM are occupied by PurR at exponential state. The peak height of the identified PurR-binding region is the log 2 enrichment ratio calculated from Cy5 (IP DNA) and Cy3 (mock IP DNA) signal intensity of the probe corresponding to the identified region. (C) Overlaps between PurR-binding regions of exponential phase in the presence (blue) and absence (red) of adenine. (D) Sequence logo representation of the PurRDNA binding profile. (E) Comparison of ChIPchip results and gene expression profiles.

the absence of exogenous adenine (Figure 1).

We performed a hybridization of the immunoprecipitated DNA (Cy5 channel) and mock immunoprecipitated DNA (Cy3 channel) onto the high-resolution whole-genome tiling microarrays, which contained a total of 371034 oligonucleotides with 50-bp tiles overlapping every 25bp on both forward and reverse strands [36, 37]. The normalized log2 ratios obtained from the hybridization identify the genomic regions enriched in the IP-DNA sample compared with the mock IP-DNA sample and thereby represent a genome-wide map of in vivo interactions between PurR protein and E. coli genome (Figure 1A). Using a peak finding algorithm, 35 and 13 unique and reproducible PurR-binding regions were identified from the hybridizations in exponential phase in the presence and absence of adenine, respectively (Table 1).

The genome-wide PurR-binding maps obtained from two different conditions, i.e. exponential growth phase in the presence and the absence of exogenous adenine, indicated that the PurR association on the E. coli genome is dramatically sensitive to the addition of adenine. For instance, PurR occupancy for the promoter regions of carAB, purC and purMN transcription units showed a great differential ratio between those two conditions (Figure 1B, Table 1). At the previously characterized PurR-binding promoter regions of pyrD, purB, purR, cvpA-purF-ubiX, guaBA, purL and purHD, we only observed the PurR-binding in the presence of exogenous adenine. Only 37% of binding sites (13 of 35) overlapped under the conditions in the absence and presence of exogenous adenine, and 62% of binding sites (22 of 35) were found in the presence of exogenous adenine (Figure 1C). Adenine can be converted to IMP through the intermediate formation of adenosine, inosine, and Hx catalyzed by purine nucleoside phosphorylase (deoD) and adenosine deaminase (add)(2). Thus, this observation indicates that the addition of exogenous adenine increased in the intracellular level of Hx, which led to the formation of the PurRHx complex that functions in transcriptional regulation[22]. A total of 22 new PurR-binding regions were identified in this study, whose roles were involved in various cellular processes (Table 1). Prior to this study, 15 PurR-binding regions had been characterized by DNA-binding experiments in vitro and mutational analysis in vivo, 87% (13 of 15) of which were identified in this study (transcription units in bold characters in Table 1). The exceptions were pyrC and glnB promoters, whose cellular functions are related to the pyrimidine biosynthesis and nitrogen metabolism, respectively. It is unclear why those PurR-binding regions were missed from our analysis.

We next assessed the locations of the PurR-binding regions against the current annotated genome information[36]. The PurR-binding regions were observed only within intergenic (i.e., promoter and promoter-like) regions. Therefore, there exists a strong preference for the PurR-binding target to be located within the noncoding intergenic regions, similar to that observed for Lrp-binding sites[37]. To identify common DNA sequence motifs of the PurR-binding regions, we used the MEME suite tool[38]. The sequences of PurR-binding regions were used to generate the position specific probability matrix and to rescan the entire genome with the FIMO program. We then analyzed only those sites which were located in the PurR-binding regions and fell below a stringent cut-off (P-value ¡0.0001). This revealed a total of 28 conserved sequences spread across 35 binding regions (Table 1). The identified sequence motif (ACGNAAACGTTTGCNT) was consistent with the previously characterized 16-bp palindromic binding site of the PurR (Figure 1D) (2). Based on the fact that the increase in the intracellular adenine levels enhances PurR binding to its DNA targets and the coverage of the known binding regions in our data, we concluded that PurR-binding regions identified here are bona fide binding sites.

2.3.4 Genome-scale determination of causal relationship

Currently, a total of 22 genes have been characterized as members of PurR regulon to be directly repressed by PurR (33). From our ChIPchip analyses, we significantly expanded the size of the PurR regulon to comprise 53 target genes (Table 1). To determine the causal relationships between the binding of PurR and the changes in RNA transcript levels of genes in the PurR regulon, we integrated the information on the binding regions of PurR with transcriptomic analysis. Among 53 target genes in PurR regulon determined by ChIPchip analyses, we determined 23 genes (43%) differentially expressed in response to the purR deletion and the addition of exogenous adenine with more than a 2-fold change and an FDR value ¡0.05 (P-value=0.0056) from ANOVA analysis (Figure 1E and Supplementary Table S4). The genes directly repressed by PurR in response to the exogenous adenine (23 genes) include codB, codA, purT, xanP, yieG,pyrL, pyrB and pyrI encoding cytosine NCS1 transporter, cytosine deaminase, phosphoribosylglycinamide-formyltransferase, xanthine NCS2 transporter, adenine transporter, PyrL leader peptide, aspartate carbamoyltransferase and aspartate carbamoyltransferase regula- tory subunit, respectively, as newly found members.

The remaining 57% of the genes had a direct association with PurR, lacking significant changes in RNA transcript levels. The cellular functions of most of the remaining genes were not clustered in purine and pyrimidine metabolic pathways, indicating that the changes in their transcript levels require additional regulatory signals such as transcription factors. Surprisingly, none of the genes were directly activated by PurR. On the contrary, PurR completely represses target genes involved in the IMP biosynthetic pathway. This suggests that most of the repression is a direct interaction, but the transcriptional activation is indirect (Supplementary Tables S1, S3 and S4).

In general, the PurR-binding sites are located in the promoter region between position 35 and 10 promoter elements, indicating that the binding of PurR regulates transcription initiation (2). In the case of purB and purR, PurR binds to the open reading frame so that it blocks transcription elongation[39]. Therefore, the binding position of PurR is of great interest in order to understand its regulatory mechanism. We calculated the distance between PurR-binding motifs and the transcription start sites (TSSs) based upon the TSSs recently published[36]. Of 32 promoter regions directly regulated by PurR, 14 regions (44%) include the PurR-binding motif between 10 and 35 promoter elements.

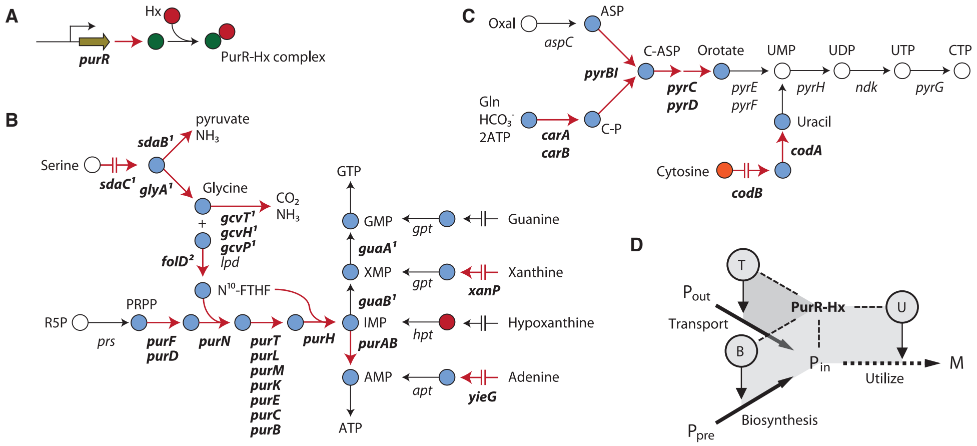


Figure 2.2: **Metabolic pathways directly regulated by PurR and its regulatory motif.** (A) Formation of PurRhypoxanthine complex. (B, C) The purine and pyrimidine biosynthesis pathways. The genes directly regulated by PurR are depicted by bold characters. Red arrows show the PurR-mediated repression. Broken arrows indicate the transporters. (D) Schematic diagram for the regulatory motif reconstruction in feedback loop.

2.3.5 Metabolic pathways directly regulated by PurR–Hx complex

Purine nucleotide metabolism plays a critical role in various cellular activities. To identify the metabolic pathways regulated by PurRHx complex, the members of PurR regulon were functionally classified and further mapped to the E. coli metabolic pathways (Figure 2). The genes with direct PurR association lacking changes in transcript levels were classified into other cellular functions. However, the genes directly repressed by PurR in response to the exogenous adenine clustered mainly into purine and pyrimidine metabolic pathways. First, PurR directly autoregulates itself and generates PurRHx complex in the presence of Hx or adenine (Figure 2A). Second, PurR completely regulates purine transport, biosynthesis, salvage and interconversion pathways (Figure 2B). Interestingly, PurR directly regulates serine transport and metabolic pathways to produce N10-formyltetrahydrofolate (N10-FTHF), which is an intermediate for the IMP biosynthetic pathway. In addition, we found that PurR directly represses xanthine (xanP), purine nucleoside (tsx) and adenine (yieG) transporters. Although the transporter for Hx is currently unknown, xanthine transporter is unable to transport Hx [40]. Lastly, PurR downregulates the genes in pyrimidine biosynthetic and transport pathways (Figure 2C). Most of the genes having direct PurR association with differential gene expression were enriched into purine and pyrimidine metabolic pathways. Interestingly, none of genes involved with purine utilization, such as apt, deoD and add, are directly regulated by PurR.

2.4 Discussion

We determined the PurR regulon in E. coli in response to the exogenous adenine stimuli by integrating genome-scale location analysis and gene expression profiles. The genome-wide map of PurR-binding sites presented here not only confirms previously characterized binding sites (15 regions) but also expands the number of known binding sites (35 regions) to a genome-wide assessment; similar to what we previously reported for Lrp- and Fis-binding sites [41, 37]. From the genome-wide mapping results, we were also able to show that: (i) a total of 35 PurR-binding regions were identified, all of which were located within noncoding regions, showing the strong binding preference of

PurR-binding to the promoter and promoter-like regions; (ii) only 37% of binding sites (13 of 35) overlapped under the conditions in the absence and presence of exogenous adenine, indicating that PurR bindings to the E. coli genome are dramatically sensitive to the addition of exogenous adenine (or hypoxanthine); (iii) the integration of these results with mRNA transcript level information indicates that the functional assignment of the regulated genes is strongly enriched in the purine and pyrimidine metabolism-related functions. In addition, most of the genes were thoroughly repressed by PurR. Interestingly, the other genes directly bound by PurR lacking differential expression in response to the purR deletion or the exogenous adenine were functionally diverse; and (iv) the PurR-binding motifs were observed at the regions of 10 and 35 promoter elements, indicating PurR regulates transcription initiation.

We discovered PurR-binding regions from the promoter regions of codBA, purT, xanP, yieG and pyrLBI with the differential gene expression. First, PurR directly regulates the de novo biosynthesis of pyrimidine. Among the genes in the biosynthetic pathway, codB and codA encode a cytosine transporter belonging to the NCS1 family of purine and pyrimidine transporters and a cytosine deaminase metabolizing cytosine to uracil and ammonia, respectively. In addition, PurR directly regulates pyrB and pyrI, encoding catalytic and regulatory subunits of aspartate transcarbamylase (ATCase), respectively, catalyzing the first reaction of the de novo biosynthesis of pyrimidine nucleotides. Considering that PurR represses carA, carB, pyrC and pyrD, the very early steps of de novo pyrimidine biosynthesis and transport are tightly regulated by PurR in response to exogenous adenine. Interestingly, RutR, the uracil responsive transcription factor, binds to the promoter region of carAB[42]. Although Shimada and co-workers demonstrated that the RutR binding site plays little or no role in the regulation of transcription, it may have an additional regulatory role along with other proteins. In the carAB promoter, at least five regulatory proteins (IHF, PepA, PurR, RutR and ArgR) are involved in the purine, pyrimidine and arginine-specific control of the promoter activity [32]. The complexity of the multicomponent regulatory mechanisms modulating carAB transcription shows the need for a cellular balance between the synthesis of pyrimidine and purine residues. Second, PurR directly regulates the transport of purine nucleotides. We observed PurR-binding peaks at the upstream regions of xanP, mdtL and yieG with differential gene expression in response to the exogenous adenine. Interestingly, PurR-binding peaks were observed for xanP and mdtL in the absence of exogenous adenine but not for yieG, suggesting that the yieG encodes a high-affinity transport system for adenine, which is dispensable in the presence of excess substrate. It has been suggested that another adenine transport system close to the genomic position of yieG operates at low affinity and is not energy dependent [34]. However, it has not been discovered which gene has the adenine transport function with low affinity. Here, we found that mdtL encoding drug MFS transporter is directly repressed by PurR and located close to yieG (4.5 kb), indicating that mdtL might be responsible for the low-affinity adenine transport. Thus, the purine transport system of the PurR regulon can be composed of two high-affinity transporters (xanP and yieG) and one low-affinity transporter (mdtL).

Transcriptional regulatory systems often regulate the formation rates and the concentration of small molecules by feedback loops that regulate the transport, biosynthesis and metabolic enzymes [36, 43]. Since adenine can be utilized by apt, encoding adenine phosphoribosyl-transferase, we were able to connect transport, biosynthesis and metabolic feedback loop pairs (Figure 2D). In the left loop, PurRHx complex represses the transcription of the transport proteins (T) for purine (xanP and yieG) and pyrimidine (codB), and biosynthetic proteins (B) for IMP (purEK, purB, purT, purF, purC, purMN, purl and purHD) and UMP (carAB, codA, pyrBI and pyrCD), reducing the influx of the purine or pyrimidine molecules (Pin) from the media (Pout) and precursors (Ppre). In the right loop, metabolic enzyme (U) responsible for converting Pin into metabolites (M) is not directly regulated by PurRHx complex; however, its transcript level is reduced by the exogenous adenine. Thus, the logical structure of the connected feedback loop (CFL) motif described by a notation that uses three signs indicating repression (R) or activation (A) for each of T, B, and U can be R-R-R. In the previous studies[37, 43], the B component (i.e. biosynthesis) was not included in the logical structures. The R-R-R motif demonstrates that the influx and efflux are repressed for flow homeostasis[43], which means that the exogenous adenine cannot be utilized as nutrient molecules. In the case of nutrient molecules and homeostasis, the logical structures of CFL would have been A-A/R-A and R-A/ R-A, respectively[43]. Since the R-R-R motif is uncommon for the regulation of small molecules in living cells, transcriptional regulatory networks for maintaining the levels of the purine and pyrimidine molecules may be more complex than previously thought.

Previously, the PurR was classified into the group of general TFs based on the functional diversity of the genes in its regulon[24]. Compared to the other general TFs in E. coli such as Fnr [44], Crp[45] and Lrp[37], the functions and size of the PurR regulon are limited. However, cellular functions of the genes are highly enriched into the purine and pyrimidine transport and biosynthesis, indicating that the direct effect of PurR on the E. coli metabolism is local, but via the balance of cellular purine content, it plays a critical role in metabolism. Now we may need to select a new list of global transcription factors in E. coli.

2.5 Materials and Methods

2.5.1 Bacterial strains and growth conditions

All strains used are E. coli K-12 MG1655 and its derivatives. The E. coli strain harboring PurR-8myc was generated as described previously [46]. Deletion mutant (purR) was constructed by a Red and FLP-mediated site-specific recombination system [47]. Glycerol stocks of E. coli strains were inoculated into M9 minimal medium supplemented with 2 g/l glucose and cultured overnight at 37C with constant agitation. The cultures were inoculated into 100 ml of the fresh M9 minimal medium in either the presence or absence of 100 mg/ml adenine and continued to culture at 37C with constant agitation to mid-log phase.

2.5.2 Transcriptome analysis

Samples for transcriptome analyses were taken from exponentially growing cells. From the cells treated by 2 vol of RNAprotect Bacteria Reagent (Qiagen), total RNA was isolated using RNeasy kit (Qiagen) with DNaseI treatment in accordance with manufacturers instruction. AffymetrixGeneChipE. coli Genome 2.0 arrays were used for genome-scale transcriptional analyses. Complementary DNA (cDNA) synthesis, fragmentation, end-terminus biotin labeling and array hybridization were performed as recommended by Affymetrix standard protocol. Raw CEL files were analyzed using robust multi-array average for normalization and calculation of probe intensities.

2.5.3 ChIP and microarray analysis

To identify PurR-binding regions in vivo, we isolated the DNA bound to PurR protein by ChIP. Cultures at mid-log phase were cross-linked by 1% formaldehyde at room temperature for 25min. After cell lysis and sonication, the cross-linked DNA-PurR complex was immunoprecipitated by using the specific antibody against myc-tag (9E10, Santa Cruz Biotech) and Dynabeads Pan Mouse IgG magnetic beads (Invitrogen) followed by stringent washings as described previously[41]. After reversal of the cross-links by incubation at 65C overnight, the samples were treated by RNaseA (Qiagen) and proteaseK (Invitrogen) and then purified with a PCR purification kit (Qiagen). Then, the amplified ChIP DNA samples were labeled and hybridized onto whole-genome tiled microarrays (Roche-NimbleGen).

2.5.4 Data analysis

To identify PurR-binding regions, we used the peak finding algorithm built into the NimbleScanTM software. Processing of ChIPchip data was performed in three steps: normalization, IP/mock-IP ratio computation (log base 2), and enriched region identification. The log2 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 bi-weight mean. The log2 ratio of Cy5 (IP DNA) to Cy3 (mock-IP DNA) for each point was calculated from the scanned signals. Then, the bi-weight mean of this log2 ratio was subtracted from each point. Each log ratio dataset from triplicate samples was used to identify PurR-binding region using the software (width of sliding window = 300 bp). Our approach to identify the PurR-binding regions was to first determine binding locations from each data set and then combine the binding locations from at least five of the six data sets to define a binding region[36].

2.5.5 Motif searching

The PurR-binding motif analysis was completed using the MEME and FIMO tools from the MEME software suite[38]. We first determined the proper binding motif and then scanned the full genome for its presence. The elicitation of the motif was done using the MEME program on the set of sequences defined by the PurR-binding regions. Using default settings, the previously determined PurR motif was recovered and then tailored to the correct size by setting the width parameter to 16bp. We then used these motifs and the PSPM (position specific probability matrix) generated by MEME to rescan the entire genome with the FIMO program.

2.6 Acknowledgements

Chapter 2 in full is a reprint of a published manuscript: Cho, B.K., Federowicz, S.A., Embree, M., Park, Y.S., Kim, D., Palsson B.. The PuR regulon in Escherichia coli K12-MG1655. Nucleic Acids Research. 2011 Aug;39(15):6456-64. doi: 10.1093/nar/gkr307. The dissertation author was the second author of this paper, responsible for the computational analysis. The other authors were Byung-Kwan Cho, Mallory Embree, Young-Seob Park, Donghyuk Kim, and, Bernhard Ø. Palsson.

Chapter 3

Deciphering the transcriptional regulatory logic of amino acid metabolism

3.1 Abstract

Although metabolic networks have been reconstructed on a genome-scale, the corresponding reconstruction and integration of governing transcriptional regulatory networks has not been fully achieved. Here we reconstruct such an integrated network for amino acid metabolism in Escherichia coli. Analysis of ChIP-chip and gene expression data for the transcription factors ArgR, Lrp, and TrpR showed that 19/20 amino acid biosynthetic pathways are either directly or indirectly controlled by these regulators. Classifying the regulated genes into three functional categories of transport, biosynthesis, and metabolism leads to elucidation of regulatory motifs constituting the integrated networks basic building blocks. The regulatory logic of these motifs was determined based on the relationships between transcription factor binding and changes in transcript

levels in response to exogenous amino acids. Remarkably, the resulting logic shows how amino acids are differentiated as signaling and nutrient molecules, and thus revealing the overarching regulatory principles of this stimulon.

3.2 Introduction

Transcriptional regulatory networks (TRN) in bacteria govern metabolic flexibility and robustness in response to environmental signals[48]. Thus, causal relationships between transcript levels for metabolic genes and the direct association of transcription factors (TFs) at the genome-scale is fundamental to fully understand bacterial responses to their environment [14, 37]. In particular, the molecular interaction between small molecules ranging from nutrients to trace elements and TFs governs the TRN and ultimately regulates the related metabolic pathways. From the causal relationships, a small set of recurring regulation patterns, or network motifs [37, 49] were identified and reconstructed to describe the design principles of complex biological systems. One primary discovery from this effort was the connected feedback circuit which coordinates influx (biosynthesis and transport) and efflux (metabolism) pathways that are jointly regulated by a TF sensing the relevant small molecule[37]. For example, a part of the global TRN is comprised of certain TFs (ArgR, Lrp, and TrpR) that sense the presence of exogenous amino acids (arginine, leucine, and tryptophan, respectively) and, in response, regulate the expression of a number of target genes5. Upon addition of these amino acids to the environment, the TFs exhibit enhanced, reversed, or unaffected regulatory modes[37, 50, 51, 52]. These TF responses make these amino acids not just nutrients but also signaling molecules[53].

Previously discovered network motifs[37, 49] represent a significant step forward in our understanding of complex biological behavior. However, they fail to appropriately elucidate the system wide response since they were either based upon incomplete information[49], or were only specific to a single transcription factor and regulon[37]. This has resulted in an inability to appropriately understand complex regulatory phenomena existing across multiple transcription factors and regulatory signals. Hence, it is necessary to achieve a full elucidation of these interactions with systematic and integrated experimental analysis.

Comprehensive elucidation of the causal relationships between transcription factors and genes is achievable by integrated analysis of expression data obtained from microarray or sequencing (e.g., RNA-seq)[54] with direct TF-binding information from chromatin immunoprecipitation coupled with microarrays or sequencing (ChIP-chip or ChIP-seq)[37, 55]. Here we obtain and integrate genome-scale expression profiling and ChIP-chip for each TF to reconstruct regulons involved in amino acid metabolism at the genome-scale. The elucidated regulatory logic falls into two categories that differentiate the role of amino acids as signaling and as nutrient molecules. Therefore, the reconstruction of the regulatory logic of the network motif allows us to establish the physiological role of each TF regulon and to determine how they govern amino acid regulation in E. coli. The integration of these multiple regulons into a unified network led to the first full bottom-up genome-scale reconstruction of a stimulon.

3.3 Results

3.3.1 Genome-wide identification of TF-binding regions

ArgR, Lrp, and TrpR are TFs involved in amino acid metabolism in E. coli[50, 51, 56] responding to arginine, leucine, and tryptophan, respectively. The binding of the small effector molecule (here being the amino acids) to these TFs carries out the genomes regulatory code by enhancing or decreasing the TFs affinity for a specific genomic region and concurrently modulating the transcription of downstream genes. In the case of Lrp, the direct analysis of in vivo binding was fully described[37] using chromatin immunoprecipitation coupled with microarrays (ChIP-chip) experiments. A total of 141 binding regions were analyzed, representing coverage of 74% of the previously identified regions[37]. However, similar genome-scale data for the other two major TFs in amino acid metabolism, ArgR and TrpR were unavailable. To determine their binding regions on a genome-wide level in an unbiased manner, we employed the ChIP-chip approach to E. coli cells harboring 8myc-tagged ArgR or TrpR protein[46]. The resulting log2 ratios obtained from the ChIP-chip experiments identify the genomic regions enriched in the IP-DNA sample compared with the mock IP-DNA sample and thereby represent a genome-wide map of in vivo ArgR- and TrpR-binding regions (Fig. 1a).

Using a previously described binding region detection algorithm[36] 61 and 8 unique and reproducible ArgR- and TrpR-binding regions were identified, respectively (Supplementary Table 1 and Supplementary Table 2). The 61 ArgR-binding sites detected included 13 sites previously characterized by DNA-binding experiments in vitro and mutational analyses in vivo[57, 58]. For example, the ArgR-arginine complex transcriptionally represses gltBD, artPIQM operon, and artJ gene encoding arginine transport systems[59, 60]. Our results confirmed that the ArgR-arginine complex binds to each of these promoter regions (Fig. 1b). In addition, the ArgR occupancy level at the promoter of the artJ gene is greater than that of artPIQM operon in the presence and absence of exogenous arginine (Supplementary Table 1). This result is in good agreement with the de-repression/repression ratio of 28 for PartJ and 3.2 for PartP previously reported for repressibility of the artJ and artP promoters18. Also, this result is consistent with recent microarray and qPCR experiments showing a significant arginine and ArgR-dependent down-regulation of both the artJ (about 50-fold) and artPIQM mRNA levels (about three to six-fold)[59]. In the case of TrpR, a total of five associations have been determined by

DNA-binding experiments in vitro and mutational analyses in vivo[51, 61] all of which were also identified in our study (Fig. 1a and Supplementary Table 2). For instance, TrpR directly binds to the promoter regions of aroH and mtr involved in biosynthesis and transport of aromatic amino acids (Fig. 1b). Against the current genome annotation[36], all of the ArgR- and TrpR-binding regions were observed within intergenic regions, i.e., promoter and promoter-like regions. The same preference was observed for Lrp-binding sites (Supplementary Table 1 and 2)[37]. DNA sequence motifs for each of the transcription factors were also re-derived based solely upon the ChIP binding regions and were in full agreement with previously described motifs (Supplementary Fig 2). Based on the fact that the increase in the intracellular arginine and tryptophan levels enhances ArgR and TrpR binding to its DNA targets 20 [62], the confirmation of previously discovered sequence motifs, and the full coverage of the known binding regions in our data we concluded that ArgR- and TrpR-binding regions identified here are bona fide binding sites.

Interestingly, as with gltBD, artPIQM, potFGHI, and mtr (Fig. 1b), we observed that Lrp directly binds to nine ArgR- and one TrpR-binding regions (Fig. 1c and Supplementary Fig. 1). For example, the direct binding of Lrp to the promoter region of the gltBD operon encoding glutamate synthase resulted in the activation of its transcription. In contrast, the role of ArgR-binding represents the negative regulation of the operon. Integrating binding regions and changes in transcript levels, the reciprocal mode3 in the transcriptional regulation of ArgR and Lrp was observed for cellular functions including putrescine transport (potFGHI), arginine transport (artPIQM), leucine response protein (lrp), arginine biosynthesis and utilization (argA and astCADBE), the formation of nucleoid (stpA), as well as glutamate biosynthesis and transport (gltBD and gltP). While Lrp activates the tryptophan transport (mtr), TrpR represses its transcription. In addition to confirming previously identified ArgR- and TrpR-binding regions, we found 48 and 3

Fig. 1

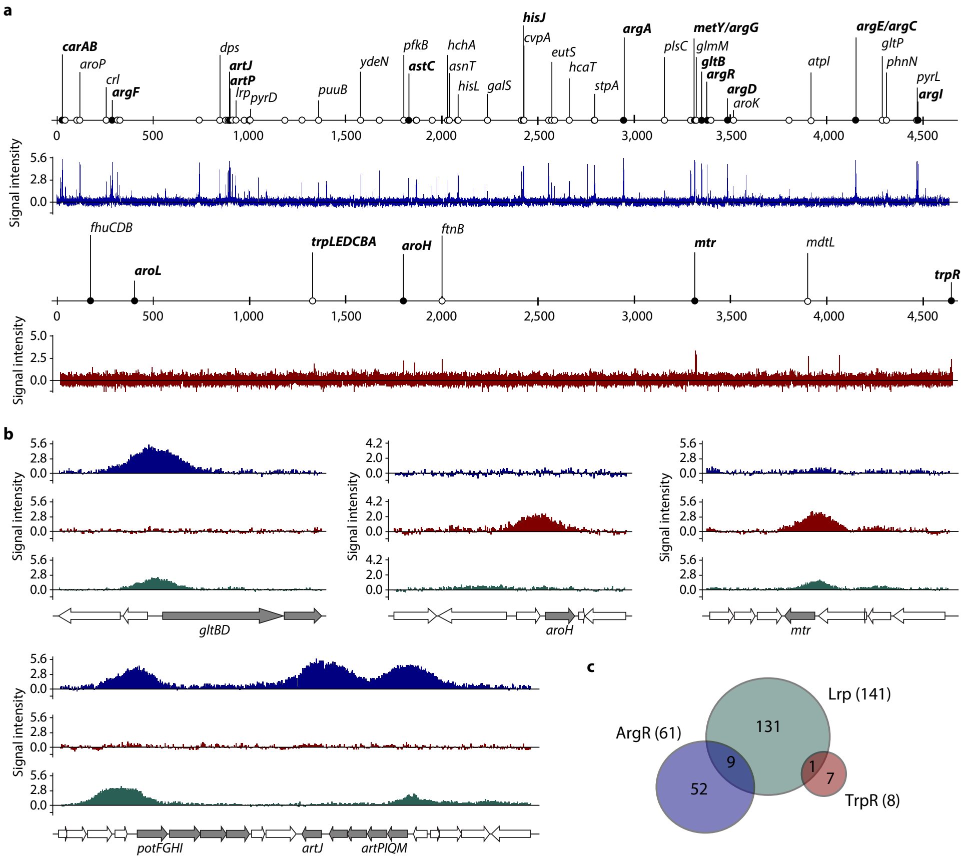


Figure 3.1: **Genome-wide distribution of ArgR- and TrpR-binding regions (regulatory code analysis).** (a) an overview of argR- and TrpR-binding profiles across the E. coli genome in the presence of exogenous arginine (blue track) and tryptophan (yellow track). Enrichment fold on the y axis was calculated from Cy5 (IP Dna) and Cy3 (mock control) signal intensity of each probe and was plotted against each location on the 4.64-Mb E. coli genome. Circles indicate previously identified (black) and newly determined (white) binding regions. Bold targets indicate previously identified binding regions. (b) Examples of genuine argR- (blue track), TrpR- (yellow track) and lrp-binding (green track) areas on the selected regions. The gltBD operon is regulated by both argr and lrp, whereas the aroH gene is regulated solely by TrpR, and mtr by both TrpR and lrp. Both argr and lrp regulate the potFGHI and artPIQM operons; however, only argr regulates artJ. (c) overlaps between argR-, lrp- and TrpR-binding regions.

novel ArgR- and TrpR-binding regions, which include the promoter region of potFGHI, encoding putrescine ABC transporter (Fig. 1b).

3.3.2 Identification of regulons: topological analysis

A regulon is defined as a group of genes whose transcription is controlled by a transcriptional regulator. The arginine regulon describing the genetic and regulatory organization of the genes involved in arginine biosynthesis in E. coli was used as an example in proposing the definition of the regulon in 1964 [59, 63]. However, it has not been included in the definition of regulon whether each regulation is direct or indirect. So far, a total of 37, 56, and 10 genes have been characterized as members of regulons directly regulated by ArgR, Lrp, and TrpR, respectively[57][58]. Based upon regulatory codes described above, we significantly extended the size of these regulons and obtained 140, 283, and 15 target genes for each regulon. Since ArgR directly controls the transcription of lrp, the regulon size of each transcription factor can be described as ArgR (423) ¿ Lrp (283) ¿ TrpR (15). These regulons represent a hierarchical structure that can be used to identify the indirect effect of the TFs. For example, thrLABC operon involved in the threonine biosynthesis is directly activated by Lrp, either in the absence or presence of exogenous leucine. We observed that ArgR indirectly represses this operon in response to exogenous arginine; i.e., transcriptional repression without the direct binding of ArgR. It is therefore possible to partially elucidate the indirect regulation by ArgR based on the hierarchical regulatory network. ArgR represses Lrp leading to the indirect repression of the thrLABC operon. As shown in this example, integrated analysis of ChIP-chip and expression profiles allows us to fully understand the hierarchical TRN including the indirect regulatory effects.

Next, we classified the 438 target genes based on their functional annotation and found that most of these functions ($\tilde{8}2\%$) were assigned to amino acid metabolism and

Fig. 2

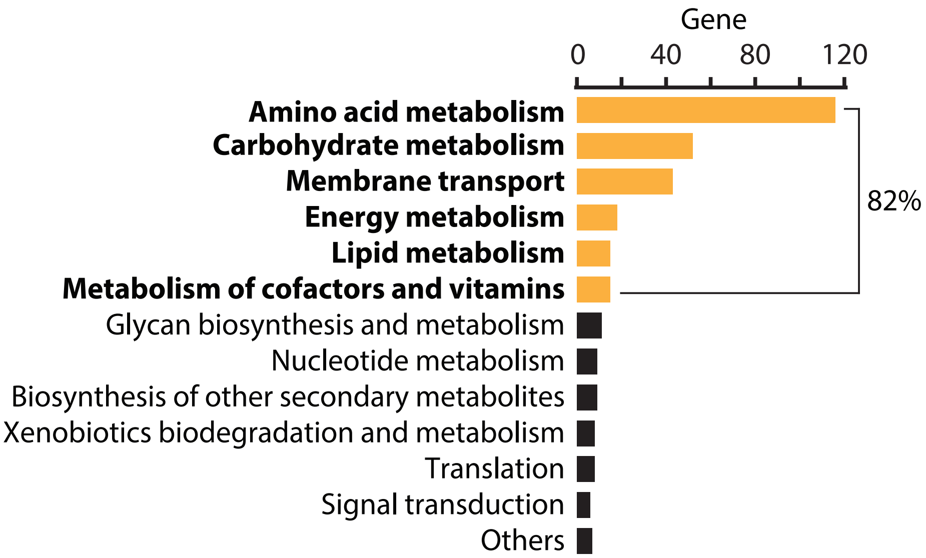


Figure 3.2: **Functional classification of genes directly regulated by ArgR, Lrp and TrpR.** The functions of regulon members are strongly enriched in amino acid metabolism, carbohydrate metabolism, membrane transport (mostly amino acid related) and energy metabolism (yellow bars).

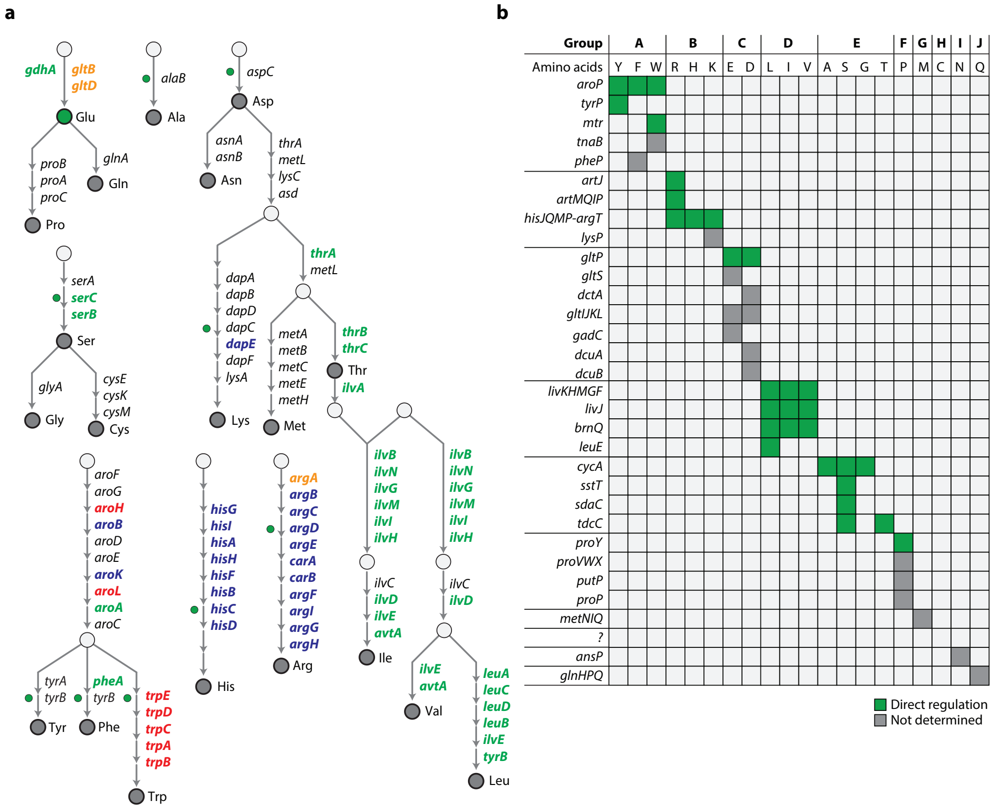
transport, as well as carbohydrate, nucleotide, and energy metabolism (Fig. 2). We are then able to show (Fig. 3) that 19/20 amino acid biosynthetic pathways are directly or indirectly controlled by these three TFs. To do this we first mapped the directly regulated genes to known amino acid biosynthetic pathways and transport systems to determine their direct metabolic roles (Fig. 3a, b). ArgR directly regulates the transcription of all genes involved in the biosynthesis of arginine and histidine. It also regulates gltBD, aroB, aroK, and dapE involved in glutamate, aromatic amino acids, and lysine biosynthesis, respectively. The genes encoding the enzymes for the biosynthesis of branched chain amino acids are comprehensively regulated by Lrp, which also controls the transcription of gltBD and gdhA encoding glutamate synthase and glutamate dehydrogenase (glutamate biosynthesis), serC and serB encoding phosphoserine transaminase and phosphatase (serine biosynthesis), thrABC operon for aspartate kinase, homoserine kinase, and threonine synthase (threonine biosynthesis), argA for N-acetylglutamate synthase (arginine biosynthesis), and aroA for 3-phosphoshikimate-1-carboxyvinyltransferase (the chorismate formation for aromatic amino acid biosynthesis). TrpR regulates the transcription of genes involved in tryptophan biosynthetic pathway (trpLEDCBA operon), as well as aroH and aroL. In addition, it has been determined that TyrR directly regulates several genes in the aromatic amino acid biosynthesis (aroF, aroG, aroK, aroA, tyrA, and tyrB) in response to exogenous tyrosine[57, 58]. Taken together, these four TFs control the biosynthesis of 12 amino acids. Furthermore, the biosynthesis of proline, glutamine, glycine, cysteine, and methionine is through branched biosynthetic pathways of glutamate, serine and aspartate (Fig. 3a). The remaining three amino acids (i.e., alanine, aspartate, and asparagine) are synthesized from glutamate as an amino donor (green dots in Fig. 3a). Therefore, biosynthetic pathways for all amino acids are directly or indirectly controlled by these four TFs.

Next, we classified the amino acids into ten groups based on the substrate specificity of each transport system, which are A (tyrosine, phenylalanine, tryptophan), B (arginine, histidine, lysine), C (glutamate, aspartate), D (leucine, isoleucine, valine), E (alanine, serine, glycine, threonine), F (proline), G (methionine), H (cysteine), I (asparagine), and J (glutamine) (Fig. 3b). This was done based on the primary literature and EcoCyc[58]. As expected, the amino acids in the same group have a similar chemical structure, e.g. aromatic amino acids and branched chain amino acids in group A and group D, respectively. Transport systems for groups G-J are highly specific and were therefore classified into individual groups.

3.3.3 Determination of causal relationships: functional analysis

In general, genes for amino acid biosynthesis are repressed by each corresponding TF, whereas catabolic operons such as astCADBE, tdh-kbl, and gcvTHP are induced in response to the exogenous amino acids[56, 64]. To determine the causal relationships between binding of a TF and the changes in RNA transcript levels of genes in the regulons, we integrated the binding regions of ArgR, TrpR, Lrp, and TyrR with the publicly available transcriptomic data (Fig. 4)[37, 59]. We then determined activation or repression based upon the regulatory modes described previously[37]. Among genes in the ArgR regulon, about 18% genes were directly activated in response to the exogenous arginine, which include aroP and gltP genes encoding aromatic amino acids and glutamate/aspartate transporters. On the other hand, ArgR represses about 70% of its regulon members, including potFGHI, artJ, artPIQM, and hisJQMP encoding putrescine, arginine, lysine, ornithine, and histidine ABC transporters (Fig. 4). ArgR represses genes involved in the arginine and glutamate biosynthesis pathways, and unexpectedly, it directly down-regulates genes involved in histidine, aromatic amino acids, and lysine biosynthesis pathways. In case of amino acid utilization, ArgR induces astCADBE and puuEB operons encoding the metabolic pathways for arginine and putrescine, respectively. The remaining 12% of its regulon members had a direct association with ArgR without differential gene expression. Most of the remaining genes are currently annotated as genes of unknown function (Supplementary Table 1).

Fig. 3



Group	A			B			C		D			E				F	G	H	I	J
Amino acids	Y	F	W	R	H	K	E	D	L	I	V	A	S	G	T	P	M	C	N	Q
aroP	🟩	🟩	🟩																	
tyrP	🟩																			
mtr			🟩																	
tnaB			▨																	
pheP		▨																		
artJ				🟩																
artMQIP				🟩																
hisJQMP-argT				🟩	🟩	🟩														
lysP						▨														
gltP							🟩	🟩												
gltS							▨													
dctA								▨												
gltIJKL							▨	▨												
gadC							▨													
dcuA								▨												
dcuB								▨												
livKHMGF									🟩	🟩	🟩									
livJ									🟩	🟩	🟩									
brnQ									🟩	🟩	🟩									
leuE									🟩											
cycA												🟩	🟩	🟩						
sstT													🟩							
sdaC													🟩							
tdcC													🟩		🟩					
proY																🟩				
proVWX																▨				
putP																▨				
proP																▨				
metNIQ																	▨			
?																				
ansP																			▨	
glnHPQ																				▨

🟩 Direct regulation
▨ Not determined

Figure 3.3: **Delineation of amino acid biosynthetic pathways and transport systems in E. coli (topological analysis).** (a) The amino acid biosynthetic pathways. The genes are directly regulated by argr (blue), lrp (green) and Trpr (red), respectively. The genes shown in bold black (gltB and gltD in glutamate biosynthesis and argA in arginine biosynthesis) are regulated by both argr and lrp. orange dots indicate the biosynthetic reactions, which use glutamate as an amino donor. (b) The amino acid transport systems. The genes encoding each transport system can be classified into ten groups (aj) based on amino acid specificity. The transport systems in orange are directly regulated by argr, lrp or Trpr. For others (gray), the transcriptional regulation has not been determined.

Gene expression profiles validated that Lrp directly regulates 283 genes. 45% and 55% of the Lrp-regulated genes were repressed and activated in response to the addition of the exogenous leucine[37]. As expected, Lrp controls the transport, biosynthetic and utilization pathways more globally than other transcription factors do. This is due to its known role as a global regulator of metabolism and nucleoid structure[56]. Lrp represses the transport systems for branched chain amino acids (brnQ, livKHMGF, and livJ), dipeptides (dppABCDF), and lipoproteins (lolCDE) but it activates a whole set of other transporters. Transporters that are activated by Lrp are aromatic amino acids (tyrP and mtr), arginine (artMQIP), glutamate (gltP), alanine, serine, glycine and threonine (cycA, tdcC, sdaC, and sstT), proline (proY), putrescine (potFGHI), dipeptide (dtpB), and oligopeptides (oppABCDF) (Fig. 4). In terms of amino acid biosynthetic pathways, Lrp represses all genes but the thrLABC operon for threonine biosynthesis. For amino acid utilization, Lrp activates all pathways for aromatic amino acids, arginine, aspartate, branched chain aromatic amino acids, alanine, glycine, serine, threonine, methionine, and putrescine. In case of the TrpR regulon, a total of 15 genes are directly regulated, of which 13 genes are repressed (Supplementary Table 2)[58, 65]. TrpR also represses mtr encoding the tryptophan transporter as well as aroH, aroL, and trpABCDE involved in the tryptophan biosynthesis pathway. While TyrR activates the transport systems for aromatic amino acids (aroP, tyrP, and mtr), it represses tyrosine biosynthetic pathway comprising of aroG, aroL, aroF, tyrA, and tyrB (Fig. 4).

3.3.4 Elucidation of regulatory logic

Based on the integrated analysis of TF-binding locations and gene expression profiles, we were able to connect transport, biosynthesis, and utilization of amino acids, and generate the connected bidirectional circuits (Fig. 5a). In the left feed-back circuit, TF-amino acid (TF-AA) complexes regulate the transcription of the transporters (T) and biosynthesis pathways (B), facilitating the influx of the amino acid molecules (AAin) from amino acids in the media (AAout) and precursors (AApre). In the right feed-forward circuit, TF-AA complexes control transcription of utilization genes (U) responsible for converting AAin into metabolites (M). Thus, the logical structures of the connected bidirectional circuit motifs can be described by a notation that uses three signs indicating repression (R) or activation (A) for each of T, B, and U (Fig. 5b). For example, the A-R-A circuit motif indicates that the transcription of transport, biosynthesis, and metabolic genes are activated, repressed, and activated, respectively, whereas the R-R-A circuit motif demonstrates that the transcription of both transport and biosynthesis are repressed and the metabolic genes are activated. The possible logical structures of the connected circuit motifs can be characterized depending on how the TF-AA complex activates or represses both influx (T and B) and efflux (U) in response to the exogenous amino acids. Based on the connected circuit motifs, we analyzed the behavior of logical structures of the transcription of transport, biosynthesis, and metabolic genes in responses to the exogenous arginine and leucine (Fig. 5b).

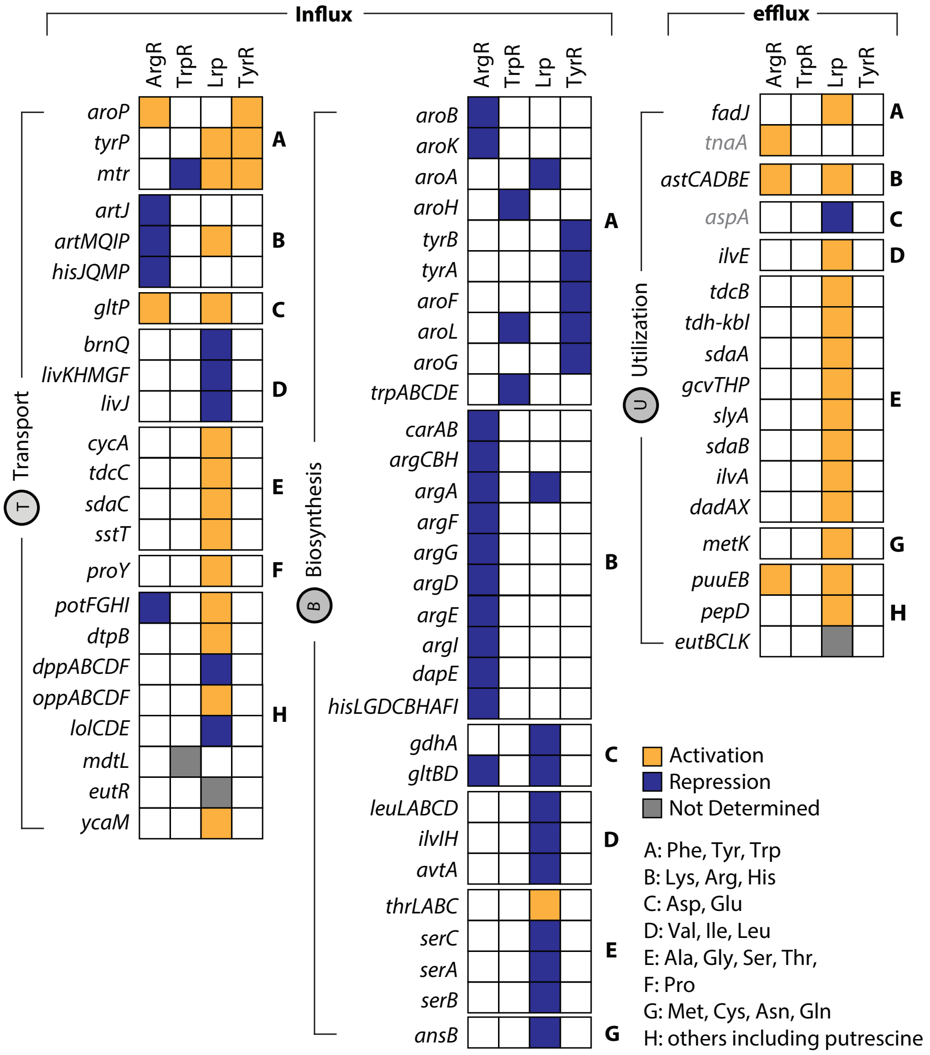
Surprisingly, there are only three influx-efflux combinations found between amino acid groups and TFs (Fig. 5c). For example, the connected circuit motif controlled by ArgR-arginine complex shows the R-R-A logical structure for group B amino acids (lysine, histidine, and arginine), whereas the logical structure of the motif is switched to A-R-R for glutamate and aspartate and A-R-A for other amino acids. On the other hand, the connected motif controlled by Lrp-leucine complex indicates the R-R-A logical structure for group D (valine, leucine, and isoleucine) and is again switched to A-R-R for glutamate and aspartate and A-R-A for other amino acids. For glutamate our primary observation was that the utilization was repressed given its role as a substrate for nine biosynthetic pathways (Fig. 3,4). However we acknowledge that the regulation is highly complex and not universally repressed. This logically follows from the critical and centralized role it plays throughout the metabolome[66]. Overall, we conclude that for two global transcription factors (ArgR and Lrp) in amino acid regulation, the connected circuit motif has an R-R-A logical structure for signaling molecules (i.e., arginine for ArgR and leucine for Lrp) and the A-R-A and A-R-R logical structures for other amino acids (Fig. 5c).

3.4 Discussion

We reconstructed the regulons of ArgR, Lrp, and TrpR in E. coli individually and then integrated them to form the first genome-scale reconstruction of a stimulon. First, we set out to comprehensively establish the TF-binding regions on the E. coli genome experimentally and furthermore to elucidate any DNA sequence motif(s) correlated with the TF regulatory action. Second, we significantly extended the size of each regulon and obtained 140, 283, and 15 target genes for each regulon. Third, using changes in transcript levels on a genome-scale, we identified the regulatory modes for individual gene governed by each TF in responses to exogenous arginine, leucine, and tryptophan. The integrated analyses indicate that the functional assignment of the regulated genes is strongly enriched in the amino acid metabolism-related functions. As suggested previously, many of these genes are likely to be involved in the feast or famine adaptation for survival in nutrient-rich or depleted environments [37, 53]. Fourth, we assigned the regulated target genes to three functional categories; transport, biosynthesis, and metabolism of amino acids. The classification allowed us to identify the connected circuit motif as a basic building block of the integrated network. Finally, we determined the regulatory logic of the connected circuit motif based on the causal relationships between the association of TFs and changes in transcript levels. These fall into two categories and thus allow for the differentiation between amino acids as signaling and nutrient molecules.

In general, transport systems along with biosynthetic and metabolic pathways

Fig. 4



Transport	ArgR	TrpR	Lrp	TyrR	
aroP	🟧			🟧	A
tyrP			🟧	🟧	
mtr		🟦	🟧	🟧	
artJ	🟦				B
artMQIP	🟦		🟧		
hisJQMP	🟦				
gltP	🟧		🟧		C
brnQ			🟦		D
livKHMGF			🟦		
livJ			🟦		
cycA			🟧		E
tdcC			🟧		
sdaC			🟧		
sstT			🟧		
proY			🟧		F
potFGHI	🟦		🟧		H
dtpB			🟧		
dppABCDF			🟦		
oppABCDF			🟧		
lolCDE			🟦		
mdtL		▨			
eutR			▨		
ycaM			🟧		

Biosynthesis	ArgR	TrpR	Lrp	TyrR	
aroB	🟦				A
aroK	🟦				
aroA			🟦		
aroH		🟦			
tyrB				🟦	
tyrA				🟦	
aroF				🟦	
aroL		🟦		🟦	
aroG				🟦	
trpABCDE		🟦			
carAB	🟦				B
argCBH	🟦				
argA	🟦		🟦		
argF	🟦				
argG	🟦				
argD	🟦				
argE	🟦				
argI	🟦				
dapE	🟦				
hisLGDCBHAFI	🟦				
gdhA			🟦		C
gltBD	🟦		🟦		
leuLABCD			🟦		D
ilvIH			🟦		
avtA			🟦		
thrLABC			🟧		E
serC			🟦		
serA			🟦		
serB			🟦		
ansB			🟦		G

Utilization	ArgR	TrpR	Lrp	TyrR	
fadJ			🟧		A
tnaA	🟧				
astCADBE	🟧		🟧		B
aspA			🟦		C
ilvE			🟧		D
tdcB			🟧		E
tdh-kbl			🟧		
sdaA			🟧		
gcvTHP			🟧		
slyA			🟧		
sdaB			🟧		
ilvA			🟧		
dadAX			🟧		
metK			🟧		G
puuEB	🟧		🟧		H
pepD			🟧		
eutBCLK			▨		

🟧 Activation
🟦 Repression
▨ Not Determined

A: Phe, Tyr, Trp
B: Lys, Arg, His
C: Asp, Glu
D: Val, Ile, Leu
E: Ala, Gly, Ser, Thr,
F: Pro
G: Met, Cys, Asn, Gln
H: others including putrescine

Figure 3.4: **Causal relationships between direct associations of transcription factors and the changes in RNA transcript levels of genes (functional analysis).** Regulated genes are broken down into three main categories: transport, biosynthesis or utilization of respective amino acids. They are further broken down on the basis of their amino acid specificities and pathway participation. Here we include (in gray) tnaA (Students t-test P = 3.29 102, factor of 1.95) and aspA (Students t-test P = 9.88 102, factor of 7.62) as indirectly regulated genes with differential expression but no ChIP enrichment to fully capture the utilization response to arginine. For class C, we note that the direct targets of utilization for glutamate are in fact the biosynthetic genes that are shown in that section and are pointed out in Figure 3.

convert external resources to basic building blocks to sustain life. The coordinated regulation of this primary process underlies expression of optimized metabolic states under different external conditions. Thus, we examined the logical structures of the metabolite-regulation connected circuit in response to the changes in the external amino acid availability in the reconstructed stimulon. We uncovered three unique logical structures that govern the amino acid biosynthesis and metabolism. The R-R-A logical structure was observed for signaling molecules whereas the A-R-A and A-R-R logical structures were determined for other amino acids severing as nutrient source (Fig. 5a, b). In principle, every metabolic pathway that includes transport, biosynthesis, and utilization functions could follow these logical structures. For example, the purine metabolism in E. coli contains a wide range of genes whose functions are transport (yieG), biosynthesis (cvpA-purF-ubiX, purHD, purMN, purT, purL, purEK, purC, hflD-purB, purA, and guaAB), utilization (apt), and a transcriptional regulator (purR). The metabolic functions of regulon members of PurR enriched into the purine metabolism and the connected circuit motif indicated the logical structures for signaling molecule in response to the exogenous purine[67]. It can be therefore envisioned that other potential metabolic pathways follow similar logical structures as determined for the amino acid metabolism in bacteria.

Bacterial cells import essential nutrients and inorganic ions such as galactose and iron due to the absence of the biosynthesis pathway. It is therefore of interest that the simple feedback circuit (SFL) motif, a connected circuit motif of transporter and utilization pathway by TF, is often observed in the regulatory circuits for these molecules[43]. If we assume the feedback circuit composed of influx and efflux combination, the logical structures of R-R-A, A-R-A, and A-R-R in the CFL motif can be reduced to R-A, A-A, and A-R, respectively. In E. coli, the galactose metabolic pathway is controlled by the galactose repressor (GalR) and galactose isorepressor (GalS), whereas iron homeostasis is controlled by the ferric uptake regulator (Fur)[68, 69]. In the case of galactose metabolism, both GalR and GalS directly repress the transcription of galP encoding galactose permease. In a similar way, GalR partially represses the mglBAC operon encoding high-affinity, ABC-type transport system. When galactose is available in the medium, the DNA-binding by both GalR and GalS is inhibited, followed by the activation of those genes along with the genes for galactose utilization[69]. Therefore, the SFL motif exhibits the A-A logical structure, confirming the exogenous galactose as nutrient. In the iron homeostasis system in E. coli, intracellular iron binds to Fur, forming the active TF complex, which in turn activates the production of iron-using metabolic enzymes and also shuts down expression of iron transporters. Interestingly, the SFL motif for Fur regulon exhibits the R-A logical structure, similar to amino acids serving as signaling molecules described above. Therefore, we can conclude that iron acts as signaling molecule rather than nutrient.

In summary, we have described an integrative analysis of genome-scale data sets to comprehensively understand the basic principles governing a stimulon in the TRN of E. coli. The overarching regulatory principle elucidated enabled us to differentiate between metabolites as signaling and nutrient molecules. This important distinction between seemingly similar metabolites is non-intuitive and could only be determined through genome-scale systems analysis. Similar analysis of other stimulons and large-scale regulatory networks may reveal that this regulatory principle is general. Thus, this approach to the analysis of regulation at the network level may reveal other fundamental non-obvious regulatory principles at work in genome-scale regulatory networks.

3.5 Methods

3.5.1 Bacterial strains and growth conditions

All strains used are E. coli K-12 MG1655 and its derivatives. The E. coli strains harboring ArgR-8myc, Lrp-8myc, and TrpR-8myc were generated as described previously13. Glycerol stock of ArgR-8myc strains were inoculated into W2 minimal medium containing 2 g/L glucose and 2g/L glutamine, and cultured overnight at 37 oC with constant agitation30. The cultures were inoculated into 50 mL of the fresh W2 minimal media in either the presence or absence of 1 g/L arginine and continued to culture at 37 oC with constant agitation to an appropriate cell density. E. coli strains harboring Lrp-8myc and TrpR-8myc were grown in glucose (2 g/L) minimal M9 medium supplemented with or without 20 mg/L tryptophan or 10 mM leucine, respectively [37, 70].

3.5.2 Chromatin immunoprecipitation and microarray analysis

To identify ArgR-, Lrp-, and TrpR-binding regions in vivo, we isolated the DNA bound to ArgR protein from formaldehyde cross-linked E. coli cells harboring ArgR-8myc by chromatin immunoprecipitation with the specific antibodies that specifically recognizes myc tag (9E10, Santa Cruz Biotech)[41]. Cells were harvested from the exponential growth conditions in the presence or absence of exogenous arginine or tryptophan. The immunoprecipitated DNA (IP-DNA) and mock immunoprecipitated DNA (mock IP-DNA) were hybridized onto the high-resolution whole-genome tiling microarrays, which contained a total of 371,034 oligonucleotides with 50-bp tiles overlapping every 25-bp on both forward and reverse strands[37, 36]. A ChIP-chip protocol previously described was used[41, 71] and microarray hybridization, wash, and scan were performed in accordance with manufacturers instruction (Roche NimbleGen).

3.5.3 qPCR

To monitor the enrichment of promoter regions, 1uL immunoprecipitated DNA was used to carry out gene-specific qPCR3. The quantitative real-time PCR of each sample was performed in triplicate using iCycler (Bio-Rad Laboratories) and SYBR green mix (Qiagen). The real-time qPCR conditions were as follows: 25uL SYBR mix (Qiagen), 1uL of each primer (10 pM), 1uL of immunoprecipitated or mock-immunoprecipitated DNA and 22uL of ddH2O. All real-time qPCR reactions were done in triplicates. The samples were cycled to 94 oC for 15 s, 52 oC for 30 s and 72 oC for 30 s (total 40 cycles) on a LightCycler (Bio-Rad). The threshold cycle values were calculated automatically by the iCycler iQ optical system software (Bio-Rad Laboratories). Primer sequences used in this study are available on request.

3.5.4 ChIP-chip and expression data analysis

To identify TF-binding regions, we used the peak finding algorithm built into the NimbleScanTM software. Processing of ChIP-chip data was performed in three steps: normalization, IP/mock-IP ratio computation (log base 2), and enriched region identification. The log2 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 bi-weight mean34. The log2 ratio of Cy5 (IP DNA) to Cy3 (mock-IP DNA) for each point was calculated from the scanned signals. Then, the bi-weight mean of this log2 ratio was subtracted from each point. Each log ratio dataset from duplicate samples was used to identify TF-binding regions using the software (width of sliding window = 300 bp). Our approach to identify the TF-binding regions was to first determine binding locations from each data set and then combine the binding locations from at least five of six datasets to define a binding region using the recently developed MetaScope software (Kim et al. Submitted). Raw gene expression CEL files were gathered from GEO for ArgR with accession GSE4724 and for Lrp from a previous study3. They were normalized using background corrected robust multi-array average implemented in the R affy package. To detect differential expression between the wild type and TF deletion strains we applied a two-tailed unpaired students t-test with Microsoft excel between the experimental triplicates for the wild type and gene deletion strains. This was followed by a false discovery rate39 adjustment using the R statistical software package. Before performing the FDR correction we removed all genes which exhibited an expression level below the background across all experiments. The background level was calculated as the average expression level across all intergenic probes. We then only considered genes meeting a 5% FDR (false discovery rate)-adjusted P-value cut-off to be differentially expressed. To make calls for activation or repression we followed the methodology laid out previously[37].

3.5.5 Motif searching

The ArgR-, Lrp-, and TrpR-binding motif analysis was completed using the MEME and FIMO tools from the MEME software suiteBailey:2009eu. We first determined the proper binding motif and then scanned the full genome for its presence. The elicitation of the motif was done using the MEME program on the set of sequences defined by the ArgR-, Lrp-, and TrpR-binding regions respectively[72]. Using default settings the previously determined ArgR[59], Lrp[37], and TrpR[51] motif were recovered and then tailored to the correct size by setting the width parameter to 18-bp, 15-bp, and 8-bp respectively. We then used these motifs and the PSPM (position specific probability matrix) generated for each by MEME to rescan the entire genome with the FIMO program. The sequence logo generated from these sites.

3.6 Acknowledgements

Chapter 3 in full is a reprint of a published manuscript: Cho, B.K., Federowicz, S.A., Park, Y.S., Zengler, K., Palsson B.. Deciphering the transcriptional regulatory logic of amino acid metabolism. Nature Chemical Biology. 2011 Nov 13;8(1):65-71. doi: 10.1038/nchembio.710. The dissertation author was the second author of this paper, responsible for the computational analysis. The other authors were Byung-Kwan Cho, Young-Seob Park, Karsten Zengler, and, Bernhard Ø. Palsson.

Chapter 4

The OME Framework for genome-scale systems biology

4.1 Abstract

A central aim of systems biology is to elucidate emergent properties of living organisms through data analysis and computational model simulations. However, it remains challenging to manage and integrate data and models in workflows that extract novel biological insights. Here we present the OME Framework (operational, metabolic, expression) is a systems biology and bioinformatics software framework that helps to enable a broad array of iterative and evolving systems biology workflows. The framework consists of three core modules: models, components, and datasets. These modules are defined by an underlying database schema, which is implemented into common Python objects through the object-relation mapper (ORM) SQLAlchemy (sqlalchemy.org). The models module stores, manages, and provides access to the rich structure of genome-scale biological models. The components module stores and makes accessible a full spectrum of biological parts, from metabolites and enzymes to genes and genomes. The datasets

module provides a generic structure for storing any type of biological data that maps to a given component or model. The key features of the OME Framework arise from the interactions among these three modules; connected modeling integrates models with detailed genomic and biochemical data, connected analysis integrates high-throughput data analysis with the same set of genomic and biochemical data, and data-mapped modeling provides an implicit mapping between experimental data and genome-scale biological models. While the current implementation of the OME Framework exists in Python with SQLAlchemy, the underlying schematics and software design can be seen as an addition to the community of standards for systems biology. The OME Framework makes many contributions including the simple handling of arbitrarily complex biological models, a generic and extensible universal biological parts database, and an explicit mapping between experimental data and computational models. Taken together, the OME Framework takes a significant step in helping to reduce to practice the design-build-test loop at a genome-scale. The code is fully open source and ready to be forked on GitHub (http://github.com/SBRG/ome).

4.2 Introduction

The discipline of biological sciences is undergoing a continuous and accelerating transformation with its merger into the computational and engineering sciences. The biology that many in the field have been trained on may be hardly recognizable in ten to twenty years. One of the major drivers for this transformation is the blistering pace of advancements in DNA sequencing and DNA synthesis. This has resulted in unprecedented amounts of new data, information, and knowledge.

Many software tools have been developed to deal with aspects of this transformation and each is sorely needed [73, 74, 75]. However, few of these tools have been forced to deal with the full complexity and scale of genome-scale models along with high throughput genome-scale data. This particular situation represents a unique challenge, as it is simultaneously necessary to deal with the vast breadth of genome-scale models and the dizzying depth of high-throughput datasets. It has been observed time and again that as the pace of data generation continues to accelerate, the pace of analysis significantly lags behind[76]. It is also evident that, given the plethora of databases and software efforts [77, 78, 79, 80, 81, 82, 3, 83], there is a significant challenge in working with even genome-scale metabolic models, let alone next-generation whole cell models [84, 12, 85]. We work at the forefront of model creation (ME model) and systems scale data generation [86, 87, 88]. The OME Framework was borne out of a practical need to enable genome-scale modeling and data analysis under a unified framework to drive the next generation of genome-scale biological models.

Here we present the OME Framework exists as a set of Python classes. However, we want to emphasize the importance of the underlying design as an addition to the discussions on specifications of a digital cell. A great deal of work and valuable progress has been made by a number of communities [84, 89, 90, 91, 92, 93, 94] towards interchange formats and implementations designed to achieve similar goals. While many software tools exist for handling genome-scale metabolic models or for genome-scale data analysis, no implementations exist that explicitly handles data and models concurrently. The OME Framework structures data in a connected loop with models and the components those models are composed of. This results in the first full, practical, implementation of a framework that can enable genome-scale design-build-test. Over the coming years many more software packages will be developed and tools will necessarily change. However, we hope that the underlying designs shared here can help to inform the design of future software.

4.3 Implementation

The OME Framework is currently a set of Python classes that act as an object-relational mapper (ORM) between Python objects and their corresponding relational tables in the OME database. This is implemented almost fully with the SQLAlchemy ORM (sqlalchemy.org) such that database table definitions are made through the Python class attributes. This has the benefit of providing a clear and centralized location for the database schema definition. It also enables the use of the framework across multiple database backends, including but not limited to SQLite, PostgreSQL, MySQL, and Oracle databases. This becomes important for rapid deployment and integration into existing hardware setups where it may not be practical or desirable to install a new database.

The framework is semantically split into three major classes: models, components, and datasets (Figure 1). It also contains a fourth class, base, which is necessary for the usage of any individual class. The base class contains the database connection information, configuration settings, and table definitions for elements that are common across models, components, and datasets. This includes genome and chromosome tables, a synonyms table, and the actual components table among others.

In practice, each module can be used independently without any need for the other modules. However, upon installation, the OME database namespace will become populated with all of the tables from each module. This is because the native install currently contains example data, including genome-scale metabolic models for E. coli (iJO1366) and S. cerevisiae (iMM904), the associated genbank genome annotation files, and gene expression and ChIP-chip datasets[86]. If you drop the database and follow the data_loading examples you will be able to load any of the namespaces individually.

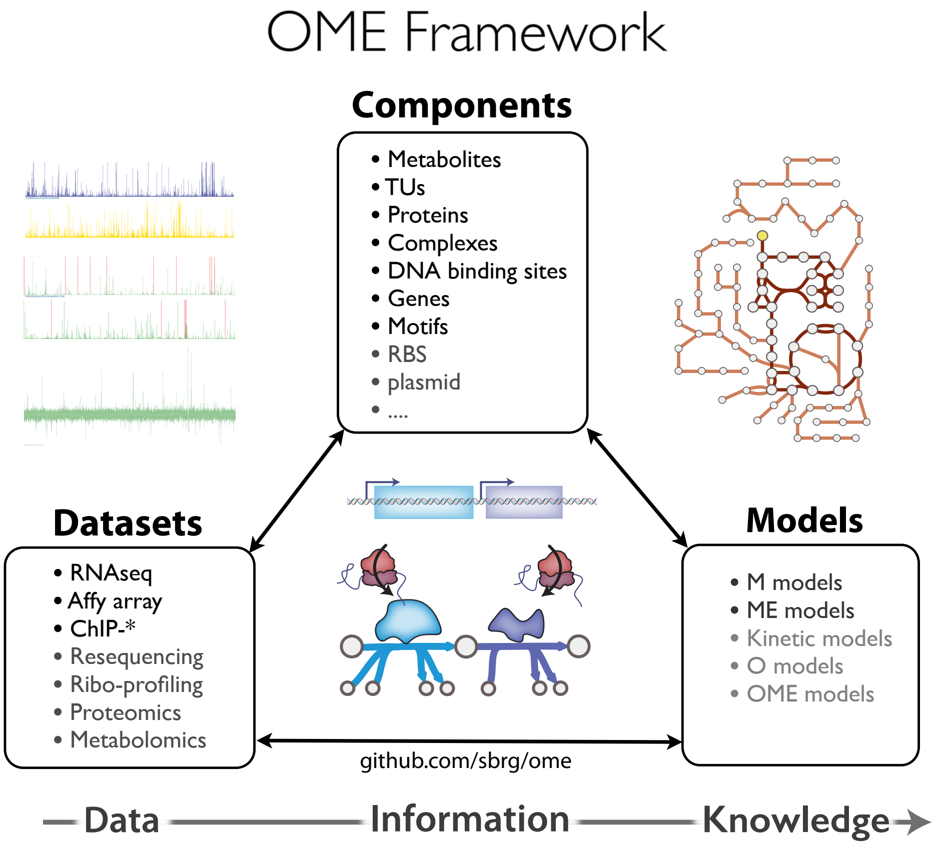


Figure 4.1: **Birds eye view of the OME Framework.** The OME Framework consists of three core modules; datasets, components, and models. Raw and processed data from experiments or analyses along with associated metadata are stored in the datasets module. These data are automatically mapped to an array of stored biological components. These biological components then form the basis as products and reactants in the reactions that make up genome-scale models. Finally, through components, all experimental data and analyses are implicitly mapped to genome-scale biological models. This overall structure can also be seen to reflect the generic progression from data to information to knowledge which itself mirrors the metabolic reconstruction process.

4.3.1 Models

The models class (Figure 2) stores, manages, and provides access to the rich structure of genome-scale biological models. The basic schema and design of the models subsection aims to provide a few key elements.

The first is the storage and versioning of a large number of biological models. The models table (Figure 2) and its resulting interconnections enable an arbitrarily large number of models to be stored along with a corresponding time stamp to provide appropriate versioning. This simple versioning is crucial to help ensure quality and provide a scaffold for future libraries to implement a more rigorous tracking of model changes and curation efforts.

The second is consolidation of redundant elements into universal reaction and component tables. By separating out model_reactions and model_components from standalone reaction and component tables we can achieve the effect of universal metabolite and reaction tables. These tables are crucial for reconstruction efforts and general utility. The knowledge of the full number of unique metabolites or reactions in a given set of models is often crucial information for various algorithms and design approaches.

The third is the ability to enable comparisons among stored models. The inherent structure of the models schema enables an easy comparison among multiple models. Part of this analysis comes part in parcel with the aforementioned universal reaction and component tables. However, one can also easily write additional sets of queries to determine outliers in specific models or to determine common reaction structures among many models.

The fourth is the ability to store a wide variety of biological models without being limited to constraint-based metabolic models. The key feature of the models schema is the use of the reaction_matrix table along with universal reactions and components tables. Since a component is defined through the polymorphic star based schema introduced in (Figure 3). A reaction can contain any arbitrary type of component in any type of arbitrary stoichiometry. This is a key feature that enabled earlier versions of the OME framework to handle and manage the large and structurally diverse models, such as ME models[14, 15].

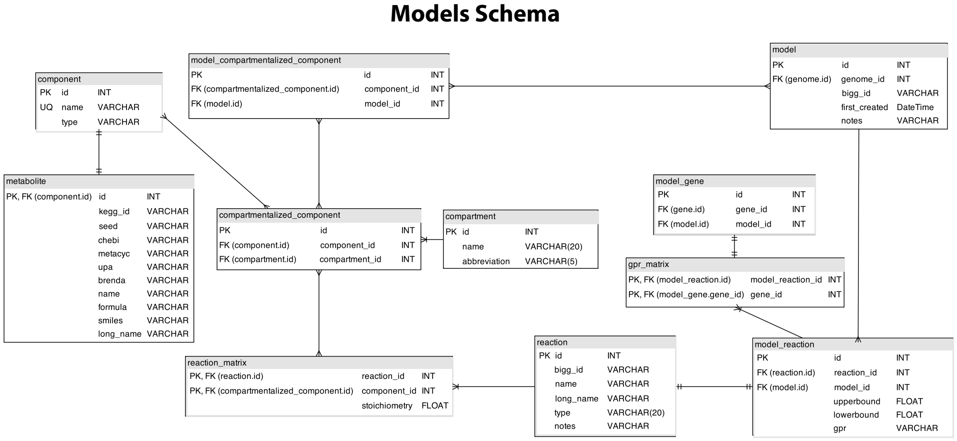


Figure 4.2: **Relational database schema for the datasets module.** The datasets module has a star based schema in which the core dataset table is the parent of any individual dataset type. Each dataset has a strain, environment, and a datasource. However, specific datasets, e.g. ChIPExperiment can also have specific values such as the antibody or protocol used. In this fashion the appropriate metadata can be captured for any specific datatypes without having to add the redundant information of strain, environment, and and datasource for each new table. This also results in a simplified mapping to components as a minimal number of tables can be used to link to datasets without needing many individual tables linking to each individual datatype

4.3.2 Components

The components class (Figure 3) stores and makes accessible a full spectrum of biological parts, from metabolites and enzymes to genes and genomes. The key contribution in the architecture of the components module is the use of a Star Schema that provides two crucial advantages.

The first is the ability of a component to be a complex (e.g., an enzyme complex or ribosome), which enables the creation of complexes of arbitrary components. The result is a simple and generic structure upon which biological parts can be broken down and aggregated into arbitrary modules and sub-modules. Standard and well-known complexes like protein-protein and protein-dna complexes are of course stored. However, GPRs (Boolean representations of gene-protein-reaction associations), which have been previously stored in a string format can now be represented properly in a tree based structure as components of components. While the SQL schema to enable complexes of complexes (the complex_composition table) is relatively succinct, the implementation can quickly become unwieldy and impractical. However, SQLAlchemy can succinctly enabling object based tree traversal. This is done by encapsulating the recursive SQL queries in SQLAlchemy attributes, which allows them to be accessed as normal class properties. The second is to drastically reduce the relational complexity in integrating components with models and data. The most obvious way to design the models and components aspects of the OME framework is to develop a large number of linker tables that exist between each component and the table it belongs to. An example would be for metabolites to have a table called metabolite_reaction that would contain a reaction_id and a metabolite_id and thus provide a normalized mapping between reactions and metabolites. However, imagine you also wanted to include a protein into a reaction as the enzyme to catalyze it. You could create another table called protein_reaction and again put the reaction_id and the corresponding protein_id into the table. However, this strategy would lead to tables for rna_reaction, dna_reaction, etc., and the result would be an inefficient and overly complex schema.

4.3.3 Datasets

The datasets class (Figure 4) provides a generic structure for storing any type of biological data that maps to a given component or model. The design of the datasets module is similar to that of the components module. A Star Schema is used with a

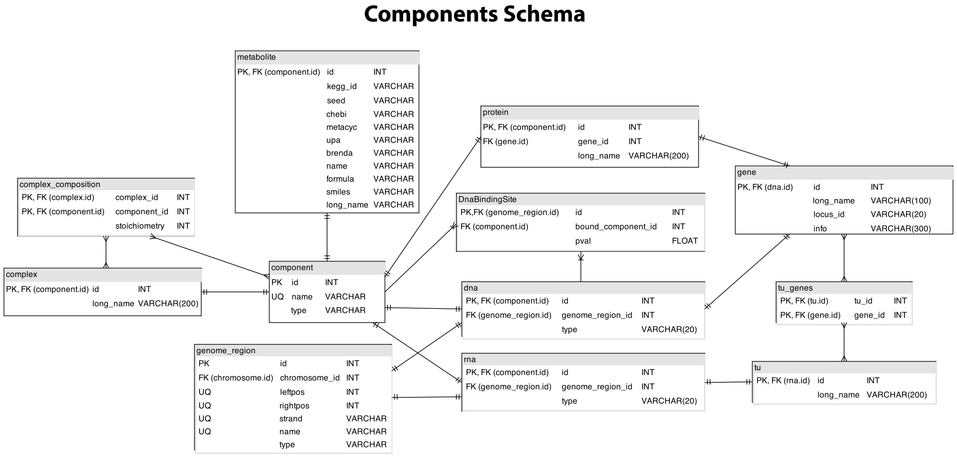


Figure 4.3: **Relational database schema for the components module.** The components module also has a star based schema in which the core components table is the parent of any individual component type. Each component simply has a unique name and an ID that is then inherited by each of the specific tables common biological parts. The definition of a component is taken to be any physical entity within a cell. This distinguishes a genome region from a component as the genome region is taken as a piece of information describing the component of the underlying DNA strand. This design decision simplifies model building because the only that are allowed to be part of a reaction in a model are components.

generic datasets object that is then inherited by each individual experiment or analysis type. The first main advantage provided by the datasets module is the ability to store an arbitrary array of experimental and computational datasets in a common schema. This enables analysis and integration across diverse datatypes in a controlled and reproducible manner. It also enables the storage of both experimental and computationally generated data side by side. This particular use case is discussed further as a key result but in short this provides for the rapid assessment of predictive computational algorithms. This will be crucial for systems-level biological engineering and the design-build-test cycle. The second key advantage is that the use of a star based schema allows for a large number of differing and diverse datatypes to be associated with a similarly large number of components while requiring only three primary tables. These three linker tables, genome_data, reaction_data, and metabolite_data act as interface points for a number of datasets on one end and components on the other end. In a traditional schema, the addition of a new data type would require generating a new linking table for each component type to be linked. With the current breadth available in the OME framework (n=9 components, m=8 dataset types) this would require 72 linking tables, which are instead represented with only 3.

4.4 Results and Discussion

The three main sets of features in the OME framework (Figure 5) lie along the interactions of each of the models, components, and dataset classes. These features can be utilized by a variety of different types of users (Figure 6).

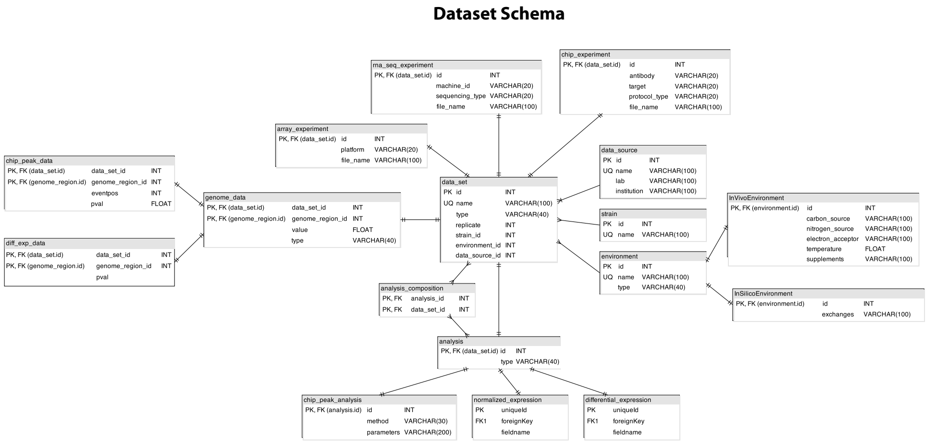
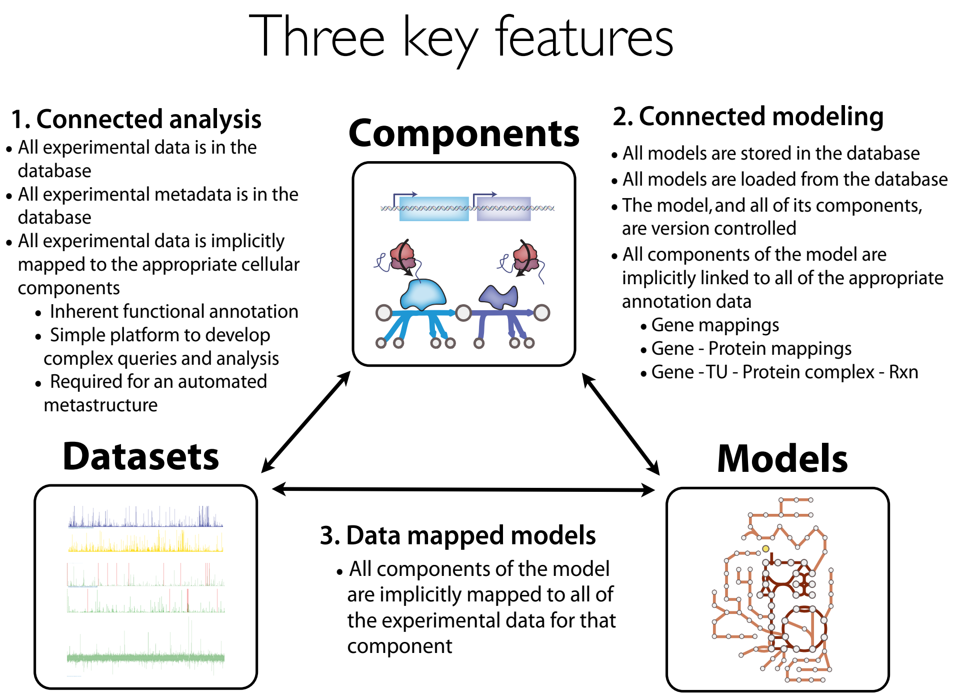


Figure 4.4: **Relational database schema for the models module.** The models module has a simple schema which provides two crucial benefits. The first is the simple architecture of the reaction_matrix table which enables the storage of arbitrary reaction types and stoichiometrys. The second is the division introduced between components and reactions which belong to a model vs. those which are universal. The reaction table and compartmentalized_component table are Universal in the sense that they will form a non-redundant set of reactions and components which are then used in the model_reactions and model_compartmentalized_component tables to specify the exact content for a given model.



Three key features

Components

1. Connected analysis

- All experimental data is in the database
- All experimental metadata is in the database
- All experimental data is implicitly mapped to the appropriate cellular components
    - Inherent functional annotation
    - Simple platform to develop complex queries and analysis
    - Required for an automated metastructure

2. Connected modeling

- All models are stored in the database
- All models are loaded from the database
- The model, and all of its components, are version controlled
- All components of the model are implicitly linked to all of the appropriate annotation data
    - Gene mappings
    - Gene - Protein mappings
    - Gene -TU - Protein complex - Rxn

Datasets

Models

3. Data mapped models

- All components of the model are implicitly mapped to all of the experimental data for that component

Figure 4.5: **The three primary features of the OME Framework arise from the interactions of its parts.** Connected analysis helps enables efficient and rapid analysis of high throughput data by linking to relevant genomic and metabolic annotation information. Connected modeling enables the use of the same detailed genomic and metabolic annotation information alongside a metabolic model. Data-mapped modeling used components as a bridge to provide implicit mappings between and gene, reaction, or metabolite in a model with corresponding experimental or computational data.

4.4.1 Connected modeling

Connected modeling integrates genomic and biochemical annotation data with constraint-based biological models to enable a vast array of tasks including model curation, quality control, and systems analyses. Model curation workflows on top of the OME framework include automated verification of reaction and metabolite IDs, along with a universal reactions and metabolite database. Quality control of constraint-based models is enforced genomically through strict gene-protein-reaction (GPR) mappings to genbank sequence files along with mass and charge balance for individual reactions and models via cobrapy[82] integration. Systems level analyses are enabled by the connected modeling features, such as integration of protein complexes, transcription unit architecture, and the simple but powerful integration of diverse annotation sources and ID mappings for genes, reactions, and metabolites. Current import functions support loading of various MetaCyc[94] and KEGG[92] datatypes if desired.

4.4.2 Connected analysis

Connected analysis integrates genomic and biochemical annotation data with a rich variety of high and low throughput biological assays, providing the ability to carry out highly complex analyses in a reproducible and efficient manner. The primary feature of connected analysis is to provide ID mappings and standardization among experiments, experiment metadata, and the annotation information of the biological parts relevant to an experiment. The mechanism for dealing with a large array of components and datasets without a code explosion lies in providing three primary links between biological parts and experiments or analyses related to these parts; genome_data, reaction_data, and component_data. The wealth of available component information combined with the huge quantities of experimental data make it possible to automate transcription unit and regulatory network annotation. The availability of well structured metadata alongside component data also enables a wide array of machine learning and other data mining based approaches in a reproducible and efficient manner.

4.4.3 Data-mapped models

Data mapped models provide a crucial practical link between large scale biochemical models and high-throughput biological data. The current primary features comes through the automated mapping of GPRs to the genome_data table which enables instant profiles of gene expression, differential gene expression, and gene regulation data for a given reaction in a metabolic model. These can be accessed as class attributes for a given reaction enabling researchers to rapidly probe and compare computed Flux Balance Analysis (FBA) flux states through cobrapy[82] to experimentally measured data. The datasets component of the framework is also designed to handle both real experimental data, and simulated computational data. Most importantly, the design provides for experimentally-measured environmental conditions to be implicitly mapped to the mathematical constructs used to represent that environment in-silico. This enables the storage of many potentially computationally intensive simulations, relevant experimental data, and seamless analysis and comparisons of the experimental data and the simulations. The design of the framework also provides handling of reaction and metabolite data to enable more direct comparison of stored fluxes and even concentrations of kinetic models.

4.4.4 Polyomic data integration

The OME framework provides a number of highly common operations as database views and convenience functions, including gene expression analysis and multi-omic integration. These views are demonstrated in the OME_views examples. Please also refer to the data_loading examples for details on loading the data that comprises these views. Note that the currently available metadata in these views and the database structure is geared towards microbial cell culture. Future efforts will replace the current simplistic metadata handling with the ISA (isatools.org)[95] standard. The first crucial database view is GeneExpressionData which provides rapid and simple access to arbitrary volumes of gene expression data. The GeneExpressionData view is currently defined by nine columns that include the gene locus id, the gene common name, the type of gene expression data (currently RNA-Seq or Affymetrix arrays), the average value across all replicates, the standard deviation of this value across replicates, the strain used in the experiments, the carbon source used in the experiments, the nitrogen source used in the experiments, the electron acceptor used in the experiments, and finally any supplements used in the experiments. The second common view is DifferentialGeneExpressionData which builds off of the GeneExpressionData to present all possible pairs of pre-computed differential expression statistics across environmental and genetic perturbations. The columns of DifferentialGeneExpressionData mirror those for GeneExpressionData but also include a Bonferroni corrected p-value and the log fold change calculated between the expression profiles. This difference calculation is shown with a / on the given column that is used for the differential (e.g. glucose/fructose) in the carbon_source column, corresponding to differential expression calculated between two samples in which the first was grown with glucose as a carbon source and the second is grown with fructose as a carbon source. The log2 fold change is then calculated as sample1-sample2, in this case glucose-fructose. The third, fourth, and fifth common views represent the integration of ChIP-chip, ChIP-seq, or ChIP-exonuclease data with genome annotation and gene expression data. ChIPPeakGene first integrates ChIP binding regions with the transcription unit architecture or gene annotation to provide a mapping between a ChIP peak and the genes it will regulate. This view is then extended and combined with the previously discussed GeneExpressionData to provide another view, ChIPPeakGeneExpression, in which quantitative information for any ChIP peak is paired with the expression data for downstream genes. With this view a user can make instant correlation plots to compare ChIP peak intensity with gene resultant expression. Finally the same process as above is repeated with the only difference being the integration of DifferentialGeneExpressionData with ChIPPeakGene instead of GeneExpressionData. This highly useful view, ChIPPeakDifferentialGeneExpression view, comprises a data-driven reconstruction of a regulatory network. This is accomplished via the automatic integration of a ChIP peak with the differential gene expression that is caused by the binding of that peak. This occurs because the only differential expression values shown correspond to the difference between a wild type strain and a gene deletion strain for the transcription factor that was responsible for the ChIP peak. This view is tailored for paired ChIP* and transcription factor knock-out gene expression data. Example data are included to enable this view and the data corresponding to Figure 4 of [87] and Fig 3 of[86].

4.4.5 Arbitrary creation and traversal of biological complexes and modules

The OME framework can deal with arbitrary complexes that currently include, but are not limited to, protein-protein complexes, DNA-protein complexes, protein-small molecule complexes, and protein-RNA complexes. This is accomplished with the simple architecture of the component_composition table in Fig 3. The architecture of this table enables the creation of components of components. Since a component can be any number of physical biological entities, this results in the creation of arbitrary biological complexes or modules. Practically, performing recursive SQL queries directly on a database can be difficult. However, the current implementation is able to harness the power of SQLAlchemy to generate the effect of a graph-like object oriented tree structure. The complexes_example notebook (http://nbviewer.ipython.org/github/sbrg/ome/blob/master/examples/complexes_example.ipynb) illustrates these features. Essentially, every component is given a parent and a child attribute which corresponds to the direct parent or children components. A further recursive SQLAlchemy function is then written as a function, all_children, which can traverse the graph and return all children of a parent. In this manner a full tree can be iteratively traversed or further recursive functions can be written to speed the traversal.

4.4.6 Optional NoSQL data storage for extreme scalability

The OME framework has a distributed architecture that enables it to store and analyze enormous quantities of data. It does this by simply leveraging NoSQL type in-memory scalable databases alongside the rich relational ORM. The primary NoSQL table is a simple genome_data table that stores genomic position, an arbitrary numeric value, and a dataset ID. The dataset ID can thus provide all of the necessary information about the given piece of data through queries onto the datasets module. Users can thus rapidly query out sets of experiments based upon well structured metadata and then use the resulting dataset ID to pull out any genome_data values corresponding to a genome region of interest. This general pattern of Polyglot Persistence is becoming more and more commonplace as an efficient solution to solve difficult heterologous data problems [75](martinfowler.com/bliki/PolyglotPersistence.html).

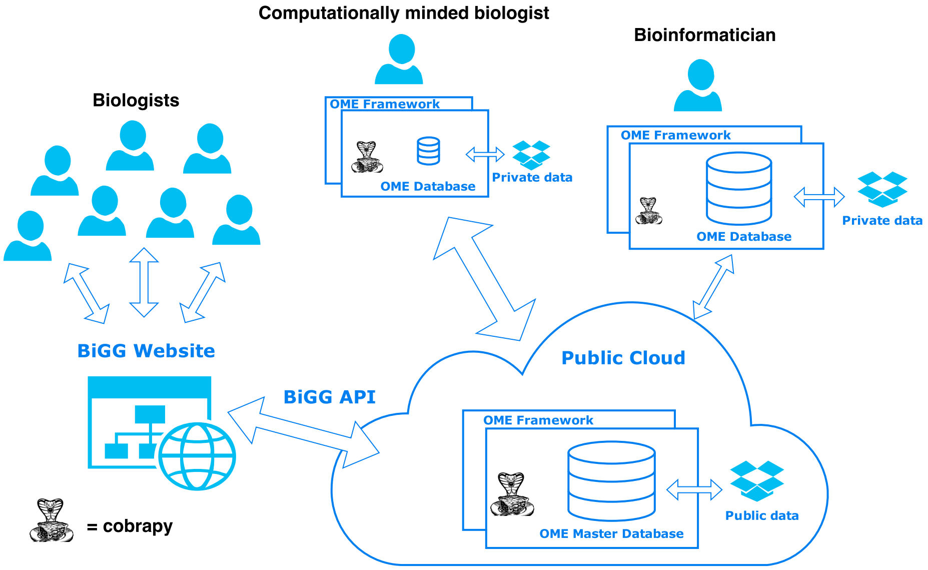


Figure 4.6: **The OME Framework is designed to serve three overall use cases.** (A) The first three use cases rely to varying degrees on publicly provided data from UCSD. The non-computational biologists can access much of the underlying data and features through the next-gen BiGG website and corresponding BiGG APIs. The computationally minded biologist can easily install a local copy of the OME Framework but rely heavily on the public preformatted data files provided from the main BiGG hosted instance. Finally, the bioinformatician can install a local copy along with large quantities of private data, annotation information, or models.

4.4.7 Use case I: Loading a metabolic model, genomic annotation, and various genomic datatypes

The most basic use case is simply loading data into and out of a local version of the OME Framework. This is demonstrated in the accompanying notebook (http://nbviewer.ipython.org/github/sbrg/ome/blob/master/examples/data_loading.ipynb) which can be found in the examples directory. Once the framework is installed (http://github.com/SBRG/ome/INSTALL.md) navigate to the ome_data directory in your home folder. Note that you can easily change the location or name of the ome_data folder by simply altering the settings.ini file as explained in the installation documentation. You will notice example data for S. cerevisiae and E. coli including representative metabolic models in the models directory [96, 97] corresponding genbank genome annotation files in annotation/genbank, and a few RNA-Seq datasets in rnaseq/fastq and rnaseq/bam. It is recommended to run through the loading of this test data before adding your own data or including additional organisms. However, if you are feeling brave then go ahead and start piling in your own datatypes. Note that even though folders exist for metabolomics, fluxomics, motifs, and proteomics, each directory is currently empty in the example installation. Currently supported datatypes are shown in the documentation http://ome.readthedocs.org/en/latest/ome.html for datasets. Navigate back to the data_loading ipython notebook and begin following the instructions, based upon how and what kind of a data you are loading. In general there are two ways to load data into and out of the database. The first is to rerun the entire loading script called upon installation, bin/load_db.py. This will reliably scan each directory in the ome_data folder and load in any datasets, models, and annotation information that is found there. This is the easiest and most simple way to load data into the OME database. This is the recommend approach for new users. However, this can be very slow for experienced users who already have a large amount of data loaded in. This occurs because the loading functions prevent duplication and thus are forced to check the existing database before attempting to insert any new data. This checking of existing values to avoid duplications is generally unavoidable given inherit redundancies and imperfections in biological annotation and data sources.

4.4.8 Use case II: Data mining and machine learning across diverse datatypes

The OME Framework can be easily used as a scaffold for machine learning and data mining applications of diverse biological data sources. The advantages lie in a native python implementation that results in grokking of the popular pandas library and resultantly numpy, scipy, matplotlib, scikit-learn and the rest of the rich and ever expanding python ecosystem. This is utilized in the accompanying data heatmaps notebook in which the classic clustering of gene expression data use case is illustrated. In the example, groups of genes are queried out based on the groups table, which can store arbitrary collections of genome region objects. The automatically loaded gene groups tables correspond to model subsystems in the example models. However, one can also readily load KEGG pathways or other resources if available. The clustering is then done with scipy and visualized interactively with D3 code based off of a common example (http://bl.ocks.org/ianyfchang/8119685).

4.4.9 Use case III: Integration with Escher for data visualization on metabolic maps

Escher is a web-based visualization tool for building, sharing, and embedding metabolic maps [28]. Escher comes preloaded with numerous metabolic and also enables users to build their own metabolic maps for any organism. Escher also supports a rich variety of options for visualizing datasets on top of metabolic maps, with support for gene-, reaction-, and metabolite-associated data. The OME framework integrates with Escher by associating any loaded dataset with the identifiers in any Escher map for a given organism, so that visualizations can be generated immediately in a Python session (via a web browser) or in the IPython notebook. As demonstrated here and in the previous example, the OME Framework simplifies the generation of an array of web-based visualization tools, and integration with the IPython notebook means that code and visualizations can be executed in the same, simple interface.

4.4.10 Use as a scaffold for whole cell models

The OME framework can be used as a scaffold to construct whole cell biological models. The original database design for the OME Framework happened in two highly related but distinct software projects, which have both played a critical part in multiple significant publications[12, 85, 86]. In particular, the OME Framework grew out of the distinct effort to combine the original ME database which continues to power the ME models, and the TRN database, which was designed to handle the ever increasing breadth of genomic data types and resultant analyses. In this way the underlying schema and design has (as indicated through its name) always held the primary intention of existing as the backbone for the development of whole cell biological models.

4.5 Conclusion

The OME Framework is a systems biology and bioinformatics framework that helps to enable a broad array of iterative and evolving systems biology workflows. Its key features arise from the rich interaction among its constituent parts: models, components, and datasets. However, each individual module can be used in isolation without the necessity to dive into the full feature set. For example, models can be used alongside COBRApy as a model storage utility. Similarly, the datasets and components modules can also be used individually to simply store and retrieve datasets or biological parts.

The structure of the OME Framework facilitates the transformation of data to information to knowledge. After collection, data must be processed and integrated with other data or broken down further until it becomes a piece of information. Information is then assembled over time into larger constructs that represent fundamental knowledge about a system. The OME Framework embodies this process at the genome-scale. As we move towards ever more accurate and comprehensive digital representations of organic systems, a key factor in the rate of this progress will be the ease and efficiency in which the feedback loop between data generation and model consistency is closed.

Finally, most biological analyses and analysis tools are focused on working with a relatively limited subset of data concerned with a specific study or question of interest. However, as the sequencing and synthesis of DNA continues to surge ahead it becomes clear that we will soon generate 10 replicates instead of 3, cross reference 10 historic datasets instead of 1, and most importantly begin to get a level of quantitation that will enable true biological engineering. The OME Framework is a small step towards a future that amasses, manages, and processes biological data at a systems scale.

4.6 Acknowledgements

Chapter 4 in full is a reprint of a submitted manuscript: Federowicz, S.A.*, King, Z.A.*, O'Brien, E.J.*, Ebrahim, A. *, Lerman, J.A.*, Lu, J. *, Sonnenschein, N., Latif, H., Lewis, N.E., Palsson, B.. The OME Framework for Genome-Scale Systems Biology. Submitted. * Indicates equal contribution. The dissertation author was the primary

author of this paper responsible for the research. The other authors were Zak King (equal contributor), Edward O'Brien (equal contributor), Ali Ebrahim (equal contributor), Joshua Lerman (equal contributor), Justin Lu (equal contributor), Niko Sonnenschein, Haythem Latif, Nathan Lewis and, Bernhard Ø. Palsson.

Chapter 5

Integrated Analysis of Molecular and Systems Level Function of Crp Using ChIP-exo

5.1 Abstract

A fundamental challenge of systems biology is to unify detailed molecular mechanisms with genome-scale predictive models. Meeting this challenge for transcriptional regulatory networks has proven difficult. Here, ChIP-exo, a derivative of ChIP-seq, is applied to help bridge this gap for transcription initiation in bacterial systems. σ70 ChIP-exo profiles revealed patterns of DNA protection at genome-scale under in vivo conditions that strongly corroborate in vitro DNA footprinting studies. Aligning and orienting the ChIP-exo data relative to the TSS identified a unimodal distribution centered at +20 on the template strand and multimodal distributions located between -35 and +5. Both strands indicate the capture of stable intermediates of transcription initiation that occur post-closed complex formation. Similar profiles are observed for ChIP-exo

performed on activators but not for repressors. Activators are found to provide little protection at the operator site whereas repressors are centered on them. Furthermore, genetic perturbations of RNAP/Crp interactions highlight the importance of RNAP/Crp interactions for stabilization of the ternary complex. This suggests that post-recruitment, Crp is associated with RNAP but independent of the operator site. Finally we are able to confirm recent physiological models of Crp regulation at the genome-scale and provide an initial link between promoter level mechanisms and systems level regulation.

5.2 Introduction

A longstanding goal of molecular biology is to link structure with function. This structure-function relationship plays out at many levels with one of the most complex being the link between the structural attributes of proteins and the overall functions of pathways or systems level features of a cell. This link proves especially difficult because it comprises many smaller, yet equally important and difficult links. One crucial smaller link is the relation between a proteins structure and its individual function [98] and another is the link between the structure of a biomolecular network and its corresponding function [99]. Thus, attempting to provide links and cohesion between the structure of individual components of a cell and the complexes of those components with the overall function of the cellular system as a whole will continue to be a grand challenge for many years to come. Here, we seek to provide an initial link in that direction by synthesizing protein structural level detail of transcriptional initiation complexes with systems level features and physiological models of cellular behavior.

One crucial technique in biological sciences is chromatin immunoprecipitation (ChIP) has been paired with microarrays (ChIP-chip) [100] and nextgen DNA sequencing (ChIP-seq) [101] to comprehensively unravel the molecular networks of protein binding to DNA. This approach has proved to be highly successful and has resulted in an explosion of knowledge about transcriptional regulation. However, both ChIP-chip and ChIP-seq are limited to a resolution of the size of DNA fragments that are identified to about 50-100 nucleotides. ChIP-exonuclease [102] was developed in order to significantly increase this resolution, reaching the level of a dozen or so nucleotides. By utilizing a 5 exonuclease to chew down the ends of DNA not protected by a cross-linked transcription factor, it is possible to achieve single nucleotide resolution binding profiles. While ChIP-exo has been extensively deployed in mammalian systems [102], it has yet to be applied in detail to transcriptional regulation in bacteria. Since much of the foundation of our knowledge of transcriptional regulation comes from bacteria, and E. coli in particular, we sought to assess whether or not ChIP-exonuclease would enable new or complementary insights that could be based off the rich body of detailed structural and biochemical work in bacteria.

One transcription factor in particular, Crp, has long been studied as the canonical transcriptional activator in bacteria. Many detailed studies have shown precisely how Crp carries out regulation at three separate classes of promoters [103]. Similarly, recent work has also shown the precise physiological mechanisms by which Crp regulates the entirety of metabolism and the cell [6]. We thus sought to provide a link between this detailed structural knowledge and detailed systems knowledge by fully elucidating the Crp regulatory network in precise chemical detail. Previous studies have been carried out with transcriptomics [104, 105] ChIP-chip [45, 106], and genomic SELEX [107]. However, none of these studies were able to assay a full range of physiologically relevant conditions or pair in vivo ChIP measurements with expression profiling to generate a full set of regulatory events. Similarly a vast array of studies has detailed the individual architecture and structure-function relationship at a number of classically studies promoters in E. coli [103, 108, 109, 110, 111, 112]. However, these studies did not enjoy the benefit of massively parallel assessment of detailed structural binding information from ChIP-exonuclease.

In this study, we carry out ChIP-exonuclease for both σ70 and Crp for wild type cells under glucose, fructose, and glycerol conditions. In addition we study the effects of canonical Crp mutation in two crucial regions, Ar1 and Ar2, on ChIP-exo binding profiles under glycerol conditions. We also carry out paired RNA sequencing across wild type, Δcrp, and strains harboring Ar1 and Ar2 mutations to elucidate the full Crp regulon.

5.3 Results

5.3.1 ChIP-exo data provides in vivo mechanistic insights into bacterial activation

Applying the ChIP-exo technique to interrogate transcriptional mechanisms illustrates the utility of this approach in providing genome-scale insights into promoter activity in vivo. By trapping actively transcribing complexes, stable intermediates for a given promoter are revealed on the path to transcriptional elongation. ChIP-exo results for σ70 and for Crp, the classic model for transcription factor mediated activation, are discussed.

5.3.2 ChIP-exo on sigma70 serves as a method for locating the TSS

The predecessors of ChIP-exo, ChIP-chip and ChIP-seq, are enriched at bacterial promoters when the σ factor is targeted [113, 37, 114, 115, 116, 2, 117]. Ecocyc annotated TSSs where checked for σ70 ChIP-exo enrichment within 200 bp for data generated on three different substratesglucose, fructose, and glycerol (Fig. 1A and Fig. S1). Like its predecessors, ChIP-exo peaks are consistently found near promoters but

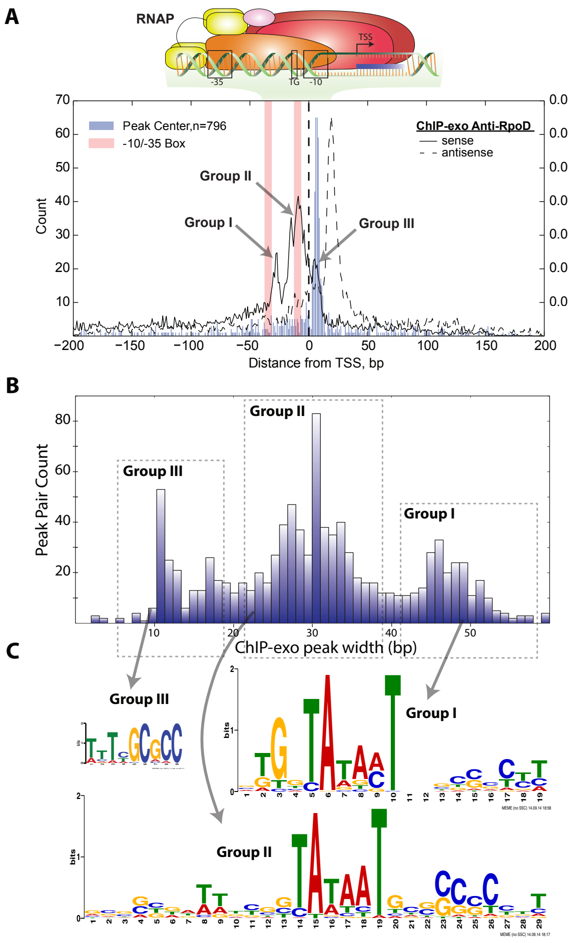


Figure 5.1: **TSS aligned and oriented σ70 ChIP-exo data reveals DNA footprint patterns consistent with stable transcription initiation intermediates.** (A) ChIP-exo peak regions were aligned and oriented relative to the TSS. The peak center (blue bars) is shown to be to consistently downstream of the TSS with a median of 5 bp. The mean distribution of the 5 position of reads (5 tags) is shown for both the template and nontemplate. The template strand distribution shows a unimodal profile that spans +207 bp and is consistent with in vitro footprinting studies characterizing the RPO, the ITC, and the TEC stable intermediates. The nontemplate strand shows a multimodal distribution with modes centered approximately +5 relative to the TSS (Group III), upstream and over the -10 promoter element (Group II), and slightly downstream of the -35 promoter element (Group I). (B) Examination of the distance between template and nontemplate strand peak maximum shows that the footprint lengths are >40 bp, 21 to 40, <20 and for Group I, Group II, and Group III respectively. (C) A motif search was performed for the -10 and -35 promoter elements for Group I, Group II, and Group III promoters. All three show σ70-like promoter sequences but with slight differences. Group I has a -35 motif that most closely resembles the consensus (TTGACA), has a highly conserved -11A, and a partial TGn motif. Group III has the least conserved -35 promoter element and no extended -10 promoter element.

only the single-nucleotide resolution of ChIP-exo provides the exact genomic location bound by RNAP holoenzyme. Surprisingly, ChIP-exo data generated on σ70 is found to be a better proxy for the TSS than it is for the -35 and -10 promoter recognition sequence elements. The median position of the σ70 peak center is 5 bp downstream of the TSS for all three carbon sources (Fig. 1A and Fig. S1). The spatial consistency of the σ70 peak center demonstrates the utility of ChIP-exo to approximate TSSs to within base pairs from where they exist and provides an orthogonal method to complement 5RACE-based TSS detection.

5.3.3 Strand oriented σ70 ChIP-exo peak distributions reveal stable intermediates in transcription initiation

The σ70 ChIP-exo peak distribution provides the bounds of protected DNA regions on the template and nontemplate strand. By aligning to the TSS as a reference position, we are able to determine the strand orientation for the σ70 ChIP-exo peaks. ChIP-exo profiles across all binding sites were calculated for the template and nontemplate strand by first calculating the density of the 5 end of tags for each individual peak region spanning 400 bp centered on the TSS. The strand orientated ChIP-exo profiles for σ70 reveal significant distinctions between the template strand and the non-template strand (Fig. 1A). The binding profiles show a unimodal distribution on the template strand whereas a multimodal distribution is seen on the non-template strand. The width of the peak regions was determined by calculating the distance between position of maximum 5 tag count on the template strand relative to position of maximum 5 tag count on the nontemplate strand (Fig. 1B). This corroborates the distribution profiles observed on the two strands and indicates that most promoters have a σ70 ChIP-exo profile that fall into one of these distances.

The activity of lambda exonuclease is 5 to 3 [118] and, as such, the protected region on the template strand is found downstream of the TSS. The unimodal ChIP-exo distribution on the template strand has a maximum 5 tag density +20 bp downstream of the TSS and approximately 30% of the mean 5 tag density is found between 207 bp. The position of the unimodal distribution on the template strand is in strong agreement with numerous in vitro footprinting studies characterizing the open complex (RPO) and the intermediates preceding RPO formation, the initial transcribing complex (ITC) and the transition to the ternary elongation complex (TEC). However, these results are not reflective of footprinting studies capturing the closed promoter complex (RPC), which typically protect promoter DNA from -5 to +5 [119]. Hydroxyl radical footprinting studies on RPc formation in the T7A1 promoter showed that the short-lived RPc complex protects DNA to approximately -5 bp [120, 121, 122]. Similar results were observed in DNase footprinting of the T7A3 promoter, lacUV5, and rrnBP1 [123]. However, the RPC complex was only observed when the temperature was dropped. The temperature dependent capture of early closed complexes has been shown to be a result greater RPO abundance at physiological temperatures [124, 121, 122]. Lastly, an RNAP mutant with deficient open complex formation was found to have DNase I footprints that extend to just +1 at the λPR promoter[125].

In agreement with the ChIP-exo data presented for the σ70 template strand, in vivo footprinting studies performed on the isomerization intermediates converting the RPc complex to RPO, the RPO complex, ITC, and early TEC uniformly conclude that the downstream boundary of protected DNA extends +15-25 relative to the TSS. Downstream boundaries centered on +20 have been observed in studies performed on the intermediates leading to RPO and the PRO complex for the T7A1 promoter [120, 122], the T7A3 promoter[123], the rrnBP1 promoter [126, 127, 128], the λPR promoter[124], and the lacUV5 promoter [123, 129]. Furthermore, the ITC and the transition to the TEC also have a downstream footprint boundary of +20-25. DNase footprinting of T7A1, tac, and lacUV5 promoters showed that the ITC has a slightly advanced footprint at +25 compared with +20 for RPO and the early TEC had a footprint at +30 [130, 131, 132].

Unlike the template strand, the ChIP-exo 5 tag distribution for the nontemplate strand is multimodal. This distribution marks the upstream boundary relative to the TSS. The dominant mode accounts for 28% of the 5 tag density and is found between -18 and -1. Therefore, promoters that belong to this mode have partial to complete protection of the discriminator sequence, the -10 promoter element, and the TGn extended -10 element but offer no protection to the -35 promoter element or any upstream promoter elements (e.g. UP element). The -35 promoter element is partially protected by the mode farthest upstream which accounts for 9% of the 5 tag density profile and spans -34 to -23 with a maximum located at -28. The upstream boundary, -3, is located in the center of the -35 element. The downstream mode accounts for 8% of the 5 tag density and is located downstream of the TSS. The boundaries of this mode are between +4 and +12 with a local maximum at +6.

The multimodal distribution on the nontemplate strand also reflects in vitro footprinting studies which find greater variability in the upstream protected region [119]. Like the template strand, the DNA protected regions of the different modes on the nontemplate strand provide little to no support that recruitment and RPC complex formation is being captured by these ChIP studies. The upstream boundary from footprinting studies conducted on RPC show periodic protected regions extending upstream of the -35 promoter element typically -55 to -12 but is highly dependent on the involvement of transcription factors and the alpha subunit [124, 120, 122]. As noted previously, the short-lived nature of RPC at physiological temperatures likely prevents its capture during ChIP studies. Furthermore, there is evidence to support that these modes are reflective of stable intermediates that occur post-recruitment and, in particular, the ITC and transition to the TEC. A detailed study on the lacUV5 promoter using DNase I, methylation protection, and exonuclease III protection across transcription intermediates showed that transitions from RPO to ITC undergoing abortive initiation retained strong protection of a region between -24 and -6 to exonuclease III digestion that was reduced to protection of the region downstream of -6 after escaping the abortive transcription phase to produce longer transcripts[133]. This is further corroborated by a recent study that showed the lacUV5 promoter has an upstream footprint boundary at -23 in the presence of σ70 compared with -13/-14 for the σ70 lacking transcribing complex[134]. A study of the T7A1 promoter using exonuclease III showed a drastic movement in the upstream-protected region from -43 to -3 in the transition from RPO to early transcribing complexes (ITC or TEC)[135]. The differentiation between the ITC and TEC can further be seen in the total length of the protected region (Figure 1B). Early TECs have been found to have footprint regions spanning 30 bp whereas the ITC has a longer footprint seen to be 50+ bp in length [136, 130, 133].

5.3.4 Promoter motif analysis of the σ70 ChIP-exo distributions

It is known that promoter sequence elements involved with RNAP holoenzyme recruitment contribute to the post-recruitment kinetics of transcription initiation[119]. Thus we examined the -10 and -35 promoter elements for the different σ70 groups (Fig. 1C) as determined by the difference in peak-pairs (Fig. 1B). σ70-like promoter motifs were found in all three groups. Though a detailed analysis of additional DNA sequence elements (e.g. UP elements, transcription factor binding sites, nucleoid-associated protein binding sits) could be more revealing, examination of just these two promoter elements revealed subtle difference between the promoter elements of each groups. Group I, having the longest distance between peak-pairs, has a motif that most resembles the -35 consensus sequence (TTGACA). Furthermore, the -10 promoter element has near perfect consensus at the critical -11-A position and a partial TGn motif characteristic of the extended -10 promoter element. Conversely, Group III has the most divergent -35 motif from consensus and no appreciable motif for the extended -10 promoter element.

5.3.5 ChIP-exo profiling of a canonical transcriptional activator: Crp

Transcription activation in bacteria was further studied by conducting ChIP-exo studies on Crp, the most studied and best characterized transcription factor[137]. These profiles revealed differences in the DNA protection patterns observed among the different classes of Crp activators. Three representative examples of ChIP-exo profiles generated on glycerol are shown for each of the three Crp Classes (Fig. 2A). The deoC promoter is a Class III promoter with two Crp binding sites flanking a CytR regulatory site that represses the activating action of Crp[138]. The ChIP-exo protected regions are in close proximity with the three operator sequences with protected regions near -40 and -90 as previously seen in vitro[138]. However, markedly different profiles are observed in the Class I (tnaC) and Class II (gatY) promoters that often have no exonuclease protection to the Crp binding site, but instead, have strong protection of the region surrounding the TSS. In fact, these regions correspond greatly with the ChIP-exo profiles generated for σ70 under the same condition but no observed σ70 ChIP-exo peak was detected for the repressed deoC promoter.

The results for these individual promoters are consistent with those observed at the genome-scale. Examination of the mean 5 tag distribution of Crp ChIP-exo data oriented relative to the TSS illustrates that Class I and Class II activators have a peak center that aligns greatly with the TSS and not the Crp binding site (Fig. 2B). To confirm that ChIP-exo was enriching Crp regulated promoters, the predicted Crp binding sites were oriented and aligned relative to the TSS (Fig. 2B). This shows three regions of elevated Crp operator sequence centered at -41.5, -61.5, and -93.5 bp upstream of the TSS corresponding with the expected positions of Class II, Class I and Class III promoters

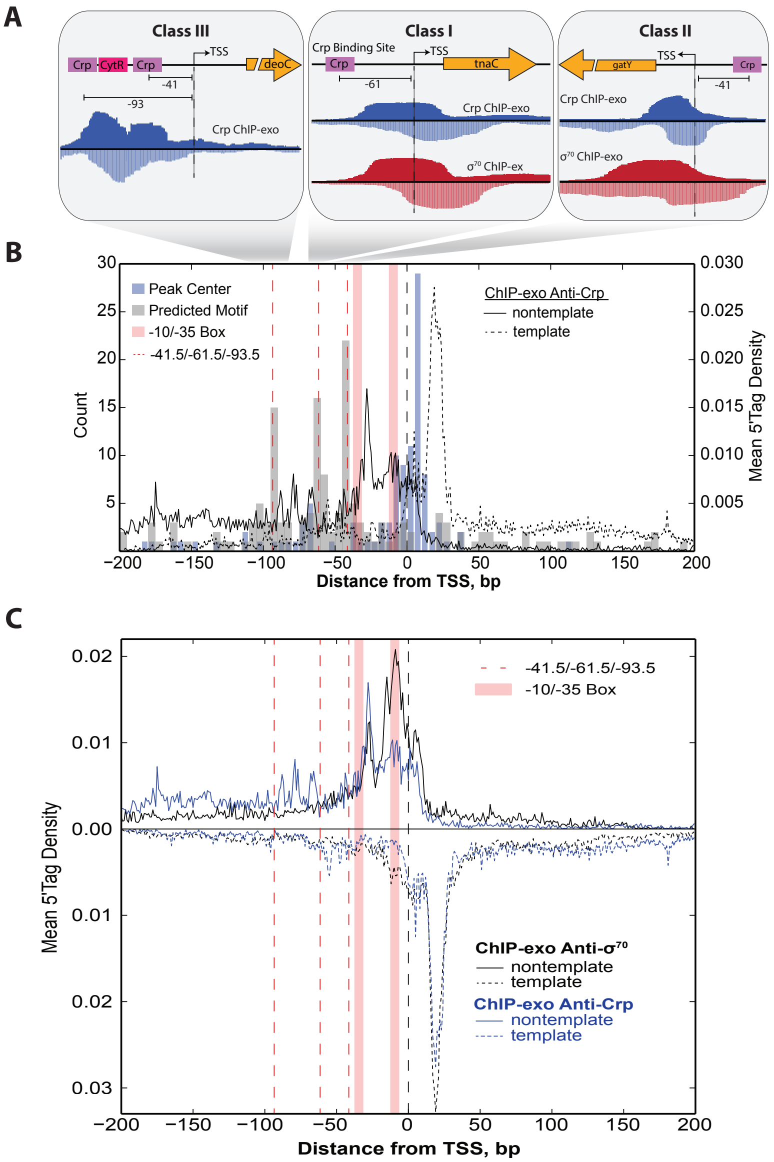


Figure 5.2: **Crp promoter classes have unanticipated ChIP-exo footprint regions.** (A) Gene tracks of Crp ChIP-exo data is shown for different classes of Crp promoters. The deoC promoter (Class III) is regulated by Crp and CytR. Peak regions are found at both Crp operators and for CytR which binds between them. However, it is often observed that for Class I and Class II promoters there is few observed reads over the Crp operator. Instead, the ChIP peak is centered on the TSS and the footprint region co-occurs with that found for σ70. Examples of this are shown for tnaC (Class I) and adhE (Class II). (B) Genome-scale analysis of all Crp peak regions relative to the TSS indicates that the peak center for a majority of Crp peak regions aligns better with the TSS rather than the Crp operator.

respectively[137]. Similar ChIP-exo profiles were obtained when wild type E. coli was grown on fructose, a Crp activating condition, but when grown on glucose, a repressing condition for Crp, few binding sites were detected (Fig. S2). We further verified that these results where not artifacts attributed to the anti-Crp antibody used to perform ChIP-exo by generating data on a Δcrp strain and no correlation was observed between biological replicate datasets indicating minimal impact due to non-specific binding (Fig. S3).

The ChIP-exo 5 tag density profile for Crp was also compared with that generated for σ70 (Fig 2C). The ChIP-exo 5 tag density for all identified Crp peak regions for cultures grown on glycerol were processed and oriented as described previously for σ70. These density profiles were then compared to the σ70 density profiles determined across the same set of peak regions. Strand orientated Crp density profiles reveal a unimodal distribution on the template strand and a multimodal distribution on the nontemplate strand analogous to those found for σ70. The template strand strongly overlaps the one observed for σ70 with a downstream boundary of protected DNA centered on +20 accounting for 33% of the aggregate density profile. However, the Crp nontemplate density profile has distinctive features. First, there is increased DNA protection on the nontemplate strand between the -93.5 and -61.5 markers. This region encompasses 13% of the total 5 tag density profile. These positions signify the center position of many Class III and Class I Crp operator sequences respectively [137]. Furthermore, Group I, Group II, and Group III modes have relative ratios of 2.5:3.3:1 compared with 1.1:3.5:1 for the same regions in the σ70 ChIP-exo 5 tag density profile. The Crp promoter bound regions appear to have a larger relative fraction of DNA protected as far upstream as the center of the -35 box. However, none of these regions indicates protection of the Crp operator sequences found for Class I and Class III promoters and only partial protection for Class II promoters due to the overlap with the -35 box.

Rifampicin (rif) prevents transcription elongation beyond 2-3 nt in length[139] and, in doing so, leaves the transcription machinery unable to advance beyond the ITC. Therefore, ChIP-exo was performed on cultures treated with rif prior to harvest followed by immunoprecipitation of Crp. The resulting mean 5 tag density profile generated on both the template and nontemplate strand closely resembles that obtained in the non-rif treated sample (Fig. S4). Therefore, this chemical perturbation of the transcriptional state had no impact on the Crp ChIP-exo distribution and no additional upstream protection of the Crp binding site was observed. This indicates that the exonuclease footprints are occurring on initiation complexes occurring prior to the TEC. This observation coupled with the evidence against the short-lived RPC complex strongly suggests that the Crp promoters studied here are being captured at stable intermediates associated with RPO formation or the ITC.

5.3.6 Distinct ChIP-exo profiles for transcriptional activators and repressors

Previously, applications of ChIP-exo have demonstrated binding events centered on the binding motifs in eukaryotic systems [102, 140]. We have previously applied the ChIP-exo protocol to characterizing the transcriptional repressor Fur in E. coli and observed similar binding profiles. To further examine the distinctions between ChIP-exo profiles of transcriptional activators and repressors, ChIP-exo was performed on c-Myc tagged strains of ArcA (repressor) and Fnr (activator) grown anaerobically on glucose minimal media. The data generated was then processed, aligned, and oriented relative to the nearest TSS (Fig. 3). ArcA which typically binds near the TSS [86] has no defined ChIP-exo 5 tag distribution on either strand though there is a noticeable increase in the 5 tag density around the TSS (Fig 3A). In contrast, Fnr demonstrates a similar 5 tag density profile as was seen for Crp and σ70 with a strong unimodal distribution on the template strand at +20 and a less defined modal distribution on the nontemplate strand (Fig. 3B).

The ArcA ChIP peak regions were aligned relative to the peak center position (Fig 3C). This yields a uniform distribution of 5 tag density with sharp peaks on the forward (+) strand and the reverse strand (-). Furthermore, plotting the predicted binding sites shows that the protected regions are centered on the ArcA motif. Lastly, the peak-pair differences for ChIP-exo profiles of ArcA and Fnr are shown (Fig. 3D). This reveals that the footprint obtained for the repressor is approximately 30 bp while the activator has a much broader footprint distribution extending to 80 bp.

5.3.7 Genetic perturbation of RNAP holoenzyme/Crp interactions

The activating properties of Crp and many transcription factors is through stabilizing interactions with RNAP holoenzyme at the promoter [137]. Molecular characterization studies and mutational analysis of Crp has revealed three activating regions (Ars) that make protein/protein interactions at specific positions in RNAP holoenzyme[141]. The first, Ar1, Crp interacts with either of the subunit at the C-terminus and the HL159 mutation to Crp prevents this interaction from forming[142][143]. This region is involved with activation at Class I and Class II promoters. The second region, Ar2, is only associated with Class II promoters and binds to the N-terminus of the subunit. This bond was shown to be severely disrupted by introduction of two mutations to Crp, KE101 and HY19. Lastly, a weaker interaction was found to occur at Ar3 between Crp and the σ factor.

We next sought to determine the impact of genetic perturbations to the RNAP holoenzyme/Crp interactions at Ar1 and Ar2. HL159, KE101+HY19, and HL159+KE101 mutation were introduced to create an Ar1, Ar2, and Ar1+Ar2 deficient mutant in Crp respectively (Fig 4A). From this point forward the mutants will be referred to as delAr1, delAr2, and delAr1delAr2. ChIP-exo was performed on these mutant strains with glycerol s the sole carbon source. In comparison with the wild type, each mutant resulted in the loss of peak regions (Fig 4B). The most drastic effect was observed in the delAr1delAr2 mutant which retained 40% of the peaks in the wild type strain. This result indicates the importance of these Ar regions on the stabilization of both Crp and RNAP holoenzyme at the promoter site. Furthermore, the characteristic ChIP-exo 5 tag density profiles (see Fig. 2C) on both strands were systematically degraded with each mutation resulting in profiles that no longer aligned well to the TSS (Fig. S5). To determine which peak regions were lost as a result of these genetic perturbations, the distribution of peak region centers was analyzed (Fig. 4C). The mutations predominantly result in a loss of peak-regions where the peak center was located near the TSS (-10 to +20 bp) and peak centers farther away from the TSS were less impacted. Lastly, the distribution of predicted binding sites were examined in the context of the different mutant strains (Fig. 4D). In agreement with expectation, modulation of Ar1 results in a drop in the predicted binding sites observed near -61.5, the typical Class I promoter distance from the TSS. This drop near -61.5 was partially recovered in the Ar2 mutant but a severe drop in the -41.5 centered binding sites occurred. This distance upstream of the TSS is associated with Class II promoters. The delAr1delAr2 mutant has a loss in peak regions with Crp binding sites matching those of Class I and Class II promoters. However, the peak regions of Class III found near the -93.5 position are unaffected by mutations in Ar1, Ar2, or both.

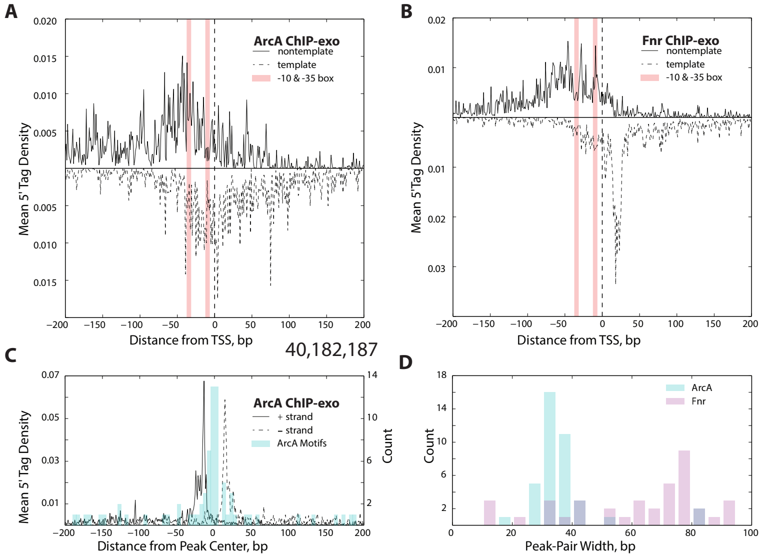


Figure 5.3: **Contrasting ChIP-exo profiles of repressors and activators.** (A) The TSS aligned ChIP-exo profile for ArcA, a predominantly repressive transcription factor, is shown to lack the characteristic distribution of mean 5 tag density observed on both the template and nontemplate strand. (B) The TSS aligned mean 5 tag density profile for Fnr, typically an activator, resembles the profile found for Crp and σ70.(C) The ArcA ChIP-exo profile is shown for all peak regions aligned to the peak center position. Also shown is a histogram of the center of the predicted ArcA binding site relative to the peak center position. This illustrates that the ChIP-exo profile is centered on the predicted binding site.(D) A comparison of the peak-pair distance is shown to illustrate the difference in resolution observed between ArcA and Fnr. ArcA, the repressor, is revealed to have shorter footprints compared with Fnr, the activator.

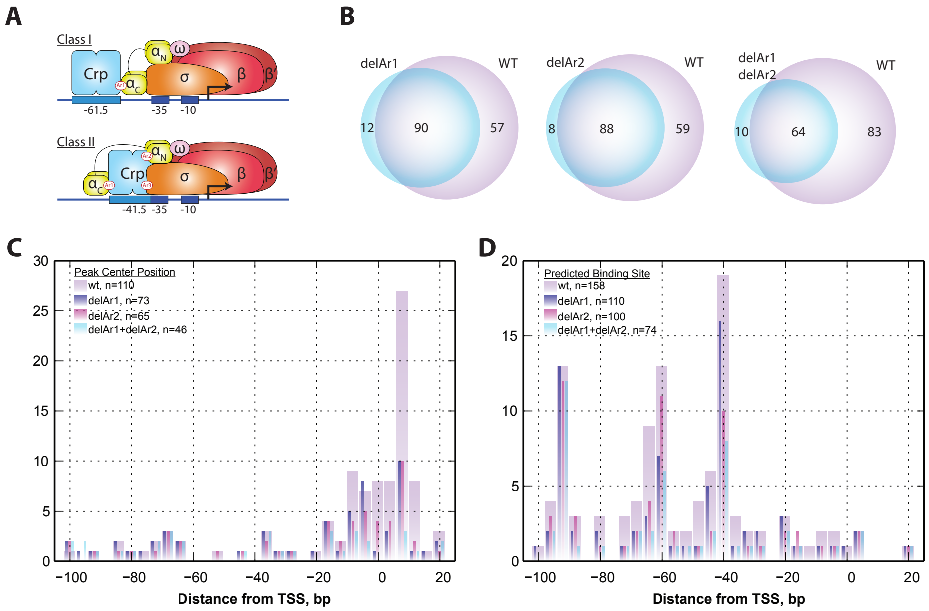


Figure 5.4: **Crp promoter classes have unanticipated ChIP-exo footprint regions.** (A) Gene tracks of Crp ChIP-exo data is shown for different classes of Crp promoters. The deoC promoter (Class III) is regulated by Crp and CytR. Peak regions are found at both Crp operators and for CytR which binds between them. However, it is often observed that for Class I and Class II promoters there is few observed reads over the Crp operator. Instead, the ChIP peak is centered on the TSS and the footprint region co-occurs with that found for σ70. Examples of this are shown for tnaC (Class I) and adhE (Class II). (B) Genome-scale analysis of all Crp peak regions relative to the TSS indicates that the peak center for a majority of Crp peak regions aligns better with the TSS rather than the Crp operator.

5.3.8 ChIP-exonuclease coupled with RNAseq delineates the full Crp regulon

In addition to the ChIP-exonuclease assays we also performed paired RNA sequencing for the wild type and a delta-crp strain under batch glucose, fructose, and glycerol conditions. We further performed RNA sequencing for the delAr1, delAr2, and delAr1delAr2 on glycerol conditions. An overview and compilation of this data is shown in fig5. We show a 91% (21/23) overlap with experimentally validated crp binding sites[106] and a 79% (23/29) overlap with previous ChIP-chip measurements that occurred in characterized Crp binding sites[45]. We further see a 65% overlap (317/486) with all reported crp targets in regulonDB[144] and a conservatively estimated 75% overlap (317/421) with binding sites that are found under similar environmental conditions. As shown in Figure 5, many of the genes that Crp regulates are carbon catabolic operons that are each specific to different utilizable carbon source. Since many of these specific operons are only activated under the presence of the exact carbohydrate that they degrade, they cannot be expected to be active or regulated by Crp under the experimental conditions used here. We thus conservatively removed only 65 of the 486 total crp targets in regulonDB to determine that 317/(486-65) 75% of comparable targets in RegulonDB are found in this study.

The organization of figure 5 is assembled based upon the two principal dimensions of metabolism[145, 146, 86], production of biomass (growth) and the production of energy (chemiosmosis). This organization is nearly identical to the classic and familiar categories of catabolism and anabolism. Each of the three subcategories of catabolism, anabolism, and chemiosmosis are then broken down further into the specific functions of which they are composed. Catabolism thus contains the transport genes to bring in catabolites, along with the recycling or secondary metabolic enzymes to break them

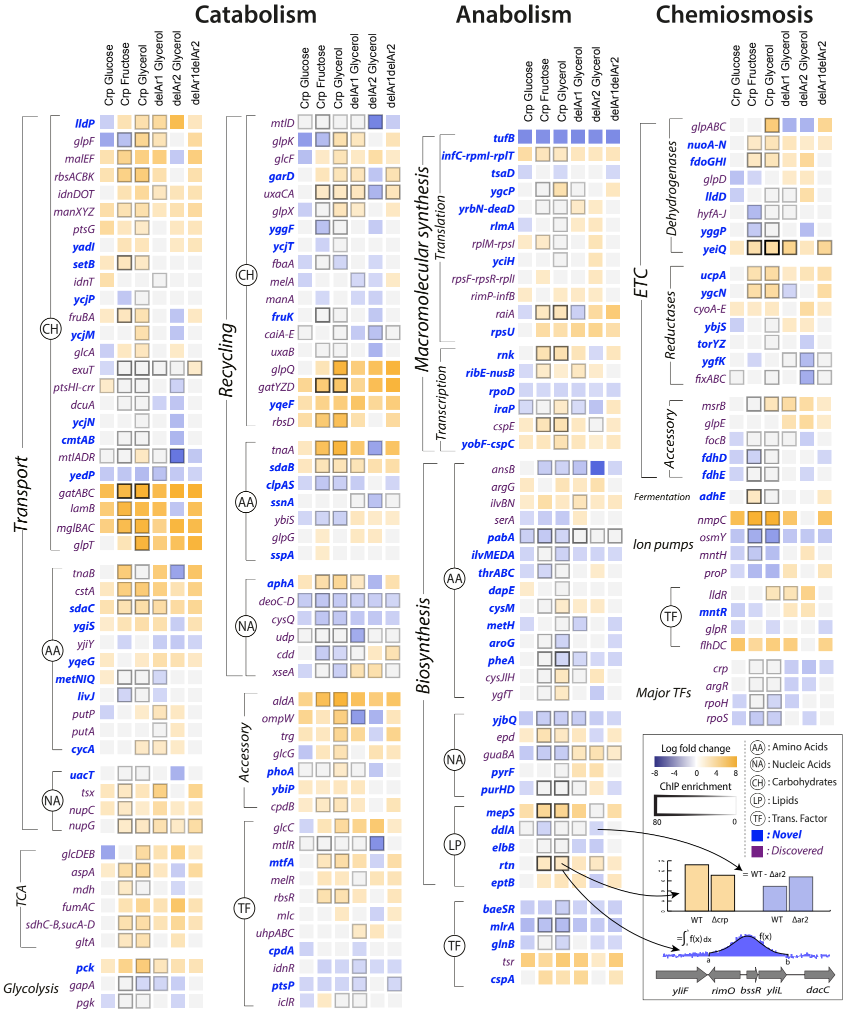


Figure 5.5: **The full Crp regulon defined through paired RNA-seq and ChIP-exonuclease data.** Specific regulation of each gene in the Crp regulon is shown as individual boxes across glucose, fructose, and glycerol environmental conditions with the addition of the Ar1, Ar2, and Ar1Ar2 genetic perturbations across glycerol conditions. Differential expression between a wild type strain and a full Δcrp strain is shown in the first three columns for glucose, fructose, and glycerol. The last three columns show differential expression between a wild type strain and a strain harboring either the Ar1, Ar2, or Ar1Ar2 mutations under glycerol conditions. ChIP-exonuclease peak density is indicated by border intensity for each gene in each column. Groupings of genes are performed similar to Federowicz, 2014 and correspond to the catabolic, anabolic, and chemiosmotic sectors of metabolism. Novel discoveries are indicated and highlight discoveries in translation associated and amino acid biosynthetic gene products. Similarly, broad activation of catabolic genes and mixed regulation of anabolic genes is consistent with previously published models You, 2013. Activation of chemiosmotic genes represents a novel addition to Crp regulatory knowledge.

down along with the central catabolic machinery to complete the degradation process. Anabolism contains both macromolecular synthetic reactions composed of transcription and translation along with biosynthetic reactions composed of specific pathways for carbohydrates, amino acids, and lipids. Chemiosmosis contains any gene responsible for maintaining the electrochemical gradient and subsequent energy supply of the cell. It is thus composed of the ETC, fermentation machinery, and any ion pumps. The most well known and considered regulatory targets of Crp lie in the catabolic sector of metabolism and contain a vast number of transport genes and recycling or secondary metabolic enzymes. We found a total of 97 regulated catabolic genes of which 30 represented novel findings. While we primarily confirm crp regulation of transport genes, we also discover 15 novel transport gene regulation events. One such discovery is ptsP, the nitrogen PTS system, which exists as a completely independent PTS [147] to the main ptsGHI which is also shown to be regulated by crp. Interestingly, pstGHI is seen to be upregulated whereas ptsP is repressed, identifying an isozyme tradeoff enforced by crp regulation. Much of the rest of the crp regulation fills gaps in our knowledge for carbohydrate transporters but also highlights crp regulation of amino acid transporters, in particular metNIJ for methionine, sdaC for serine, and livB for branched chain amino acids. We also confirm crp regulation of recycling and secondary metabolic enzymes but again add 10 novel regulatory targets. In particular, regulation of the clpAS protease, sdaC serine degradation enzyme, and fruK fructose kinase broaden the scope of crp regulation.

A lesser studied but still crucially important swath of crp regulatory targets lie in the anabolic sector of metabolism. Here we found 47 regulatory targets of which 33 represent novel discoveries. While many ribosomal genes were already known to be regulated by Crp, we discover 7 additional translation associated regulation targets. Of particular interest is tufB which carries aminoacylated tRNAs to the ribosome and is one of the most highly expressed [148] proteins in the cell. We also discover extensive regulation of amino acid, nucleotide, and lipid biosynthetic pathways. Activation of arginine and leucine production is complemented by activation of purine and pyrimidine production. One important novel finding is the repression of glnB which plays a crucial role in nitrogen metabolism by regulating the activity of glutamine synthetase. Crp repression of glnB is consistent with its role in regulating both catabolic production and anabolic demand. Additionally, crp is known to regulate rpoS, rpoH, argR, and itself. However, we also show here regulation of rpoD expression underlining crps role as globally decisive transcription factor.

Additionally, many crp regulatory targets lie in the chemiosmotic sector of metabolism including key components of the ETC and fermentation machinery. Here we see 29 regulatory targets of which 14 represent novel findings. Two crucial novel regulatory targets are the primary NADH dehydrogenase, nuoA-N, along with one of the primary fermentation enzymes, adhE, which acts as an alcohol dehydrogenase. Activation of nuoA-N again underlines how Crp acts as a global regulator. Necessarily regulating the high flux backbone throughout catabolism, anabolism, and the energy generation pathways. Similarly, regulation of adhE can enable Crp to control one of the other major routes for generation of membrane potential, especially as Crp responds to the flow of carbon through central metabolism.

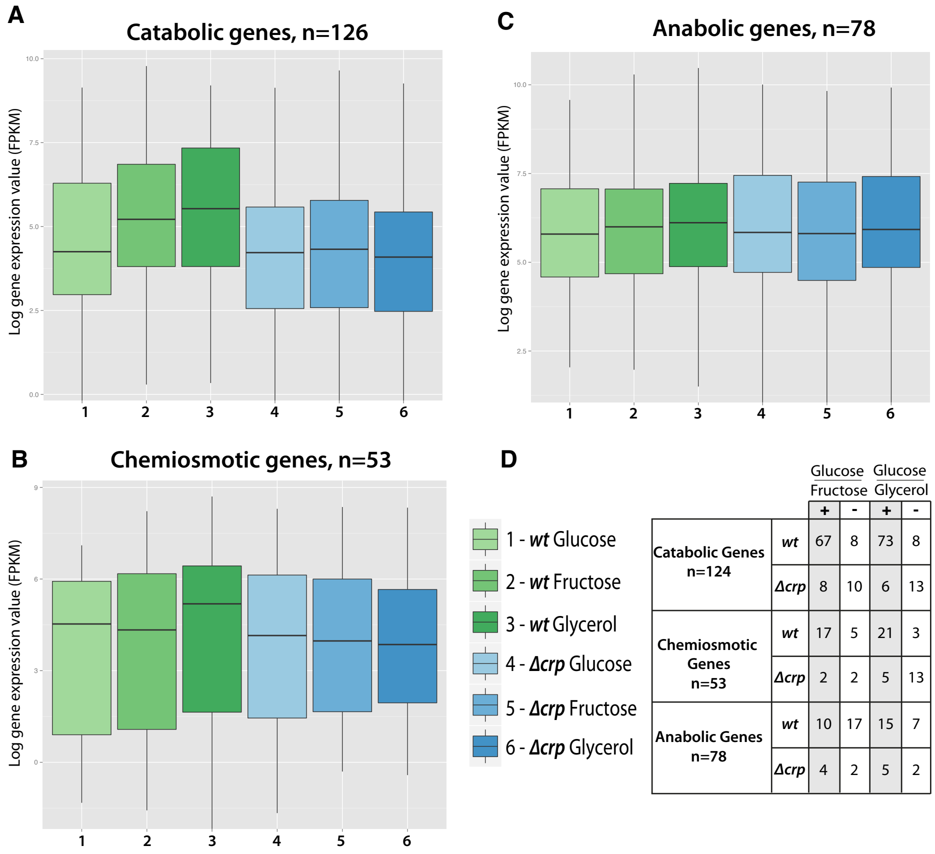
It is also important to note that another 60 regulatory events correspond to genes of unknown function and were not included in figure 5. This brings the total of crp active regulatory targets to 174 across the glycerol, fructose, and glucose conditions studied here. It also highlights that 35% of all active crp regulatory targets occur on genes of completely unknown function which concurs with previous reports [104].

5.3.9 Genome-scale analysis of crp regulon confirms physiological models

In order to understand systems level principles of Crp regulation we sought to characterize the Crp regulon in terms of recent physiological models of Crp regulation [6] and qualitative models of global regulation in microbes [86]. You et al. [6] recently elucidated the physiological function of crp at the systems scale. In particular, this development shows how crp senses carbon flow through central metabolism, and correspondingly up or down regulates the catabolic input or anabolic demand. A strong linear increase was observed between decreasing growth rates due to poorer carbon sources and the expression of catabolic genes. Similarly, a strong linear decrease was also observed among anabolic genes, presumably balancing proteome constraints among the catabolic and anabolic sectors. Here we sought to determine if the broad patterns described by the physiological model would be consistent at the genome-scale. We first grouped genes based on their classification from Figure 5 and then plotted relative expression levels across each carbon source in Figure 6. The median expression levels and quartile distributions markedly shift from glucose to fructose to glycerol conditions for catabolic and chemiosmotic genes. The same plots for strains lacking crp show unaffected expression levels across the same three substrates. This strong positive relation is also observed on historical microarray data across glucose, fructose, and acetate conditions (supplement). Further, the total number of genes which are differentially expressed at an FDR p-value ¡ .05 are shown in Figure 6d. A total of 73 catabolic genes are significantly activated and only 8 are significantly repressed between glucose and glycerol conditions. Similarly 68 catabolic genes are significantly activated and only 8 are repressed between glucose and fructose conditions. This analysis provides strong confirmation for the physiological model of You et al. by recreating their observed C-line at the genome-scale and extends its scope to include energy generation pathways. One key set of genes driving the chemiosmotic trend is the NADH dehydrogenase nuoA-N. This activation of the NADH dehydrogenase is consistent with a strategy in which Crp up regulates both catabolism and energy generation on poor carbon sources to maintain normal cell function at the expense of a higher growth rate via anabolic activity. Further, given its role as the primary dehydrogenase and a high flux backbone for energy generation, this discovery highlights the value in unbiased genome-scale assays.

Anabolic genes, in line with the same physiological model, do not increase in expression levels in response to poorer carbon sources. In the physiological model, a negative linear relation is observed for the anabolic gene sector. While we do not see a negative linear relation, we do see a clear flatlining and lack of clear response in the delta-crp strain. Some anabolic genes likely do respond along the c-line, but many do not. Part of the reason for this may lie in compensation by other transcription factors or cellular machinery at key promoters. Notably, rplM, rlmA, yrbN, and yciH are all important ribosome related genes which display very large ChIP-binding peaks, and differential expression under a delAr2 strain. However, these genes do not exhibit differential expression under the full delta-crp strain. In cases where a full crp deletion does not affect expression levels, but the Ar2 genetic perturbation clearly does, we can conclude that some form of compensation must have occurred at those promoters in the full crp deletion strain. In addition, biosynthetic genes across the amino acid, nucleotide, and lipid categories which include thrABC, metH, pyrF, purHD, ddlA, and elbB all indicate a form of compensation and hence explain the flatlining of overall expression for anabolic genes regardless of crp activity.

ArgG, ilvBN, cysJI, and cysM activation consistent with increased serine uptake through elevated transporter expression sdaC and upregulation of tsr for serine sensing. upregulation of cysteine also concomitant with tnaA activation for cysteine degradation.



		Glucose Fructose		Glucose Glycerol	
		+	-	+	-
Catabolic Genes n=124	*wt*	67	8	73	8
	Δcrp	8	10	6	13
Chemiosmotic Genes n=53	*wt*	17	5	21	3
	Δcrp	2	2	5	13
Anabolic Genes n=78	*wt*	10	17	15	7
	Δcrp	4	2	5	2

Figure 5.6: **Distribution of expression levels for catabolic, anabolic, and chemiosmotic genes reflects physiological models of Crp regulation.** Boxplots display the quartiles and median distributions of all genes in each of the categories defined in Figure 5. (A) Overall expression levels are seen to increase from glucose to fructose to glycerol conditions in wild type strains. The same transition from glucose to fructose to glycerol conditions in δcrp strains results in constant expression levels. (B) A similar trend as observed for catabolic genes is also observed for chemiosmotics genes. This trend is driven in part by novel discovery of the regulation of the NADH dehydrogenase, nuoA-N that exists as a high flux backbone for the E. coli electron transport chain. (C) Anabolic genes do not exhibit the same increase in expression across conditions. This is in line with previous observations and physiological models You, 2013. (D) The number of differentially expressed genes between glucose and fructose or glucose and glycerol are shown for wild type and δcrp strains. In general, the observation that a vast majority (e.g. 67/75 for catabolism glucose/fructose) of the differentially expressed genes at an FDR p-value ¡ .05 between glucose and fructose are being activated indicates that the differences in the boxplot medians are significant. Similarly, this fraction drops off, e.g. 10/27 are activated for anabolic genes between glucose and fructose conditions.

Crp tries to redirect serine to pyruvate to generate extra energy on poorer carbon sources. Repressing transport and biosynthesis of some branched chain amino acids (leu, ile, val) and methionine

5.3.10 Genetic perturbations provide link between promoter mechanisms and systems level regulatory features

Mutations in the functional units of Crp at the Ar1 and Ar2 regions have clear effects on Crp regulation at both the individual promoter level, and overall systems level. We showed in figure 3 how open promoter complexes are destabilized at individual promoters resulting in fewer binding peaks when the functional Ar1 and Ar2 contacts are perturbed. Interestingly, these same Ar1 and Ar2 genetic perturbations also exhibit systems level effects via the overall lessening of Crps ability to regulate the catabolic, anabolic, and chemiosmotic sectors of metabolism Fig 7c. This clear linear decrease across the catabolic and chemisomotic sectors again highlights and solidifies Crps role in activating catabolic genes along the C-line[6]. It also provides a link between the promoter level mechanisms and the systems level regulatory features.

5.4 Conclusion

In conclusion, we have shown that ChIP-exonuclease is able to provide an unprecedented window into bacterial transcriptional activation and that this information can be coupled with systems analysis. This study thus makes a significant contribution towards the longstanding goal of linking structure with function in biological sciences. We first show a striking trimodal distribution for rpoD binding on the lagging strand and a sharp unimodal distribution on the leading strand. This trimodal distribution is shown to correspond to three different groups of promoters with differing motifs and differing

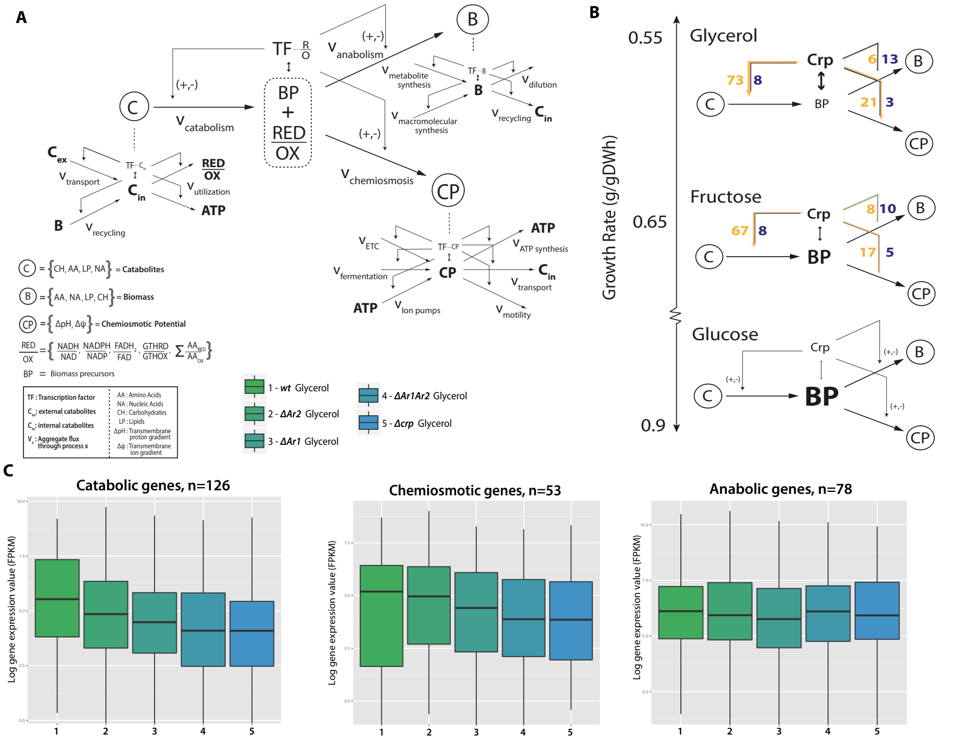


Figure 5.7: **Systems level circuitry of Crp regulation is in line with genetic perturbations.** (A) Systems level model shows the general relation between the anabolic, catabolic, and chemiosmotic metabolic-regulatory networks. Crp is able to sense multiple key biomass precursors (BP) in the form of alpha-keto acids flowing through central carbon metabolism along with cAMP and thereby regulate input catabolic fluxes and consumption anabolic and chemiosmotic fluxes. (B) Throughout the shift from glucose to fructose to glycerol conditions, carbon flux is diminished along with Crp regulation. Crp generally de-activates the catabolic and anabolic sectors in the transition from glucose to glycerol and does not heavily affect the overall expression of anabolic genes. (C) Distributions of expression levels ranging across wild type, Ar1, Ar2, Ar1Ar2, and a full δcrp strain under glycerol conditions. As can be seen, a clear linear decrease from the wild type strain with an intact Crp regulatory protein and the full δcrp provides additional external validation for the existence of the C-line.

functional properties at the systems level. This ultra-high resolution view is also shown to be in solid agreement with structural and biochemical knowledge. In particular, the -19 bp position of the unimodal leading peak aligning perfectly with the onloading of active transcription complexes.

We then repeated ChIP-exonuclease for the canonical activator in E. coli, Crp. Surprisingly the Crp binding profiles showed a remarkable degree of similarity with the rpoD binding profiles. This led us to determine that at Class I and Class II crp promoters in which Crp is known to make direct contacts with rpoD, the full crp-rpoD complex is pulled down and footprinted. This initially surprising result is in clear alignment with structural studies of crp activation and simply highlights that at base pair resolution, ChIP-exonuclease is able to assist in unraveling the dynamics and structure of nucleoprotein complexes. We are also able to show that in Class III cases where Crp does not contact rpod as well as for a transcriptional repressor, ArcA, the footprint occurs perfectly in line with the underlying sequence motif.

At the systems level we are able to confirm recent physiological studies which unraveled the true systems level function of Crp. The model of You et al. showed a feedback mechanism from central metabolism to the catabolic and anabolic genes which Crp regulates. We first confirm this model at the genome-scale by showing that the distribution of expression values for catabolic, anabolic, and chemiosmotic genes in the Crp regulon obey the same trend as the C-line. We then extend this model by showing clear and significant regulation of the energy generating high flux backbone of the cell in the form of the NADH dehydrogenase.

5.5 Materials and Methods

5.5.1 Strains and Culturing Conditions

Escherichia coli MG1655 cells and derivatives thereof were used for all experiments performed in this study. Crp-8-myc, Fnr-8-myc, and ArcA-8-myc tagged strains were those previously constructed and described in [46]. The Δcrp strain was generated using the method described. Briefly, the crp gene was deleted from start codon to stop codon using the λred, FLP-mediated site-specific recombination method and replaced with a gene conferring kanamycin resistance. The Δcrp was transformed with pKD46 and used as a basis for constructing the delAr1, delAr2 and delAr1delAr2 mutant strains using the λred, FLP-mediated site-specific recombination method. Plasmids carrying the different Ar mutant sequences were de novo synthesized using GeneArt with restriction sites at the 5 and 3 end of the gene. Linear PCR constructs carrying a homologous region upstream of the original crp start codon location, the AR mutant crp gene, an FRT site, the chloramphenicol resistance gene, the second FRT, and a second homology region targeting the region downstream of the crp stop codon were electroporated into the Δcrp pKD46 transformed cells. The chloramphenicol resistance gene was then removed from strains with a confirmed insertion of the mutated crp gene by transformation of pCP20 as previously described Datsenko, 2000. The delAr1 mutant introduces a mutation to the Ar1 region, HL159, previously determined to break the contacts between Ar1 and the subunit of RNAP Rhodius, 1997;West, 1993 . The delAr2 mutant does the same for Ar2 but introduces two mutations, KE101 and HY19 Rhodius, 1997. The delAr1delAr2 strain carries the HL159 mutation and the KE101 mutation. M9 minimal media was used for all cultures with 2 g/L of either glucose, fructose or glycerol depending on the conditions tested. For σ70, Crp, Crp8myc, Δcrp, delAr1, delAr2, and delAr1delAr2 experiments, cultures were grown aerobically in shake flasks. Rifampicin conditions were incubated in the presence of rifampicin (50 g/mL final concentration) for 20 min prior to crosslinking as previously described Cho, 2009. Fnr and ArcA experiments were conducted similarly but grown under anaerobic conditions.

Finally we show that the C-line can also be reproduced via analysis of the Ar1 and Ar2 regions. This data shows that the same mutations which destabilize complexes at the level of an individual promoter are responsible for carrying out systems level regulatory features in a consistent and coherent manner.

5.5.2 ChIP Experiments

The ChIP-exo protocol was adapted based on the method described by Rhee et al. and adapted for the Illumina platforms with the following modifications[102]. DNA crosslinking, fragmentation, and immunoprecipitation were performed as previously described[37]. Briefly, cells were crosslinked in early exponential phase in 1% formaldehyde for 30 min at room temperature. This was followed by a 5 min quenching of the crosslinking reaction by addition of glycine to a final concentration of 125 mM. Cells were then washed 3X in ice-cold TBS. Cells were lysed as previously described in the presence of protease inhibitors. Clarified lysate was then continuously sonicated at 4 C using a sonicator bell (6W) for 30 min. Cells were then immunoprecipitated with an appropriate antibody. Antibodies used in this study (all mouse derived) are: anti-Crp (Neoclone N0004), anti-σ70 (Neoclone WP004), anti-Myc (Santa Cruz Biotechnology sc-40). Immunocomplexes were captured using Pan Mouse IgG Dynabeads (Life Technologies). At this point the procedures for ChIP-chip and ChIP-seq deviate the ChIP-exo protocol. ChIP-chip was carried out as detailed in[37] while ChIP-seq undergoes end-repair, dA tailing, adaptor ligation, and PCR enrichment as is done for ChIP-exo. For ChIP-exo, the following steps were performed while the protein/DNA/antibody complexes where bound to the magnetic beads: end repair (NEB End Repair Module), dA tailing (NEB dA-Tailing Module), adaptor 2 ligation (NEB Quick Ligase), nick repair (NEB PreCR Repair Mix), lambda exonuclease treatment (NEB), and RecJf exonuclease treatment (NEB). A series of step-down washes were performed between all steps using buffers previously described. Strand regeneration and library preparation followed the approach of Rhee et al. [102] with the exception of a 3 overhang removal step after the first adaptor ligation and prior to PCR enrichment by treating with T4 DNA Polymerase for 20 min at 12 C. Libraries were sequenced on an Illumina MiSeq. Reads were aligned to the NC_000913.2 genome using bowtie2 Langmead, 2012 with default settings. Peak calling was performed using GPS in the GEMS analysis package [149] with the following parameter settings. Note that GPS was used over GEMS because GEMS peak boundaries are influenced by motif identification whereas GPS is not. ChIP-peak calls were manually curated for anti-Crp conditions on all substrates and rifampicin treated cultures. A superset of GPS peak calls across all anti-Crp conditions was analyzed for presence/absence in each individual condition. Data for ArcA and Fnr were also manually curated.

5.5.3 Gene Expression

Gene expression analysis was performed on RNA-seq data. Briefly, total RNA was isolated and purified using the Qiagen Rneasy Kit with on-column DNase treatment. The quality of the total RNA was assessed using an Agilent Bioanalyzer. Paired-end, strand specific RNA-seq libraries were constructed using the dUTP method. Briefly, total RNA was first depleted of ribosomal RNAs using Epicentres RiboZero rRNA removal kit for gram negative bacterial. rRNA depleted RNA was then primed using random hexamers and reverse transcribed using SuperScript III (Life Technologies). Second strand was synthesized using E. coli DNA Polymerase, Rnase H, and E. coli DNA Ligase and with dNTPs with dUs replacing dTs. Sequencing library construction followed with end-repair, dA tailing, adaptor ligation, removal of the second strand carrying dUs, and

PCR enrichment. Sequencing was performed on an Illumina MiSeq. Reads were mapped to the NC_000913.2 reference genome using the default settings in bowtie2[150].

5.6 Acknowledgements

Chapter 5 in full is a reprint of a manuscript in preparation to be submitted: Latif, H.*, Federowicz, S.A.*, Tarasova, J., Szubin, R., Carreri, J.U., Ebrahim, A., Zengler, K.A., Palsson, B.. Mechanistic and systems level analysis of canonical transcriptional activation in microbes using ChIP-exo. In Preparation. * Indicates equal contribution. The dissertation author was the primary author of this paper responsible for the research. The other authors were Haythem Latif (equal contributor), Janna Tarasova, Richard Szubin, Jose Carreri, Ali Ebrahim, Karsten Zengler and, Bernhard Ø. Palsson.

Chapter 6

Determining the control circuitry of redox metabolism at the genome-scale

6.1 Abstract

Determining how facultative anaerobic organisms sense and direct cellular responses to electron acceptor availability has been a subject of intense study. However, even in the model organism Escherichia coli, established mechanisms only explain a small fraction of the hundreds of genes that are regulated during electron acceptor shifts. Here we propose a qualitative model that accounts for the full breadth of regulated genes by detailing how two global transcription factors (TFs), ArcA and Fnr of E. coli, sense key metabolic redox ratios and act on a genome-wide basis to regulate anabolic, catabolic, and energy generation pathways. We first fill gaps in our knowledge of this transcriptional regulatory network by carrying out ChIP-chip and gene expression experiments to identify 463 regulatory events. We then interfaced this reconstructed regulatory network with a highly curated genome-scale metabolic model to show that ArcA and Fnr regulate ¿ 80% of total metabolic flux and 96% of differential gene expression across fermentative

and nitrate respiratory conditions. Based on the data, we propose a feedforward with feedback trim regulatory scheme given the extensive repression of catabolic genes by ArcA and extensive activation of chemiosmotic genes by Fnr. We further corroborated this regulatory scheme by showing a 0.71 r2 (p ¡ 1e-6) correlation between changes in metabolic flux and changes in regulatory activity across fermentative and nitrate respiratory conditions. Finally, we are able to relate the proposed model to a wealth of previously generated data by contextualizing the existing transcriptional regulatory network.

6.2 Author Summary

All heterotrophic organisms must balance the deployment of consumed carbon compounds between growth and the generation of energy. These two competing objectives have been shown, both computationally and experimentally, to exist as the principal dimensions of the function of metabolic networks. Each of these dimensions can also be thought of as the familiar metabolic functions of catabolism, anabolism, and generation of energy. Here we detail how two global transcription factors (TFs), ArcA and Fnr of Escherichia coli that sense redox ratios, act on a genome-wide basis to coordinately regulate these global metabolic functions through transcriptional control of enzyme and transporter levels in changing environments. A model results from the study that shows how global transcription factors regulate global dimensions of metabolism and form a regulatory hierarchy that reflects the structural hierarchy of the metabolic network.

6.3 Introduction

Regulation of metabolism in response to shifting availability of electron acceptors is a fundamental process in all of biology and is a critical subject for the understanding of pathogenesis, cancer metabolism, and industrial biotechnology. However, even in the model organism Escherichia coli, the regulatory network for this fundamental metabolic function has not been fully elucidated. It has long been known that facultative anaerobes will hierarchically utilize external electron acceptors relative to the free energy change provided by each [151, 152]. Oxygen exists at the top of the hierarchy, electron acceptors like NO3 in the middle, and lactate or acetate or other fermentation products are at the bottom[153, 154, 155]. Many detailed studies have determined that the transcription factors (TFs) ArcA and Fnr are the key players in managing this hierarchy through the activation or repression of the electron transport chain (ETC) machinery specific to an available electron acceptor[156, 157, 158, 20, 159, 160]. It is also largely understood how ArcA senses redox via the flow of reducing equivalents through the ETC, and how Fnr directly senses levels of dissolved O2[161, 162, 163] and glutathione [164]. However, it is not clear how these two TFs work together and more importantly why they regulate hundreds of gene products that lie outside of the ETC and energy metabolism [153, 155]?

Even though many biochemical details of redox regulation have been elucidated [156, 158]park, systems level principles for the global regulatory response throughout the anaerobic shift remain elusive. An important missing piece is a clear framework, or design principle, that elucidates how hundreds of transcriptionally regulated gene products are coordinately regulated to produce the necessary quantitative shifts in metabolic flux states. On the purely metabolic side, certain design principles have emerged through the analysis of stoichiometric models that identified growth and energy generation as the two principal dimensions of metabolic network function[145, 146, 165][17-19]. It was further shown that linear combinations of these two dimensions could account for observed flux patterns throughout nutrient limitations and the anaerobic shift[146, 166]. A question now becomes, what are the corresponding global TFs and how do they coordinately regulate all the gene products which enable the metabolic flux map to shift from one optimal state to another?

Here we show how the global TFs ArcA and Fnr coordinately regulate the primary metabolic dimensions of growth and energy generation. We integrated polyomic data sets and used genome-scale metabolic models to enable a mechanistic understanding of hundreds of simultaneous and individual regulatory events. This analysis subsequently provides a link between global regulatory circuits and global optimality in microbial metabolism.

6.4 Results

6.4.1 Genome-scale identification of TF regulatory events

We first identified individual TF regulatory events at the genome-scale. Side-by-side measurements of RNA transcript abundance and TF binding were carried out to determine the structure and causality in E. colis transcriptional regulatory network (TRN). ChIP-chip assays for ArcA and Fnr were performed under both fermentative and nitrate respiratory conditions (Figure 1A). Gene expression measurements were then used to determine causality of activation or repression for each ArcA or Fnr binding site under these same two conditions (see Figure 3 legend, Figure S1). We found 102, and 86 (and 143 and 132) binding regions and 58 and 54 (and 95 and 55) causal regulatory events for ArcA and Fnr under fermentation (and nitrate respiration) conditions, respectively (Figure 1A, Tables S1-S4). We then compiled the set of genomic sequences underlying these binding regions for each of the TFs and used the MEME program[38] to recover previously identified binding motifs [167, 168] (Figure 1B, Tables S5-S6). We confirmed 180 of 216 (83%) previously known regulatory events[57] and discovered 132 new binding regions relative to RegulonDB (Figure 1A), representing an increase of 74% over current knowledge of the regulatory functions of these two TFs. We further performed a detailed comparison of our results to recently published works parkmyers [16,25] to determine a 78% overlap in ArcA binding sites and a 50% overlap in Fnr binding sites under fermentative conditions (Figures S5-S7). In addition, we report 88 novel binding sites for ArcA and 52 novel binding sites for Fnr under nitrate respiratory conditions highlighting plasticity of the network throughout shifting external electron acceptors.

We then integrated transcription start sites (TSS)[58] with TF binding regions to identify promoter architectures[169]. The location of TF binding motifs within experimentally determined binding regions were used to prepare histograms of the frequency of TF binding relative to the TSS (Figure 1B). This analysis showed that ArcA spans the TSS or -35 box region and represses transcription while Fnr spans the -41.5 or alpha carboxy terminal domain and activates transcription[169]. While each of these regulatory strategies have been shown previously, here can we show that each strategy is ubiquitous at the genome-scale.

6.4.2 Transport coupled redox balancing

The NADH/NAD+ redox pair is critical because a high concentration of NAD+ is absolutely necessary to run glyceraldehyde 3-phosphate dehydrogenase and allow glycolysis to proceed. If the ratio of the NADH/NAD+ becomes too high then the cell will not be able to run glycolysis and perish. Thus a careful system of redox balancing has evolved in which respiratory metabolism is focused on the transfer of reducing equivalents through NADH and under conditions in which respiration becomes impossible the reducing equivalents are dumped onto glycolytic intermediates in the

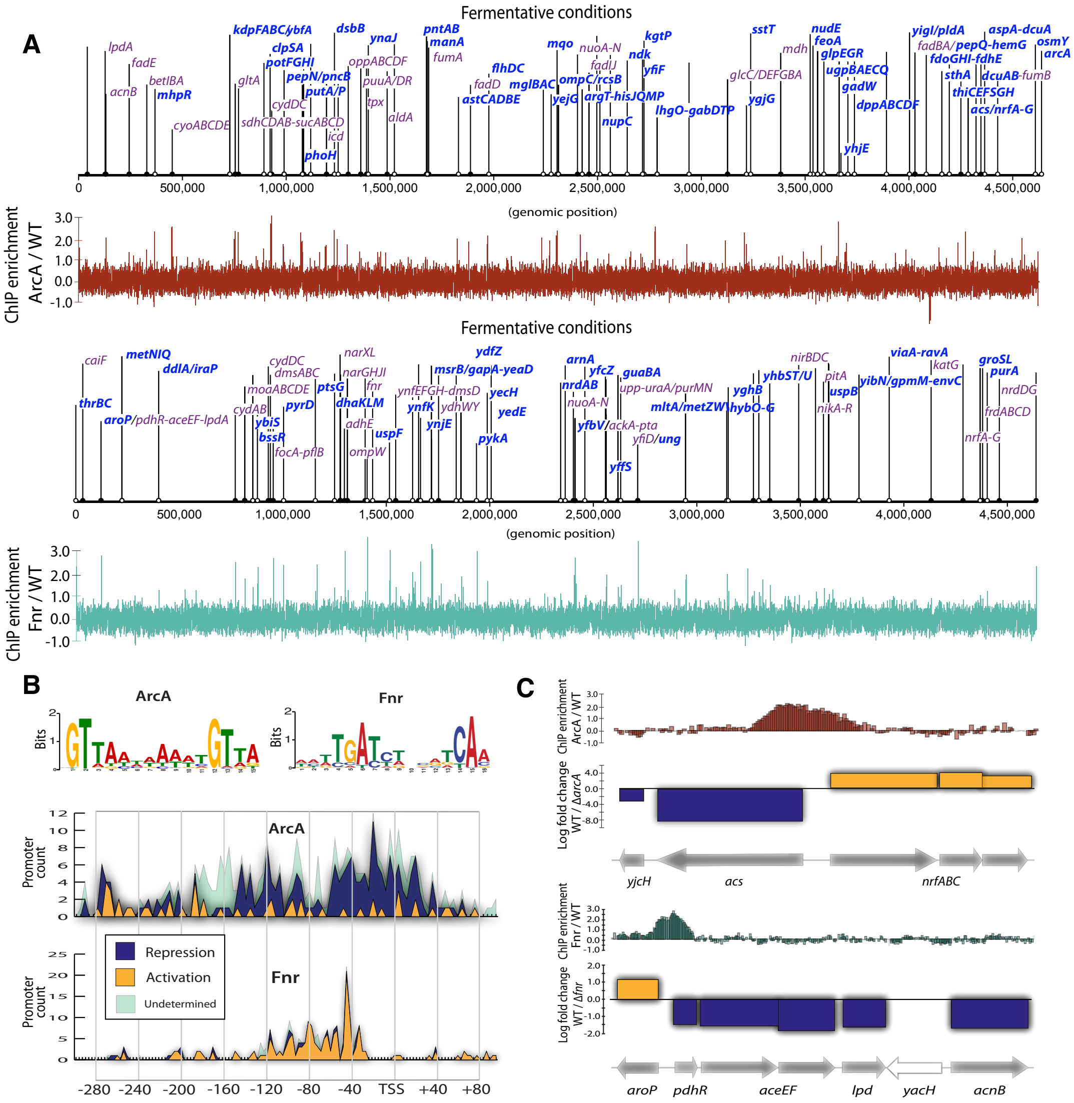


Figure 6.1: **ChIP-chip reveals hundreds of new binding regions and regulatory mechanisms.** (A) Triplicate averaged tracks of ChIP-chip intensity plotted along the length of the genome for ArcA and Fnr under fermentation. We show 83% of previously reported regulatory regions are confirmed (purple) and 132 binding regions (bright blue) are newly discovered relative to RegulonDB. All discovered peaks are shown and operon names included when ChIP peaks also corresponded to differential gene expression for a given operon. (B) Binding motifs are recovered from ChIP binding sites. Histograms of the frequency of motif occurrence relative to the transcription start site (TSS) are plotted and overlaid with gene expression data to reveal ArcA repression via blocking of the 35 box and Fnr activation via upstream binding at 41.5. (C) Transcription factor mediated bi-directional transcription is observed in which a single binding region is shown to regulate divergently transcribed transcriptional units.

process of fermentation. While the general principle of redox balancing has been widely disseminated and utilized[154] and even hypothesized to be mediated by ArcA or Fnr it was not previously possible to mechanistically determine how this phenomena proceeds at the systems level. We took a genome-scale model of E. coli metabolism and sampled it using Monte Carlo methods to determine flux range distributions for all reactions under both fully anaerobic and anaerobic with the addition of nitrate. We then looked at every reaction which utilized NADH/NAD+ and found that 14/19, and 30/35 genes encoding reactions under anaerobic and nitrate conditions were directly regulated by ArcA or Fnr. Drilling down into the 5 unregulated genes (same in both conditions) we found that one encoded fre, a constitutively expressed NAD generation enzyme, and the other four, serA, tyrA, metF, and hisD all encoded amino acid metabolic enzymes. We then took into consideration a puzzling finding of newly discovered and highly significant (fold change regulation of amino acid transporters for serine, tyrosine, methionine and histidine. We noticed that for serA and tyrA in particular, the NADH generating reactions were the subject of end product inhibition by serine and tyrosine. Thus we can hypothesize that activation of the uptake transporters for these amino acids will cause feedback inhibition of the enzymes and thus maintain the expression of critical metabolic enzymes while simultaneously modulating their redox related contributions (Fig. S4).

6.4.3 Discovery of transcription factor mediated bidirectional transcription

Novel cases of divergent transcriptional regulation were found in this data. The integration of binding regions with gene expression data revealed 42 regions where two divergent transcriptional units (TUs) were simultaneously regulated by a single binding event. Divergent transcriptional regulation has been observed previously [28] and is known to be mediated by transcription factors in certain cases. However, systematic regulation by global TFs has only been observed in limited cases [170] [29]. We observe a total of 19 inverse, 16 dual activation, and 13 dual repression events for a total of 48 events spread across the 42 regions as some recur under different experimental conditions.

Two examples (Figure 1C) highlight this hard coupling of the transcriptional regulation of seemingly unrelated but contextually dependent pathways. The acs-nrfABCDE system represents a lowest common denominator coupling between acetyl-coA synthetase (acs) acetate scavenging to acetyl-coA and usage of acetyl-coA via the TCA cycle and nrfABCDE nitrite reductase. Similarly the aroP-pdhR system couples the transport of aromatic amino acids to the regulation of pyruvate that acts as their principal precursor molecule.

The link between the acs and nrfABCD systems has been inferred/suggested in previous work which attempted to understand how E. coli could survive on acetate as a sole carbon source under anaerobic conditions [171]. In particular, E. coli cannot utilize acetate under fully anaerobic conditions because acetate must be scavenged into acetyl-coA via acs and then utilized by the TCA cycle. Anaerobically the TCA cycle cannot be used unless there is an electron acceptor in the ETC to enable oxidative phosphorylation. Thus, some usage of the TCA cycle via an alternative electron acceptor such as nitrite or nitrate is necessary for E. coli to utilize acetate and acetyl-coA anaerobically. This metabolic feature is physiologically crucial in the gut environment that is rich in fatty acids that cannot be used if E.coli does not utilize alternative electron acceptors like nitrite. Hence, the direct coupling of acs and nrfABCD through bidirectional transcriptional regulation is consistent with the necessity of a flux through the nrfABCD system in order for the acetyl-coA formed by acs to be utilized. The transcriptional coupling acts as bidirectional gate controlled by ArcA and the redox state of the cell to coordinate this evolutionarily crucial metabolic capability.

Similarly the aroP-pdhR system couples the transport of aromatic amino acids to the regulation of pyruvate that acts as their principal precursor molecule through the action of Fnr. To understand the network level connection between the aromatic amino acid transporter (aroP) and the pyruvate dehydrogenase repressor TF (pdhR) one can examine Figure 2, which shows the connection between catabolic biomass precursors and biosynthetic pathways. Tyrosine and tryptophan are both made directly from PEP that is rapidly dephosphorylated into pyruvate. The corresponding activation of aroP and repression of pdhR is consistent with an increased need for amino acid transport when the precursors for biosynthesis (PEP) are critical to maintain cellular energy levels. This characteristic is supported by a dampening of the switch upon the transition to nitrate respiration, resulting in decreased transporter expression when less pyruvate is needed for fermentation and can thus be shuttled to amino acid biosynthesis. In general, pdhR acts as a classic repressor that pops off of its binding site in the presence of pyruvate and hence allows expression of pyruvate dehydrogenase and other oxidative enzymes. Anaerobically pyruvate dehydrogenase (aceEF-lpd) is repressed regardless of pdhR by ArcA and Fnr and given that there is also a higher concentration of pyruvate it would presumably not be active. Thus, while this switch is highlighted anaerobically in that full repression of pdhR is concomitant with aroP activation its physiological significance is more prevalent under nitrate or even fully aerobic conditions in which it can function to directly couple and balance the catabolic and anabolic demands around pyruvate which acts as a critical second messenger in the aerobic-anaerobic shift[156]. It is very insightful to view such a switch as it is ramped fully up under anaerobic conditions and then turned down under nitrate respiration to maintain a physiologically crucial metabolic balance.

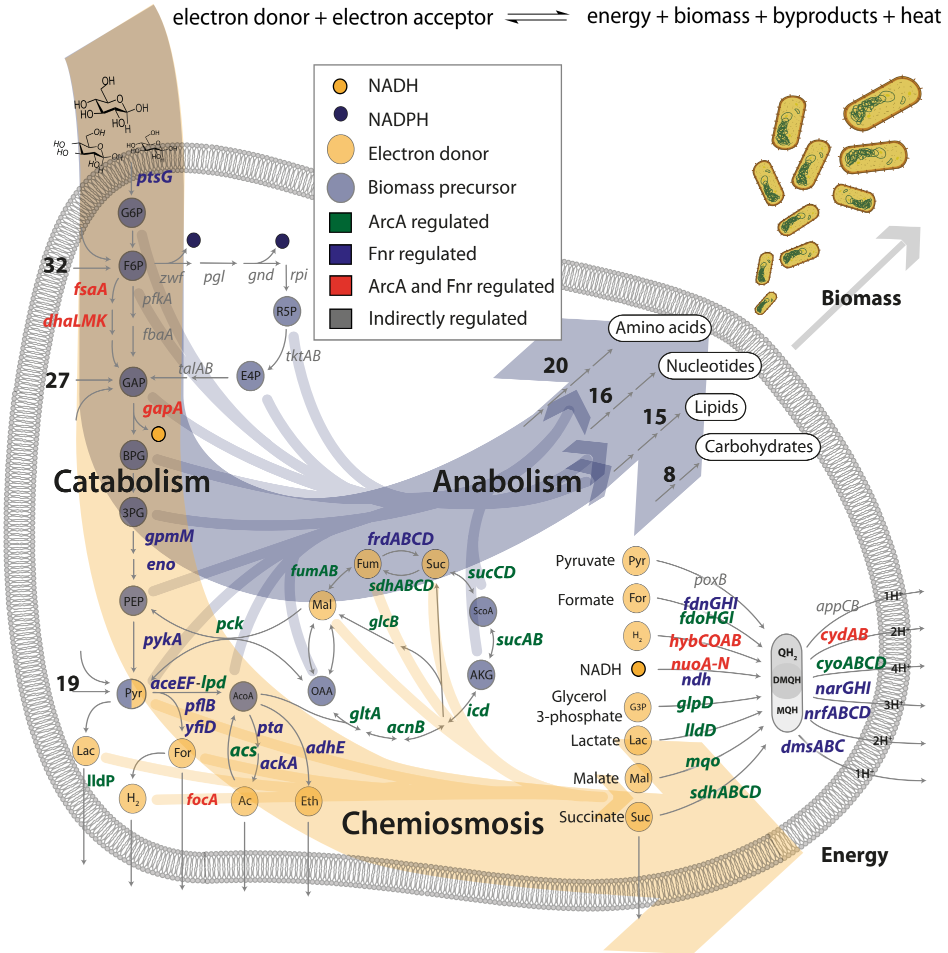


Figure 6.2: **ArcA and Fnr ubiquitously regulate the three branches of metabolism.** Transcription factor regulated gene products are shown in terms of their biological context in the metabolic network. The principal dimensions of metabolism are shown as two large arrows for the formation of biomass or energy. All of the 12 biomass precursors (10/12 regulated) and 9 primary electron donors (9/9 regulated) are shown with arrows flowing into biomass formation or chemiosmosis. The anabolic process is pictorialized with the number of genes regulated in each of the biosynthetic pathways and the chemiosmotic process is shown primarily via the electron transport chain. Numbers indicate the number of regulated genes upstream or downstream of key precursors (e.g. 19 genes encoding reactions for transport and secondary catabolism pathways are regulated upstream of pyruvate).

6.4.4 Ubiquitous regulation of the principal dimensions of metabolism by ArcA and Fnr

Previous work has identified biomass production and energy production as the two principal dimensions characterizing the overall function of metabolic networks[145, 146, 165]. This duality in function is conceptually equivalent to considering heterotrophic metabolism as the standard combustion equation (Figure 2) in which an electron donor (glucose) is broken apart with an electron acceptor (oxygen, nitrate, etc.) to form biomass, energy, waste and heat. Here we use the terms catabolism to describe oxidation of the electron donor, anabolism to describe biomass formation, and chemiosmosis to describe energy generation. The genes in each of these categories were determined by a manual curation of the E. coli metabolic model[97] and associated literature sources [154, 58]. Catabolic genes correspond to nutrient transporters, recycling machinery, and central catabolic machinery. Anabolic genes correspond to biosynthetic and macromolecular synthesis pathways. Chemiosmotic genes correspond to the electron transport chain (ETC), fermentation pathways, and ion pumps (Figure 3).

From the data sets described above, the regulation of these three classes of genes by ArcA and Fnr can be analyzed using their metabolic functions as context. ArcA and Fnr directly regulate a total of 127 catabolic genes including 49 transporter genes, 38 recycling or secondary catabolic enzymes, 33 central metabolic genes, and 7 associated TFs (Figures 2,3). In particular, recovery of all of the classic targets of ArcA and Fnr is complemented by the simultaneous discovery of transporter genes and recycling enzymes like peptidases and proteases (Figure 3). It can also be recognized that there existed many classically unknown glycolytic targets along with generally unrecognized activation of the glucose transporter ptsG. Activation of ptsG by Fnr is consistent with the fact that cells nearly double their uptake of carbon during fermentative growth compared with aerobic growth.

In anabolism, ArcA and Fnr directly regulate 54 genes including 34 metabolite synthesis genes, 14 macromolecular synthesis genes, and 6 TFs. Broad trends of nucleotide biosynthesis activation and amino acid biosynthetic activation of nucleotide precursors is consistent with redox related demands. However, perhaps the most important of these findings is the regulation of both transhydrogenases (sthA, pntAB) in E. coli. Previous work has shown that a large portion of the NADPH used for biosynthetic reactions comes from the membrane bound transhydrogenase PntAB[172] and that the soluble SthA is used for re-oxidation of NADPH under aerobic growth with excess glucose. Our data shows that ArcA activates pntAB and represses sthA in a redox-dependent fashion consistent with an increased need for NADPH under nitrate respiration relative to fermentation (Figure 3). This regulatory shuttling of reduction equivalents thus plays a critical role in maintaining the balance between growth and energy generation by increasing growth only once when energy demands are satisfied.

In the chemiosmotic category we observe regulation of 120 genes including 83 genes of the ETC, 6 for fermentation, 21 for ion pumps, 2 for motility, and 8 TFs. Nearly all of the regulation can be shown to coincide with redox related demands including regulation of ion pumps which coincides with an increased need to maintain a positive electrical gradient across the inner membrane to make up for the diminished proton gradient. We also observed strong regulation of the flhDC, gadW, and appY transcription factors. The flhDC system is a master regulator for the motility and flagellum apparatus of the cell that feeds off the chemiosmotic gradient in search of nutrients. appY and gadW are key regulators of cytochromes and acidic tolerance, respectively. After including regulation through appY we can conclude that ArcA and Fnr exhibit control either directly or indirectly over 15 out of the 16 known dehydrogenase and oxidoreductase reactions in E. coli[173] (Figures 2,3).

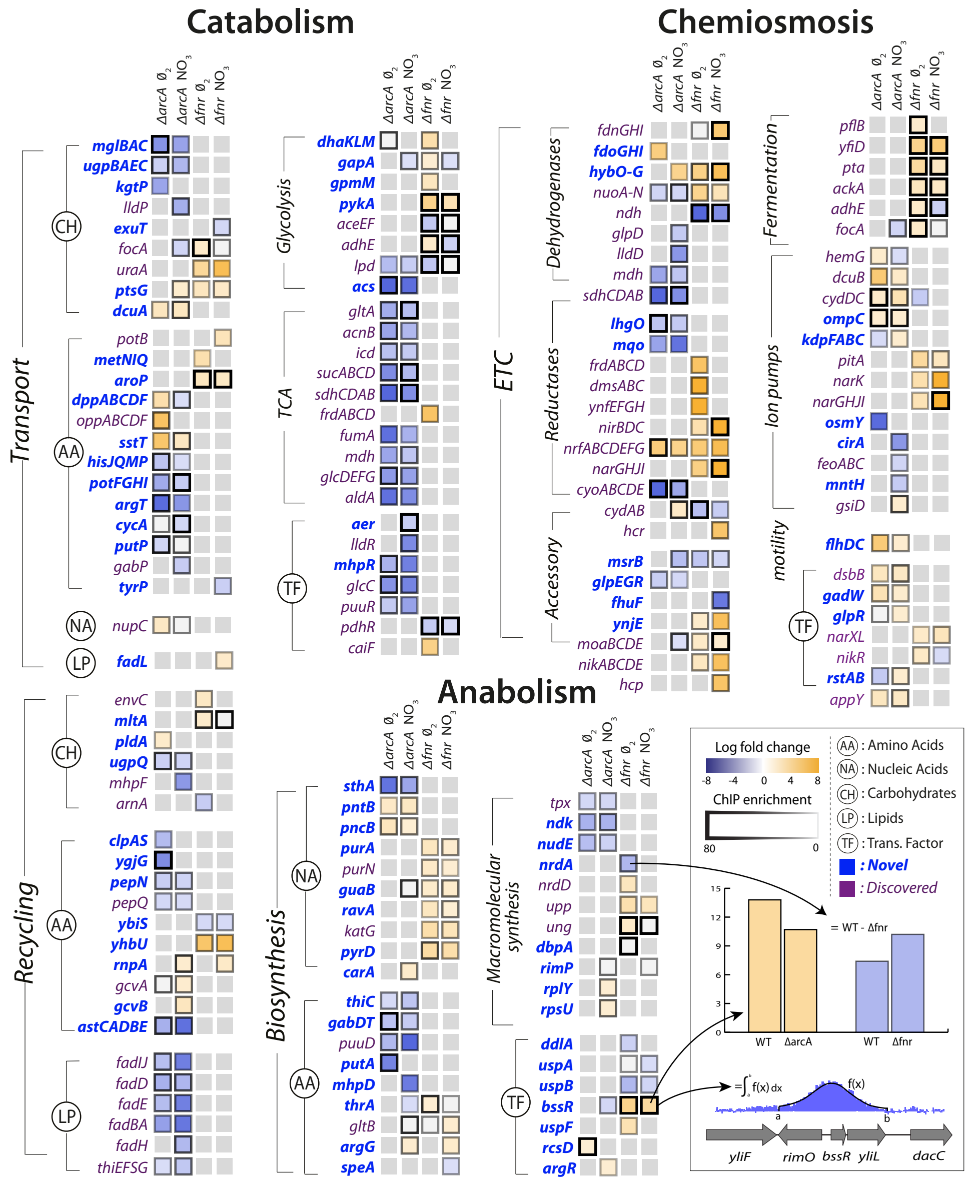


Figure 6.3: **Integration of ChIP-chip, gene expression, and biological context.** Specific regulation of each gene product by ArcA or Fnr under strictly anaerobic and nitrate respiratory conditions are shown as columns. Each box is the result of integration between ChIP-chip and gene expression data in which a TF binding peak was identified and gene expression microarrays showed differential expression upon knockout of the transcription factor in matched conditions. The genes are grouped biologically according to the principal dimensions described in figure 2. Immediate broad trends that emerge are catabolic repression by ArcA and chemiosmotic and anabolic activation by Fnr.

6.4.5 High-level architecture of the metabolic-regulatory network

Enumerating regulatory events is informative, but how do they all together form a coherent regulatory logic that produces meaningful physiological states? Network analysis of these regulatory interactions reveals a qualitative feedforward and feedback flow-based model of the primary metabolic dimensions (Figure 4A). The model input is the total set of catabolites (glucose or electron donor) available to the cell that are oxidized based on the availability of an electron acceptor into a ratio of reduced to oxidized components. These components (primarily NADH/NAD and NADPH/NADP) are then used by the anabolic machinery to generate biomass, or by the chemiosmotic machinery to generate energy as outputs. The ratio of reduced-to-oxidized components is sensed by ArcA and Fnr[151], and they can feedback and feedforward regulate the catabolic, anabolic, and chemiosmotic processes in a coordinated fashion to maintain the ratio. Consistent with this schema, it has been shown that TFs are ideal flux sensors [174].

6.4.6 Feedforward with feedback-trim architecture regulates the anaerobic shift

Analyzing the regulatory events within the context of the qualitative flow-based model reveals a feedforward with feedback-trim architecture of the overall regulatory logic. Counting the number of genes that are activated or repressed (Figure 3) provides a measure of the extent of feedforward or feedback regulation exerted (Figure 4B). Under fermentation ArcA represses 70 catabolic genes and Fnr activates 75 chemiosmotic genes. Under nitrate respiration ArcA represses 73 catabolic genes and Fnr activates 61 chemiosmotic output genes. A similar trend is observed for regulation of the anabolic circuitry in which Fnr activates 14 and 11 genes under fermentation and nitrate respiration.

This circuitry is consistent with fast sensing of oxygen by Fnr and slow but continuous sensing of redox flow through the ETC by ArcA[175].

The regulatory architecture revealed by this qualitative model is comprehensive and novel, but primarily topological. To more quantitatively assess the functions of the observed transcriptional regulatory architecture on the metabolic network that it regulates we sampled all allowable network flux states of a highly curated genome-scale metabolic model of E. coli metabolism[97] under both fermentative and nitrate respiratory conditions. This sampling of allowable flux states of the metabolic network was then integrated with the experimentally determined regulatory architecture to discern the amount of total flux (sum of flux loads across all reactions) regulated by ArcA and Fnr under each of the conditions studied. This calculation revealed that 60% and 57% (and 88% and 80%) of all metabolic flux is directly (and indirectly) controlled by ArcA and Fnr under fermentative and nitrate respiratory conditions respectively (Tables S7, S8). We further show that 69% and 62% of the catabolic fluxes producing each of the redox molecules and biomass precursors along with 71% and 69% of the downstream anabolic and chemiosmotic fluxes are directly regulated under fermentative and nitrate respiratory conditions respectively (Figure S3, Table S9-S10). From a gene level we find that 246 genes are differentially expressed (fdr $<$ .05, fold change $>$ 2) between fermentative and nitrate respiratory conditions and that 236/246 or 96% of the genes are directly (73) or indirectly (163) regulated by ArcA or Fnr (Table S12). Taken together, these measurements quantify the global metabolic regulation of flux by ArcA and Fnr and provide further evidence towards the proposed feedforward with feedback-trim regulatory architecture.

To provide more validation for the feedforward with feedback-trim architecture at the genome-scale we first assessed the set of 91 reactions that significantly differed (flux cutoff of 0.25 mmol /gDW-1 -h-1) between fermentation and nitrate respiration; gDW is denotes grams dry weight. We were then able to show that 89 of the 91 reactions were regulated directly (40 reactions) or indirectly (49 reactions) by ArcA or by Fnr (Table S11). We then calculated the change in flux for each of these 89 reactions between the two conditions along with the change in regulatory strength for the genes encoding these 89 reactions across the same conditions (Table S11). We plotted the change in flux versus the change in regulation (Figure 5A) and calculated an r2 correlation value of 0.71 (p ¡ 1e-6) for the directly regulated genes. This correlation provides quantitative evidence for the logic of the regulatory circuit in the transition from fermentation to nitrate respiration. The linear positive slope shows not only that the reactions responsible for the redox shift are regulated, but also that these reactions are quantitatively regulated to help minimize the redox ratio in concert with the quantitative model predictions. Most of the ArcA regulated reactions are de-repressed, as indicated by the lightening shade of blue under nitrate respiration (Figure 5B). Most of the Fnr regulated reactions are de-activated as highlighted by the lightening shade of yellow under nitrate respiration (Figure 5B). The broad repression of crucial catabolic genes by ArcA and activation of chemiosmotic genes by Fnr is also shown through analysis of C-13 MFA data generated between wild type and Δfnr or ΔarcA strains (Figure S8). This trend of redox ratio minimization was so strong that the only outliers resulted in identification of new biology in the form of transport-coupled redox balancing for allosterically regulated amino acid biosynthetic reactions (Figure S4, Text S1).

We then sought to show that this quantitative regulatory model was truly redox dependent and not just fermentative/nitrate respiration specific. We thus took C-13 measured flux data[176] [35] for E. coli grown aerobically in batch under either fully respiratory galactose conditions or partially fermentative glucose conditions. Even though both conditions are aerobic, we hypothesized that a similar shift in the redox ratio as observed between fully fermentative and nitrate respiration would occur given the comparison between a partially fermentative and fully respiratory condition. We made the same plot (Figure 5C) as in Figure 5a and even used regulatory strengths taken from the fermentative/nitrate shift. Only 16 flux measurements could be mapped of which only 9 showed any difference between glucose and galactose conditions. Of those 9 fluxes we were able to see a clear correlation for 7 and an overall weak but significant r2 correlation value of .26 (p = .079). This plot again shows genes regulated by ArcA being de-repressed and genes regulated by Fnr being de-activated upon the shift to more oxidative conditions (Figure 5D).

6.4.7 Hierarchy of the joint metabolic-regulatory network

An expansion of the top-level of the flow-based model contextualizes the function of the hundreds of individual gene products and provides a window into the structure of the full metabolic-regulatory network (Figure 6A). Each different type of catabolite (Figure 3, Figure 4A, Figure 6A) is maintained via production fluxes (transport or recycling) and consumption fluxes (secondary catabolism or central catabolism). The catabolism specific production set consists of genes for amino acid, carbohydrate, lipid, and nucleic acid transport and recycling. The same expansion can be performed for anabolism and chemiosmosis. For anabolism, the total biomass is a result of the sum of the rate of metabolite biosynthesis plus the rate of macromolecular synthesis[12] minus the rate of dilution and recycling. For chemiosmosis, the total gradient is a sum of protons pumped across the inner membrane via the ETC, proton equivalents pumped across the inner membrane via fermentation, and ions translocated across the inner membrane minus the usage of the gradient for ATP production, nutrient transport, and motility[177].

This expansion also accounts for the classically observed hierarchy[178] of the TRN via sensing of lower level metabolites and subsequent regulatory control of the TFs themselves or of the production or consumption pathways for sensed metabolites

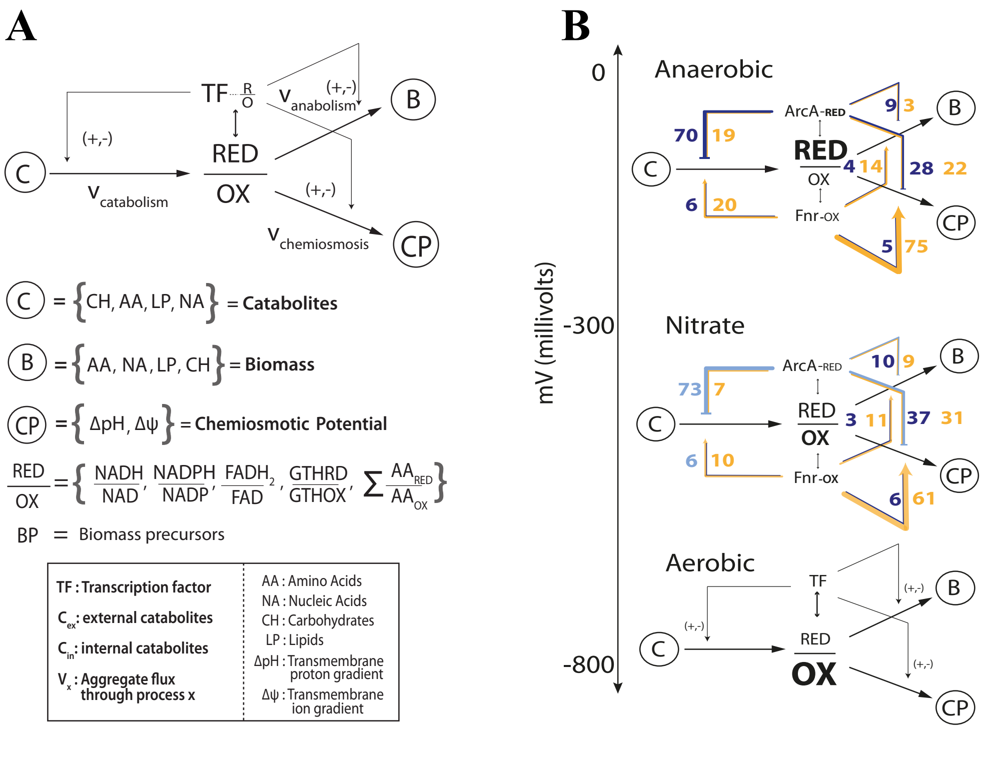

A
B
TF
R
O
(+,-)
$V_{anabolism}$
B
(+,-)
RED
OX
C
$V_{catabolism}$
(+,-)
$V_{chemiosmosis}$
CP
C = {CH, AA, LP, NA} = Catabolites
B = {AA, NA, LP, CH} = Biomass
CP = {ΔpH, Δψ} = Chemiosmotic Potential
$\frac{RED}{OX} = \left\{ \frac{NADH}{NAD}, \frac{NADPH}{NADP}, \frac{FADH_2}{FAD}, \frac{GTHRD}{GTHOX}, \sum \frac{AA_{RED}}{AA_{OX}} \right\}$
BP = Biomass precursors
TF : Transcription factor
C_{ex}: external catabolites
C_{in}: internal catabolites
V_x : Aggregate flux through process x
AA : Amino Acids
NA : Nucleic Acids
CH : Carbohydrates
LP : Lipids
ΔpH : Transmembrane proton gradient
Δψ : Transmembrane ion gradient
0
-300
-800
mV (millivolts)
Anaerobic
ArcA-RED
RED
OX
Fnr-OX
70
19
6
20
9
3
4
14
28
22
5
75
C
B
CP
Nitrate
ArcA-RED
RED
OX
Fnr-ox
73
7
6
10
10
9
3
11
37
31
6
61
C
B
CP
Aerobic
TF
RED
OX
(+,-)
(+,-)
(+,-)
C
B
CP

Figure 6.4: **Flow based model of the metabolic-regulatory network explains regulation throughout the anaerobic shift.** (A) Considering a mass balance around the ratio of reduced to oxidized molecules allows the unification of catabolism, anabolism, and chemiosmosis into a single process. The ratio of reduced to oxidized molecules is then sensed by ArcA and Fnr to elicit corresponding feedforward and feedback regulatory circuitry which allows the cell to maintain this critical ratio. (B) Mapping of the regulation of gene products (Figure 3) for each branch of the circuit reveals a broad trend of feedforward with feedback-trim regulation. Under fermentative conditions the redox ratio is high and the observed regulation works to lower the input and activate the output to bring the ratio down. Under nitrate respiration, the ratio drops and the circuit maintains a similar number of connections but is shown to decrease in gross activity levels.

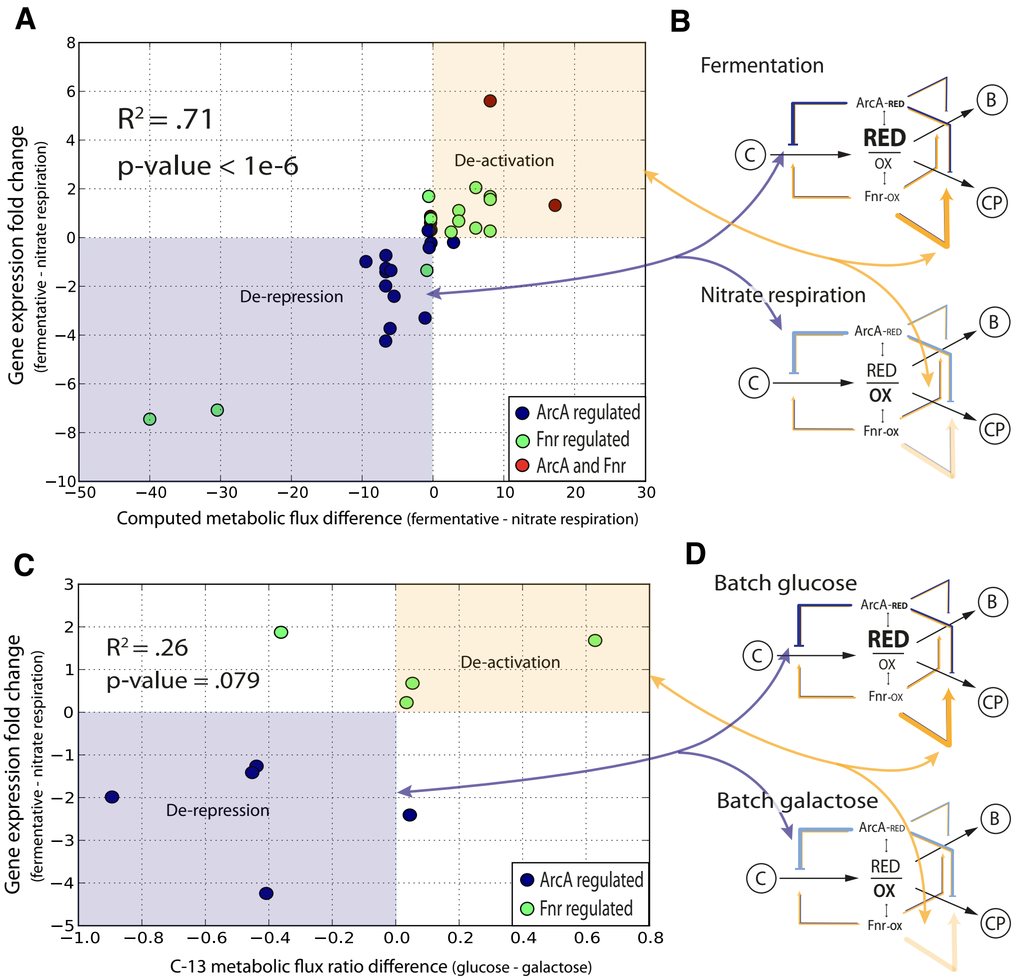


Figure 6.5: **Quantitative correlation of shifts in regulatory strength between experimental conditions with shifts in flux through regulated enzymes.** (A) The decrease in activity levels from anaerobic to nitrate is quantified by calculating a correlation between the change in flux for all altered reactions across nitrate and anaerobic conditions with the change in level of regulation across the same conditions. (B) Overlaying information for the specific regulators shows that ArcA is involved in the derepression of key reactions going from fermentation to nitrate respiration and Fnr is involved in deactivation. (C) The shift between glucose and galactose under batch growth mirrors the respiratory shift from fully fermentative to nitrate respiratory conditions. C-13 labeled fluxomic data generated for wild type cells under both glucose and galactose batch conditions is used to generate the same plot as in (A) and is even plotted against the same regulatory strengths between fully fermentative anaerobic cultures and nitrate respiring cultures. (D) One can again see that key ArcA regulated genes are de-repressed whereas Fnr regulated genes are de-activated.

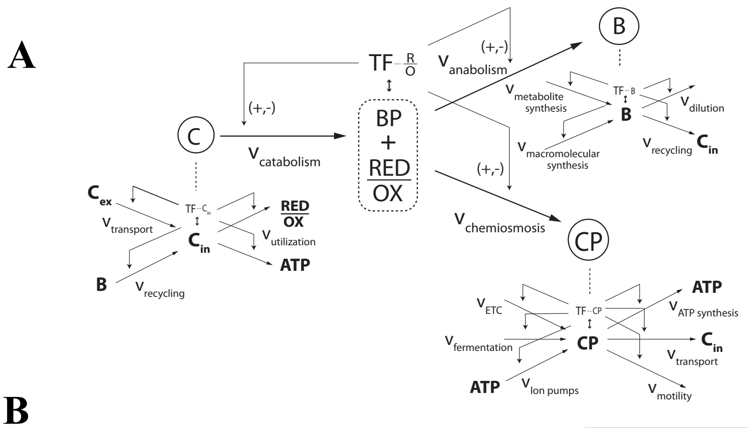
(Figure 6B). A full tracing of the TRN to explain the effects of the global TF deletion is consistent with 69% of observed differential expression (Figure S2).

6.5 Discussion

This work presents a systems level and genome-scale mechanism for the coordinate action of global transcription factors throughout an electron acceptor shift. Our mechanism accounts for the previously unexplained genes regulated by ArcA and Fnr, it predicts changes in flux patterns, and perhaps most importantly shows that the classically observed hierarchy of transcriptional regulation mirrors the hierarchy of dimensions in the metabolic network. By basing our work off of the extensive body of detailed biological literature and the more recent work of principal dimensionality in metabolic networks we are able to present a systematic and remarkably consistent genome-scale mechanism.

At the local level, we first greatly expanded the number of cases of promoter architectures[137]. This validates and highlights the importance of understanding initiation mechanisms, as they may be extendable to a systems level in future development of computational models. We were then able to make the novel discovery that 42 regions across the genome contained divergently transcribed TUs controlled by a single global TF binding region. We recognize that due to ChIP-chip resolution it is possible (and even likely) that multiple binding sites exist under the larger ChIP peak, however the local proximity still affords the same hard-coupling within the regulon. This hard coupling suggests switch like mechanisms in which sets of seemingly unrelated genes are jointly regulated to obey non-obvious systems level constraints. We identify two such cases of this in the acs-nrfABCDE operon and the aroP-pdhR operon.

To understand systems level mechanisms of transcriptional regulation we turned to



B

Catabolism

TF	# Catabolic genes	# Non-Catabolic genes
NtrC	32	1
CysB	23	0
GntR	12	0
PaaX	11	0
MalT	10	0
FadR	10	1
GalS	9	0
GalR	9	0
AraC	9	0
CdaR	9	0
AllR	8	0
NanR	6	0
UlaR	6	0
DeoR	6	0
GlcC	6	0
IdnR	6	0
MhpR	6	0
UxuR	5	0
XylR	5	0
16 TFs*	40	0

Catabolism + Anabolism + Chemiosmosis

TF	# Cat. genes	# Ana. genes	# Chem. genes
Crp	253	69	42
ArcA	97	55	68
Fnr	33	44	76
Fur	51	10	32
Cra	29	8	12
PdhR	16	21	8

Catabolism + Anabolism

TF	# Cat. genes	# Ana. genes	# Non-* genes
ArgR	41	55	11
Lrp	44	40	4
PurR	20	25	0
NagC	20	5	0
RutR	9	4	2
CytR	9	2	0
Nac	7	7	0
MetJ	3	9	0
TyrR	3	7	0
DpiA	2	8	1

Anabolism

TF	# Anabolic genes	# Non-Anabolic genes
TrpR	13	1
DnaA	8	1
ArgP	6	1
MetR	5	0
BirA	5	0
AsnC	3	0
FabR	2	0
LysR	1	0

Chemiosmosis

TF	# Chem. genes	# Non-Chem. genes
NarL	101	10
NarP	47	0
FhlA	29	0
IscR	24	1
AppY	9	0
HycA	1	0

Catabolism + Chemiosmosis

TF	# Catabolic genes	# Chemiosmotic genes
ModE	12	34
CaiF	7	4
GlpR	5	4
DcuR	3	4
LldR	1	1

Figure 6.6: **Topological organization of the joint metabolic-regulatory network.** (A) Qualitative model where levels of the hierarchy represent a coarse graining of the total metabolic network around pools of key metabolites. Each metabolite has mass balanced production and consumption fluxes and often a corresponding TF sensor that can regulate the input and output fluxes. (B) Quantitative assessment of this regulatory scheme done by curation and classification of the regulatory targets for every TF known to sense a metabolite in the iJO1366 model (Table S13). The main point is that the hierarchy of the regulatory network does in fact mirror the hierarchical dimensionality of the metabolic network. Regulators which sense a catabolite only regulate catabolic genes, regulators which sense an anabolite only regulate anabolic genes, and regulators which sense a chemiosmotic component only regulate chemiosmotic genes. However, metabolites that exist as both a catabolite and anabolite, or as both a catabolite and chemiosmotic component, tend to have regulators which regulate genes in each of the given categories. Similarly, TFs which sense molecules that are biomass precursors and energy precursors will necessarily globally regulate metabolic genes in all three categories.

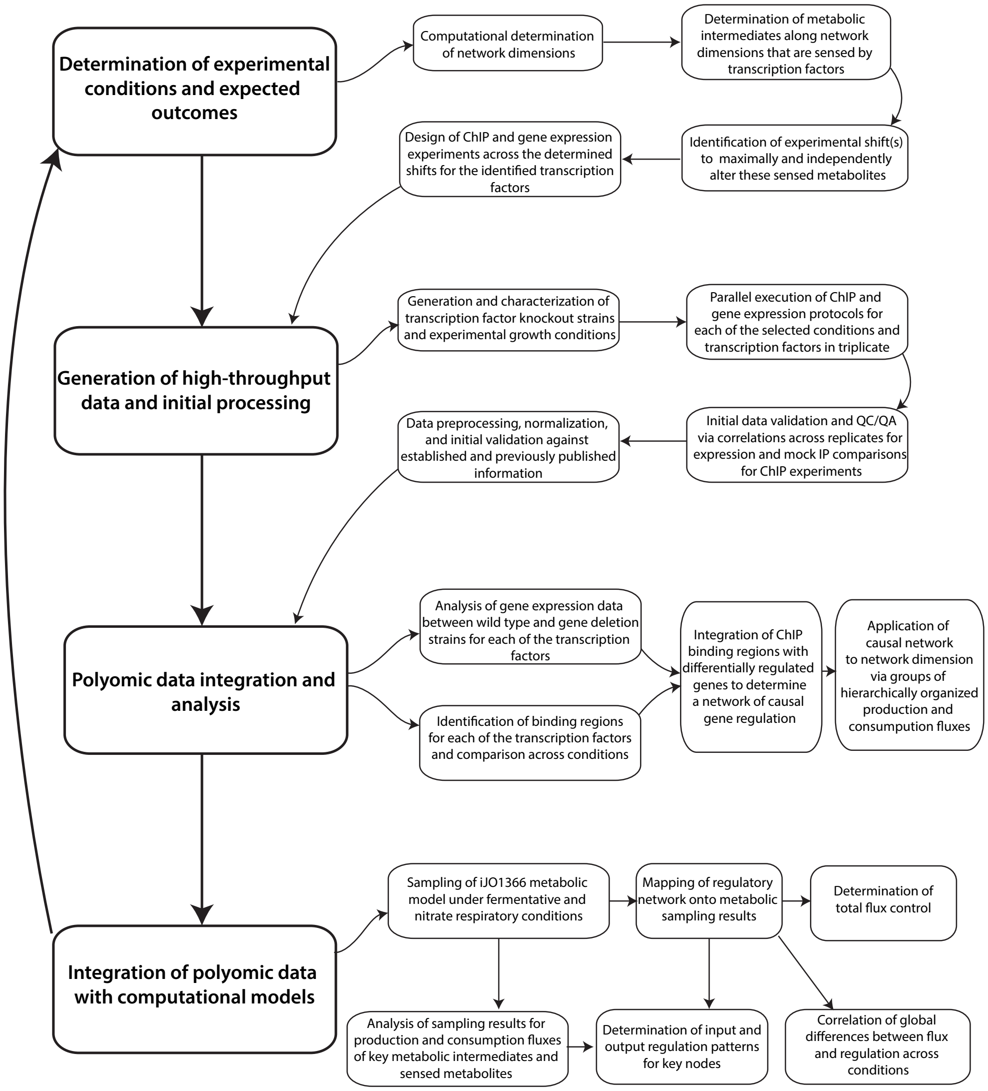

Determination of experimental conditions and expected outcomes
Computational determination of network dimensions
Determination of metabolic intermediates along network dimensions that are sensed by transcription factors
Identification of experimental shift(s) to maximally and independently alter these sensed metabolites
Design of ChIP and gene expression experiments across the determined shifts for the identified transcription factors
Generation of high-throughput data and initial processing
Generation and characterization of transcription factor knockout strains and experimental growth conditions
Parallel execution of ChIP and gene expression protocols for each of the selected conditions and transcription factors in triplicate
Initial data validation and QC/QA via correlations across replicates for expression and mock IP comparisons for ChIP experiments
Data preprocessing, normalization, and initial validation against established and previously published information
Polyomic data integration and analysis
Analysis of gene expression data between wild type and gene deletion strains for each of the transcription factors
Identification of binding regions for each of the transcription factors and comparison across conditions
Integration of ChIP binding regions with differentially regulated genes to determine a network of causal gene regulation
Application of causal network to network dimension via groups of hierarchically organized production and consumption fluxes
Integration of polyomic data with computational models
Sampling of iJO1366 metabolic model under fermentative and nitrate respiratory conditions
Mapping of regulatory network onto metabolic sampling results
Determination of total flux control
Analysis of sampling results for production and consumption fluxes of key metabolic intermediates and sensed metabolites
Determination of input and output regulation patterns for key nodes
Correlation of global differences between flux and regulation across conditions

Figure 6.7: **Workflow overview of the experimental and computational analysis process.** An integrated and iterative loop was used to generate the integrated regulatory and metabolic analysis.

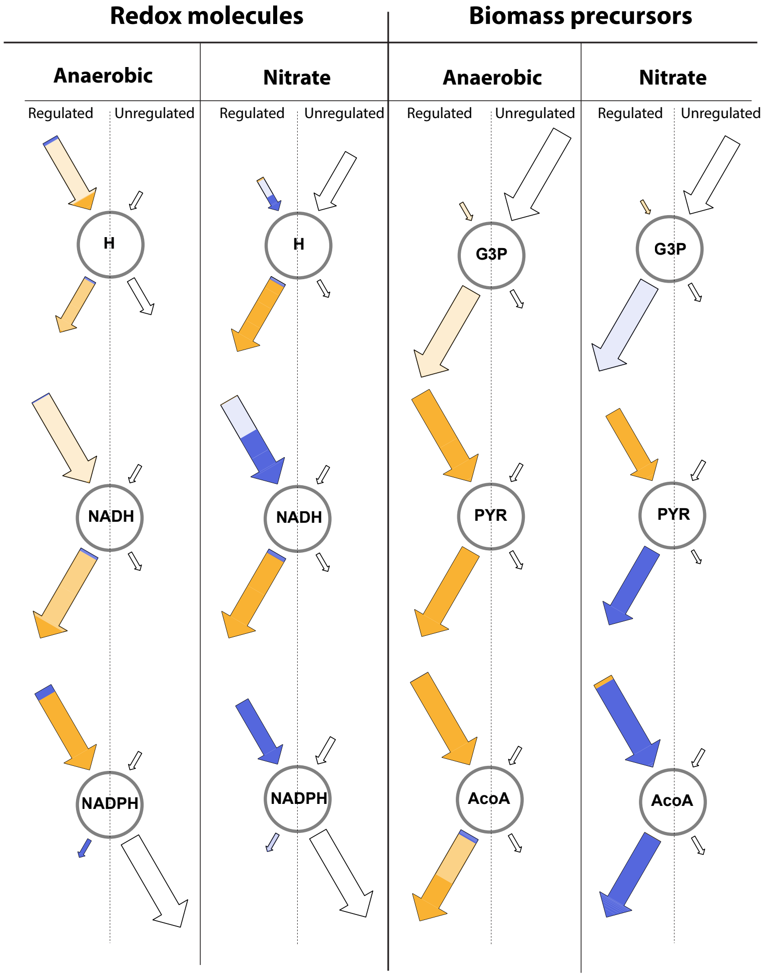


Figure 6.8: **Regulation of fluxes around key metabolic intermediates I.** Twenty-four different metabolites are profiled, including 12/13 biomass precursors, 9/9 primary electron donors, and the three primary electron carriers, H+, NADH, and NADPH. Each node map diagram shows the split between the amount of regulated vs. unregulated flux that goes into the production or consumption of each metabolite. The notable pattern is repression of the consumption and often production upon a shift to nitrate respiratory conditions. This occurs primarily as a means of negative feedback on the flux through these core nodes. In fact these diagrams fail to show that under fermentative conditions these same fluxes through core nodes are even more highly repressed. This occurs because the metabolic network at optimality is already in line with the regulation, and hence does not carry flux through many of the reactions that are shown to be repressed under nitrate respiratory conditions. This result led us to make the scatter plot of Figure 5A which more clearly displays the higher degree of repression in fermentation vs. nitrate respiratory conditions along with deactivation through the shift. All data tables and associated code is available at http://nbviewer.ipython.org/ea455904c0d7cda4bfba.

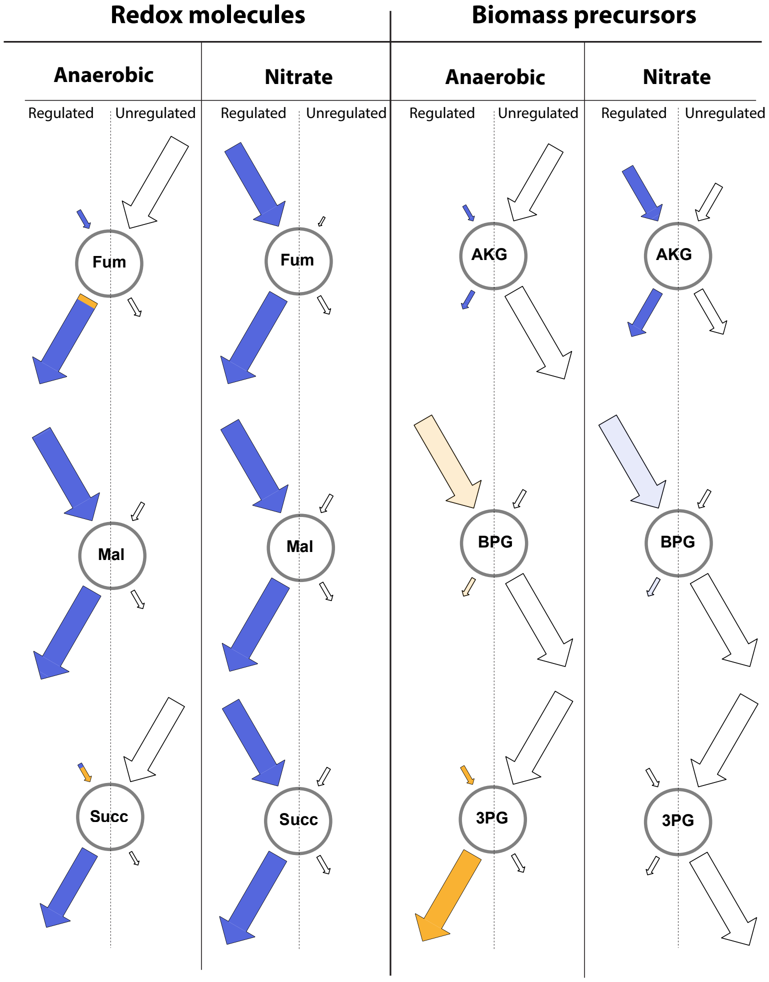


Figure 6.9: **Regulation of fluxes around key metabolic intermediates II.** Twenty-four different metabolites are profiled, including 12/13 biomass precursors, 9/9 primary electron donors, and the three primary electron carriers, H+, NADH, and NADPH. Each node map diagram shows the split between the amount of regulated vs. unregulated flux that goes into the production or consumption of each metabolite. The notable pattern is repression of the consumption and often production upon a shift to nitrate respiratory conditions. This occurs primarily as a means of negative feedback on the flux through these core nodes. In fact these diagrams fail to show that under fermentative conditions these same fluxes through core nodes are even more highly repressed. This occurs because the metabolic network at optimality is already in line with the regulation, and hence does not carry flux through many of the reactions that are shown to be repressed under nitrate respiratory conditions. This result led us to make the scatter plot of Figure 5A which more clearly displays the higher degree of repression in fermentation vs. nitrate respiratory conditions along with deactivation through the shift. All data tables and associated code is available at http://nbviewer.ipython.org/ea455904c0d7cda4bfba.

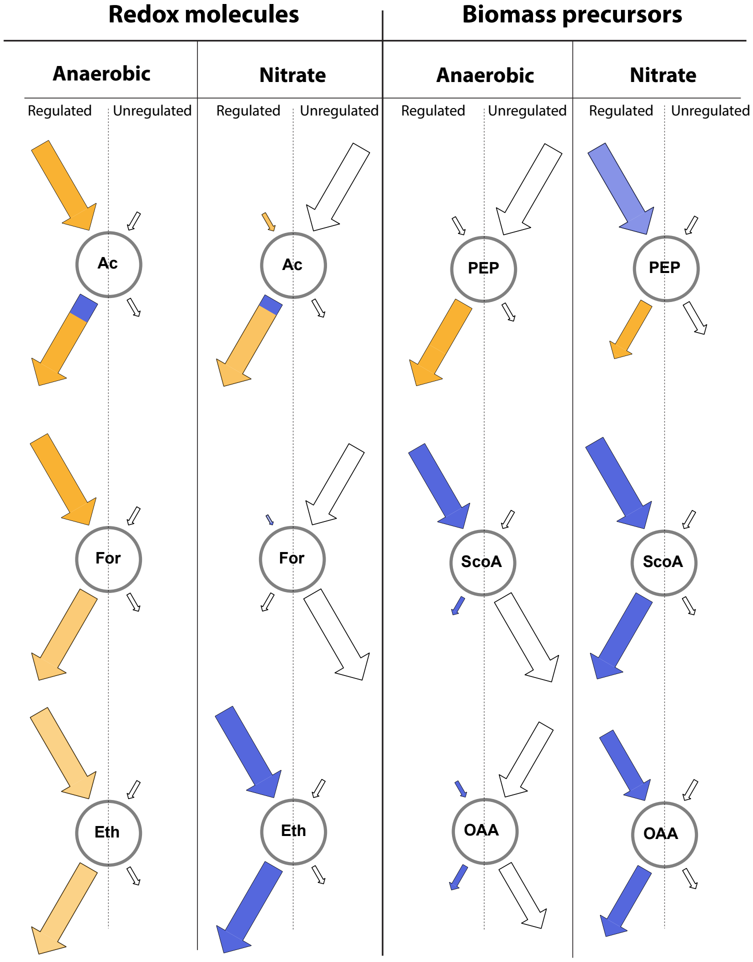


Figure 6.10: **Regulation of fluxes around key metabolic intermediates III.** Twenty-four different metabolites are profiled, including 12/13 biomass precursors, 9/9 primary electron donors, and the three primary electron carriers, H+, NADH, and NADPH. Each node map diagram shows the split between the amount of regulated vs. unregulated flux that goes into the production or consumption of each metabolite. The notable pattern is repression of the consumption and often production upon a shift to nitrate respiratory conditions. This occurs primarily as a means of negative feedback on the flux through these core nodes. In fact these diagrams fail to show that under fermentative conditions these same fluxes through core nodes are even more highly repressed. This occurs because the metabolic network at optimality is already in line with the regulation, and hence does not carry flux through many of the reactions that are shown to be repressed under nitrate respiratory conditions. This result led us to make the scatter plot of Figure 5A which more clearly displays the higher degree of repression in fermentation vs. nitrate respiratory conditions along with deactivation through the shift. All data tables and associated code is available at http://nbviewer.ipython.org/ea455904c0d7cda4bfba.

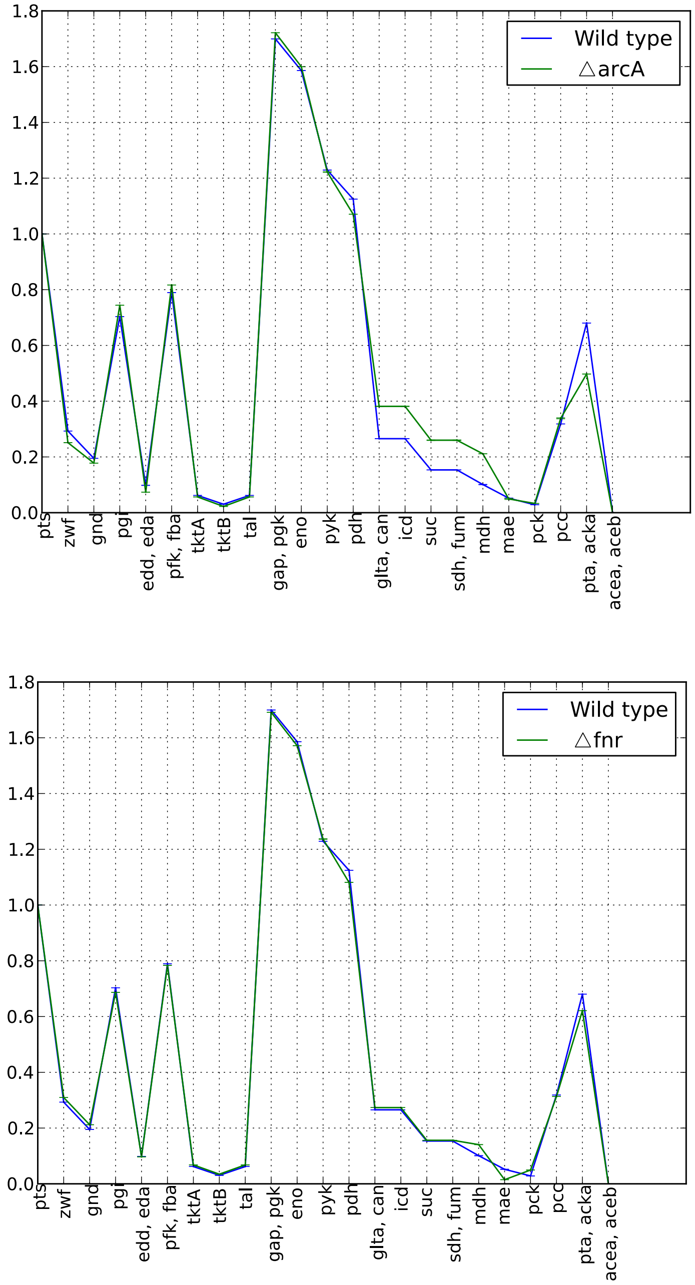


Figure 6.11: **C-13 MFA data for ArcA/Fnr regulation targets.** We compared C-13 MFA derived flux values [36] gathered for wild type strains and Δfnr or ΔarcA strains under partially fermentative glucose batch growth. It can be seen that deletion of arcA does cause de-repression of the key catabolic fluxes of the TCA cycle. This causes less flux to be directed towards the fermentative chemiosmotic pathways ultimately wasting energy.

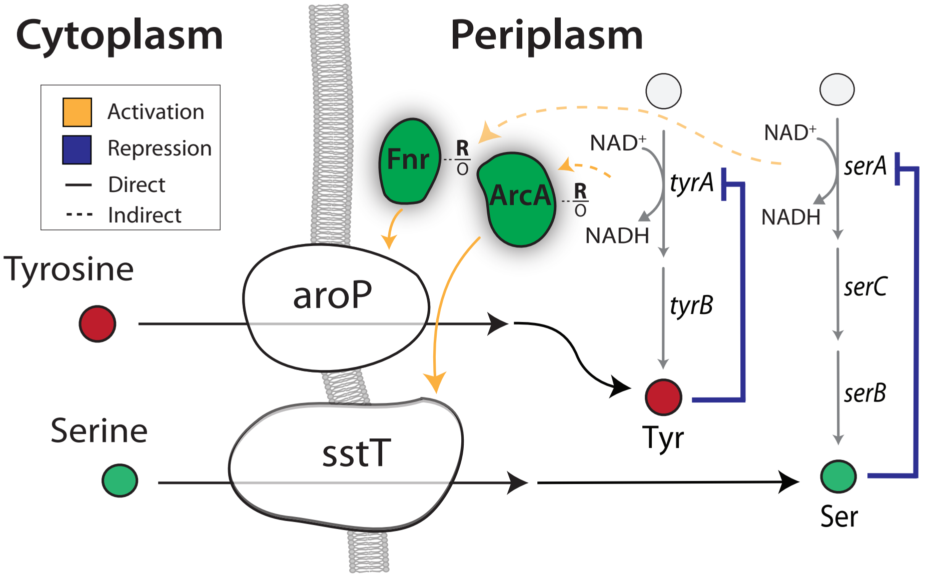


Figure 6.12: **Transport coupled redox balancing.** After sampling the metabolic model and determining all reactions that produce or consume NADH, we identified only 5 reactions that carried flux and were not regulated by ArcA of Fnr. We found that one encoded fre, a constitutively expressed NAD generation enzyme, and the other four, serA, tyrA, metF, and hisD all encode amino acid biosynthetic enzymes. We then took into consideration a puzzling finding of newly discovered and highly significant regulation of amino acid transporters for serine, tyrosine, methionine and histidine. We noticed that for serA and tyrA in particular, the NADH generating reactions were the subject of end product inhibition by serine and tyrosine. Thus we can hypothesize that activation of the uptake transporters for these amino acids will cause feedback inhibition of the enzymes and thus maintain the expression of critical metabolic enzymes while simultaneously modulating their redox related contributions.

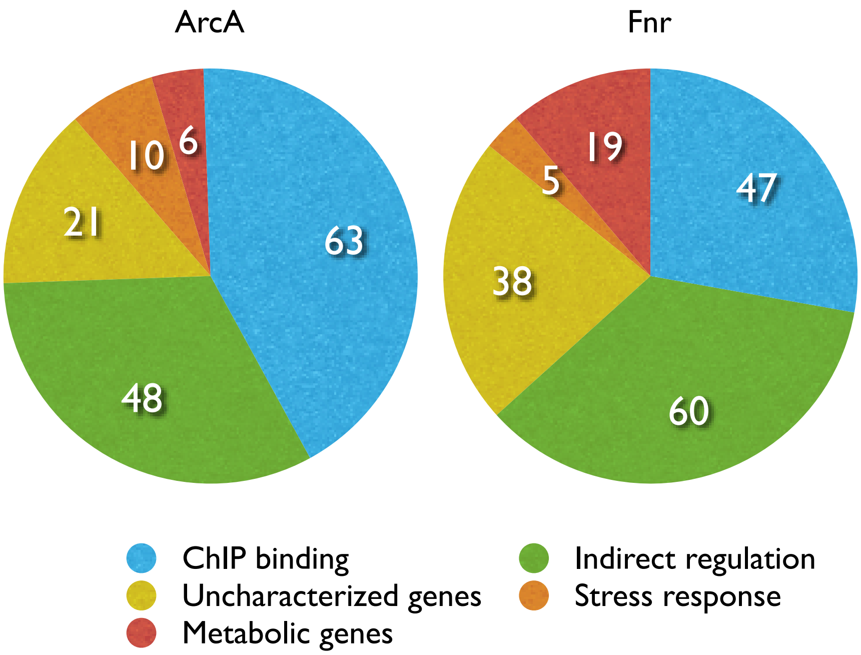


Figure 6.13: **Causative classification of genes differentially expressed log two fold between a wild type and ΔarcA or Δfnr strain under fully fermentative conditions.** After deletion of the arcA and fnr transcription factor genes, 148 and 169 genes are differentially expressed under anaerobic conditions. We then trace the regulatory network to explain the regulation of these genes. 63 and 47 are shown to be directly regulated through binding of the TFs in the ChIP-chip data. Another 48 and 60 genes are indirectly regulated via secondary network effects (Regulation by a local TF that is directly regulated by ArcA or Fnr). Finally the last three categories represent genes involved in the stress response, genes of unknown function, and other metabolic genes. Differentially regulated genes that are primarily stress response genes may represent variability in culture conditions or unknown regulatory interactions. Uncharacterized and metabolic genes likely represent unknown regulatory links.

previous work that showed the principal dimensions of a metabolic space were biomass and energy generation. We hypothesized that global regulators must play a role in regulating globally decisive metabolic dimensionality. This hypothesis is supported by broad regulation across all of these main categories and the abilities of ArcA and Fnr to sense the molecules that govern the branch point between the two dimensions.

Although we were able to make an unbiased characterization of the genes in each of the categories using the iJO1366 model we were still unsatisfied with such a coarse grained approach and sought to understand the composition of each of the categories. This led us to a hierarchical expansion and classification of pathways around key metabolic intermediates. Going on in this fashion led us to realize that the global transcriptional regulatory hierarchy plays out not only on the level of TF-TF regulation, but perhaps more importantly at the level of global TFs regulating the production or consumption fluxes of lower level metabolites which are correspondingly sensed by other intermediate regulators. In essence, the regulatory network is shaped by the underlying metabolite pools and vice versa.

After determining the broad circuitry of the metabolic-regulatory network we mapped our data onto it and discovered that a strong feedforward with feedback trim architecture dominates at the genome scale. This occurs via ArcAs strong repression of input catabolic circuits coupled with Fnrs strong activation of downstream chemiosmotic and anabolic circuitry. This circuit is corroborated by Fnrs ability to sense oxygen[163] which will diffuse quickly whereas ArcA will more continuously sense the flow of reducing equivalents through the ETC by sensing of the ratio of reduced to oxidized quinones[162]. This pattern of a fast component feeding forward for downstream planning coupled with a slower but continuous feedback sensor is a common pattern in basic process control schemes[179][40]. If coupled with other common process control patterns such as hierarchical and PID control one can envision a process control based model for the entire joint metabolic-regulatory network.

This work presents a formal integration and reconstruction of over 50 years of research on E. coli metabolism and its transcriptional regulation. The result is a detailed and coherent hierarchical view of the regulation of the principal dimensions of metabolism through a critical environmental shift. We find that the mathematical notions of optimality in metabolic functions are in line with our observations of global regulation. TRNs are not just TF-gene networks but rather TF-gene-enzyme-reaction flux networks, that are tightly integrated as levels or ratios of metabolites can drive TF activity[180][181]. The full elucidation of an electron acceptor response in the important model organism, E. coli, may have implications for similar metabolic responses in other organisms. For cancer, recent focus has shifted towards an understanding of the metabolic drivers and Warburg effect, where the hypoxia inducible factor (HIF) [182] senses the redox ratio and feedforward or feedback regulates genes producing or consuming reduction potential.

Taken together, we are able to show how the two principal dimensions of metabolism are controlled in a shifting environment by global TFs through the use of polyomic data sets and genome-scale metabolic models. This study is likely to be useful as a guide for similar studies in other organisms where the same tools for experimentation and analysis are available.

6.6 Methods

6.6.1 Bacterial strains and growth conditions

All strains used in this study were E. coli K-12 MG1655 and its derivatives. The E. coli strains harboring Fnr-8myc and ArcA-8myc were generated as described previously[46]. The deletion mutants (Δfnr and ΔarcA) were constructed by a red and FLP-mediated site-specific recombination method. Glycerol stocks of E. coli strains were inoculated into M9 minimal medium containing 0.2% (w/v) carbon source (glucose) and 0.1% (w/v) nitrogen source (NH4Cl), and cultured overnight at 37 C with constant agitation. The cultures were diluted 1:100 into fresh minimal medium and then cultured at 37 C to an appropriate cell density with constant agitation. For the anaerobic cultures, the minimal medium were flushed with nitrogen and then continuously monitored using a polarographic-dissolved oxygen probe (Cole-Parmer Instruments) to ensure anaerobicity. For nitrate respiration 20 mmol potassium nitrate was added.

6.6.2 ChIP-chip

To identify Fnr and ArcA binding regions in vivo, we used the ChIP-chip approach as described previously[37, 46]. Briefly, cells at appropriate cells density were cross-linked by 1% formaldehyde at 20 C for 25 min. Following the quenching of the unused formaldehyde with a final concentration of 125 mM glycine at 20 C for 5 min, the cross-linked cells were harvested and washed three times with 50 ml of ice-cold Trisbuffered saline. The washed cells were resuspended in 0.5 ml lysis buffer composed of 50mM Tris-HCl (pH 7.5), 100 mM NaCl, 1 mM EDTA, 1 g/ml RNaseA, protease inhibitor cocktail (Sigma) and 1 kU Ready-Lyse lysozyme Epicentre). The cells were incubated at 37 C for 30 min and then treated with 0.5 ml of 2 IP buffer with the protease inhibitor cocktail. The lysate was then sonicated four times for 20s each in an ice bath to fragment the chromatin complexes using a Misonix sonicator 3000 (output level, 2.5). The range of the DNA size resulting from the sonication procedure was 300-1,000 base pairs (bp). The specific antibodies that specifically recognizes myc tag (9E10, Santa Cruz Biotech) were used to immunoprecipitate each chromatin complex, respectively. For the control (mock-IP), 2 g of normal mouse IgG (Upstate) was added into the cell extract. The remaining ChIP-chip procedures were performed as described previously. The high-density oligonucleotide tiling arrays used to perform ChIP-chip analysis consisted of

371,034 oligonucleotide probes spaced 25 bp apart (25 bp overlap between two probes) across the E. coli genome (Roche NimbleGen). After hybridization and washing steps, the arrays were scanned on an Axon GenePix 4000B scanner and features were extracted as a pair format by using NimbleScan 2.4 software (RocheNimbleGen).

6.6.3 qPCR

To monitor the enrichment of promoter regions, 1 L immunoprecipitated DNA was used to carry out gene-specific qPCR. The quantitative real-time PCR of each sample was performed in triplicate using iCycler (Bio-Rad Laboratories) and SYBR green mix (Qiagen). The real-time qPCR conditions were as follows: 25 L SYBR mix (Qiagen), 1 L of each primer (10 pM), 1 L of immunoprecipitated or mock-immunoprecipitated 3DNA and 22 L of ddH2O. All real-time qPCR reactions were done in triplicates. The samples were cycled to 94 C for 15s, 52 C for 30s and 72 C for 30s (total 40 cycles) on a LightCycler (Bio-Rad). The threshold cycle values were calculated automatically by the iCycler iQ optical system software (Bio-Rad Laboratories). Any primer sequences used were described previously[46].

6.6.4 Transcriptome analysis

Samples for transcriptome analysis were taken from exponentially growing cells. From the cells treated by RNAprotect Bacteria Reagent (Qiagen), total RNA samples were isolated using RNeasy columns (Qiagen) in accordance with manufacturers instruction. Total RNA yields were measured using a spectrophotometer (A260), and quality was checked by visualization on agarose gels and by measuring the sample A260/A280 ratio (¿1.8). Affymetrix GeneChip E. coli Genome 2.0 arrays were used for genome-scale transcriptional analyses. cDNA synthesis, fragmentation, end-terminus biotin labeling, and array hybridization were performed as recommended by Affymetrix standard protocol. Raw CEL files were analyzed using robust multi-array average for normalization and calculation of probe intensities. The processed probe signals derived from each microarray were averaged for both the wild type and deletion mutant strains.

6.6.5 ChIP-chip and expression data analysis

To identify TF-binding regions, we used the peak finding algorithm built into the NimbleScan software. Processing of ChIP-chip data was performed in three steps: normalization, IP/mock-IP ratio computation (log base 2), and enriched region identification. The log2 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 bi-weight mean. The log2 ratio of Cy5 (IP DNA) to Cy3 (mock-IP DNA) for each point was calculated from the scanned signals. Then, the bi-weight mean of this log2 ratio was subtracted from each point. Each log ratio dataset from duplicate samples was used to identify TF-binding regions using the software (width of sliding window =300 bp). Our approach to identify the TF-binding regions was to first determine binding locations from each data set and then combine the binding locations from at least five of six datasets to define a binding region using the MetaScope software (http://sbrg.ucsd.edu/Downloads/MetaScope). Raw gene expression CEL files were normalized using background corrected robust multi-array average implemented in the R affy package. To detect differential expression between the wild type and TF deletion strains we applied a two-tailed unpaired students t-test between the experimental triplicates for the wild type and gene deletion strains. This was followed by a false discovery rate adjustment. Before performing the FDR correction we removed all genes that exhibited an expression level below the background across all experiments. The background level was calculated as the average expression level across all intergenic probes. We then only considered genes meeting a 5% FDR (false discovery rate)-adjusted P-value cut-off to be differentially expressed. ChIP binding tracks for Figure 1a and the heatmap for Figure 3 were generated using D3[183]. Related code is available at http://nbviewer.ipython.org/gist/steve-federowicz/7cceedba73982c0ae995. All raw and processed data have been deposited in NCBI/GEO under accession number GSE55367 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55367).

6.6.6 Motif searching

The ArcA and Fnr binding motif analysis was completed using the MEME and FIMO tools from the MEME software suite[38]. We first determined the proper binding motif and then scanned the full genome for its presence. The elicitation of the motif was done using the MEME program on the set of sequences defined by the ArcA and Fnr binding regions respectively. Using default settings the previously determined ArcA and Fnr motifs were recovered and then tailored to the correct size by setting the width parameter to 18-bp and 16-bp respectively. We then used these motifs and the PSPM (position specific probability matrix) generated for each by MEME to rescan the entire genome with the FIMO program.

6.6.7 Promoter architecture determination

We integrated transcription start sites (TSS) with our TF binding regions to identify promoter architectures genome wide[169, 141]. We first determined the location of motif binding sites within experimentally determined binding regions. We then calculated the distance between motif center position and previously determined TSS locations[58]. Finally, we prepared a histogram of the number of motifs that occur at varying distances relative to the TSS (Figure 1B) and included the gene expression data to determine the regulatory outcome of each binding event. The results showed that ArcA spans the TSS or -35 box region and represses transcription while Fnr spans the -41.5 or alpha carboxy terminal domain[141] and activates transcription. The histograms also reveal the previously reported trend[184] of motif frequency oscillation at a roughly 10.5 bp interval consistent with helical phasing of the DNA strand.

6.6.8 Genome-scale metabolic sampling

To perform sampling we first generated pFBA[185] constrained models of the iJO1366[97] metabolic model corresponding to fermentative and nitrate respiratory conditions. Fermentative conditions were simulated by setting the lower bound of the oxygen exchange reaction (EX_o2) to zero. Nitrate respiratory conditions were simulated by setting the lower bound for nitrate uptake (EX_no3) to -20 mmol gDW-1 h-1 (mirroring experimental addition of 20 mmol KNO3) along with the lower bound of EX_o2 set to zero. pFBA constrained models were generated by first using the convertToIrreversible() function of the COBRA toolbox[186] followed by a standard FBA for growth rate. This growth rate was then imposed as a constraint in a subsequent optimization that found the minimum sum of flux able to achieve that growth rate. Finally, using the gpSampler()[186] method we sampled each of the pFBA constrained models. All sampling runs were for a full 24 hours to ensure a mixing fraction below .55. After sampling was performed we took the average across the 7046 sampling points (2n where n = 3,523 reactions in the metabolic model). Sampling results were then interfaced with the regulatory network and metabolic model via the COBRApy project (http://opencobra.sourceforge.net/openCOBRA), iPython notebook[187], and in-house databases.

6.7 Acknowledgements

Chapter 6 in full is a reprint of a published manuscript: Federowicz, S.A.*, Kim, D.K., Ebrahim, A, Lerman, J.A., Nagarajan, H.N., Cho, B.K., Zengler, K.A., Palsson, B.. Determining the Control Circuitry of Redox Metabolism at the Genome-Scale. Plos Genetics. 2014 Apr;10(4): e1004264 doi: 10.1371/journal.pgen.1004264. The dissertation author was the primary author of this paper responsible for the research. The other authors were Donghyuk Kim, Ali Ebrahim, Joshua Lerman, Harish Nagarajan, Byung-Kwan Cho, Karsten Zengler, and, Bernhard Ø. Palsson.

Chapter 7

Conclusion

In conclusion, the regulatory networks of a cell are dense, and often extremely complicated. However, when studying first the individual components and then the symphony of these components among the full network of regulation, one can begin to see broad patterns emerge. Many of the principles seen in the individual study of regulatory networks or metabolism, such as robustness, hierarchy, feedback, and temporal positioning are clear in the joint metabolic-regulatory network. Yet, many more principles emerge when studying the joint network. Three key principles covered throughout this thesis are the congruence between optimal metabolic flux states and observed regulation data, the shared dimensionality among metabolic and regulatory networks, and the segregation of the targets of the regulatory network relative to metabolic function.

This work seeks to present one of the first full in-vivo integration and analyses of multi-factor global regulatory circuits across well-defined physiological shifts. Being able to determine the primary inputs and outputs for the quantity being sensed and controlled allows one to assess the global feedforward and feedback circuits that act to maintain homeostasis upon small perturbations and drive a state change when necessary. Doing the genome scale integration can also bring out novel regulatory mechanisms and solidifies our primary knowledge for downstream engineering. Going forward it will be possible to stoichiometrically account for every single binding site, the physiological state of the TF (ligand bound, phosphorylated etc) at that binding site, the nucleoprotein complexes formed by these TFs at each promoter, and the initiation or repression characteristics of the promoter given these constraints. This will enable an unbiased and ultimately parameter free assessment of the circuit integration or logic capacities at each individual promoter. When this is coupled to an appropriate objective function it will be possible to almost fully predict the thermodynamic and evolutionary constraints which have shaped an organism.

7.1 Transcriptional regulation of metabolism as a process control problem

To shed further light on the control circuitry presented in Chapters 2, 3, 5, and 6, we sought to classify the individual elements based upon standard control principles. The first and most striking is the maintenance of a common ratio that finds a perfect analogy with the widely studied class of ratio controllers. In practice a ratio controller often occurs in chemical plants or other settings in which two feeds (or pipes) are being combined together at a specified ratio. In practice, the simplest configuration is to have a sensor for each incoming flow, a logic controller that combines the sensory information, and then a single control valve that adjusts the flow of one of the pipes such that the ratio of the two is kept constant. While this is the simplest and ultimately linear configuration it is often the case that the control circuitry is made to be more complicated than necessary by adding in a second control valve such that the logic controller is able to control each of the flows independently. As it turns out being able to tweak both flows, instead of just one, becomes a complicated mathematical problem and can result in weird and unexpected behavior in the face of large perturbations. In our system we similarly have two flows, the flow of reduced carbon, and the flow of an electron acceptor. The ratio of these two must be kept within some range and since in this case it is either not possible or very difficult to control the flow of the electron acceptor we find that the reduction flow is primarily controlled. To make a nearly perfect analogy, we have two sensors, Fnr that senses the levels of molecular oxygen and ArcA that senses the level of reduction potential flowing through the ETC and both then work to regulate the reduction flow in a potential evolutionary selection for a simple and robust control system. We also noticed that a similar phenomenon[188] that the same type of setup appears to occur between two electron transferring Geobacter species in which the species that acts as the electron acceptor is forced to change its genome whereas the species that acts as the electron donor is able to shift is reduction delivering capacity with the regulatory circuits it has already evolved.

While the ratio controller is certainly a part of the system in question we realized that our observed system is actually much more complicated and involves a number of other common control architectures. In fact of the commonly applied control architectures including feedforward with feedback trim, cascading, and PID controllers we can classify our system as a ratio-feedforward with feedback trim-cascading PID controller. Aside from the ratio the cascading control architecture is particularly relevant to our system. As developed in figure 3b the hierarchical nature of the system in question in which subsystems of each larger flow contain there own internal controls such that minor internal perturbations are buffered from the larger system. This is in fact a very commonly used control architecture in which systems that do not or should not respond on slow time scales are separated allowing for local systems to be controlled separately. This is beneficial because it allows each of the larger and smaller systems to focus on their own individual control problems rather than being forced to have an unnecessary or unrealistic

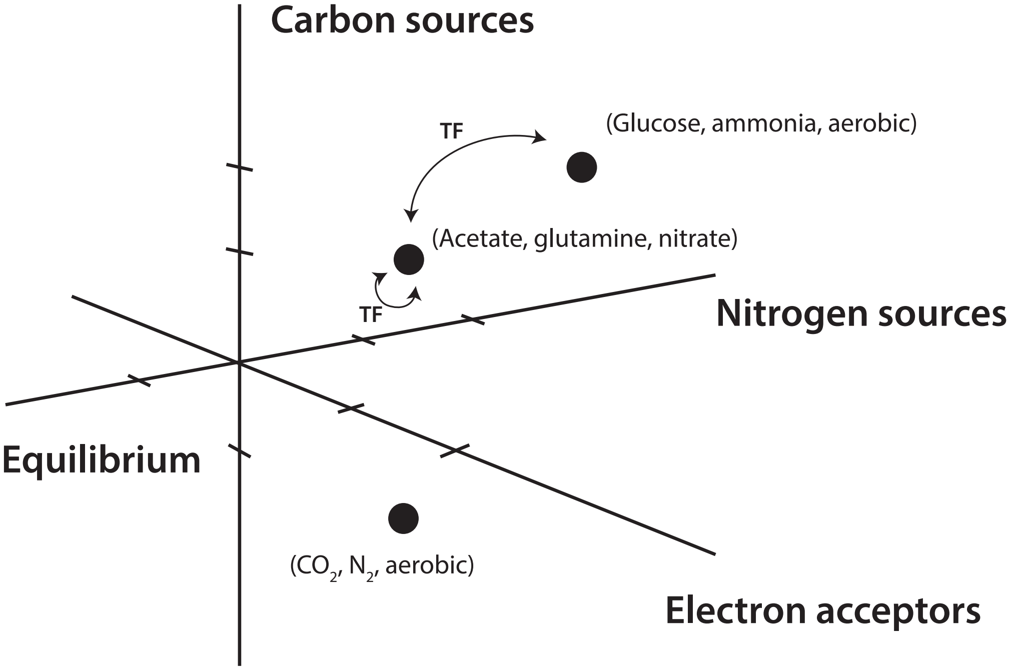


Figure 7.1: **Flow based model of the metabolic-regulatory network explains regulation throughout the anaerobic shift.** Throughout evolution bacteria have evolved two broad levels of transcription factors, transitional and homeostatic. Transitional regulators sense the external environment and cause an organism to undergo a large physiological shift whereas a homeostatic regulator will sense the internal environment and work to keep multiple cellular systems functioning around a steady state.

range of sensitivity and specificity as having one large system controlling the whole thing would force those individual components to be both incredibly sensitive and incredibly specific which is perhaps not possible. To draw on an another analogy we can look to the classic failure of a so called micro-manager who can enjoy great success in the early stages of an organization when there are only a few members but are soon swamped and inevitably inefficient or detrimental once an organization grows too large in size.

Another common class of controllers is comprised of those exhibiting primarily feedforward control while maintaining a feedback trim. This is in fact incredibly relevant to our current system as the classically observed points of control for ArcA and Fnr have always been the ETC and utilization of the redox ratio given that their effects are most dominant in this area. While the negative feedback is strong onto the catabolic circuits it can be likened to a trimming effect in that glycolysis and most of the TCA are not fully shut down but rather ratcheted down to lower the overall flow rate. In contrast elements of the ETC are fully shut down and completely replaced by other components. Also from an evolutionary perspective feedforward control represents the anticipatory response that is hardwired into every cell[**?**]. It is through this internal model that the cell has been able to continually adapt and survive in ever more ingenious ways.

In essence biology has evolved to use an incredibly intricate and robust system that certainly encapsulates and likely will far surpass our current linear and mathematical understandings of control. The fact global transcription factors regulate a large number of other transcription factors and that this forms a hierarchy is indisputable. However, the reason why global transcription factors are global is not because they are hubs in a network but rather because they are linked to the most high flux and central thermodynamic drivers of the cell. The fact that they are hubs of a regulatory network is a side effect of their fundamental regulatory role in attempting to maintain homeostasis or drive a state transition of the quantity in which they are sensing. The classically observed regulatory hierarchy is an outcome of the fundamental timescales and thermodynamic properties of the chemical reactions in a cell.

7.2 Alteration of the Minspan algorithm to provide a basis for genome-scale metabolic regulatory models

Usage of an altered version of the minspan algorithm[4] to develop a mathematical basis for computational models of metabolism coupled to transcriptional regulation. This work will be fully disclosed in my graduate thesis and has been previously described as part of my official advancement to candidacy. The invention includes determining the sparsest possible matrix N which exists as a null space of a stoichiometric metabolic network with the added criterion that each column vector of N contains reaction entries for at least one of the major flux routes ($> 5\%$ of sum total production or consumption flux) for one or more of the 12 primary biomass precursors or one or more of the 9 primary electron donors as defined in Fig 2. of Federowicz, et al. [86]. This altered null space matrix results in a set of pathway vectors that traverse the metabolic network and simultaneously adhere to the primary global dimensions of metabolic networks which exist as biomass production and energy production. This set of precursor included spanning vectors (bowspans) are then agglomeratively clustered with the distance metric as the angle of obliqueness which can be formalized as the cosine similarity. The resulting clustering will generate two primary clusters corresponding to the primary dimensions of biomass and energy production along with a hierarchical tree that reconstructs the metabolic network couplings such that a branch point in the hierarchical tree will correspond to one or more intermediate regulatory proteins, regulatory nucleic acids, or regulatory small molecules that carry out transcriptional, translational, or allosteric regulation on the enzymes which exist as part of the bowspans that comprise the portion of the

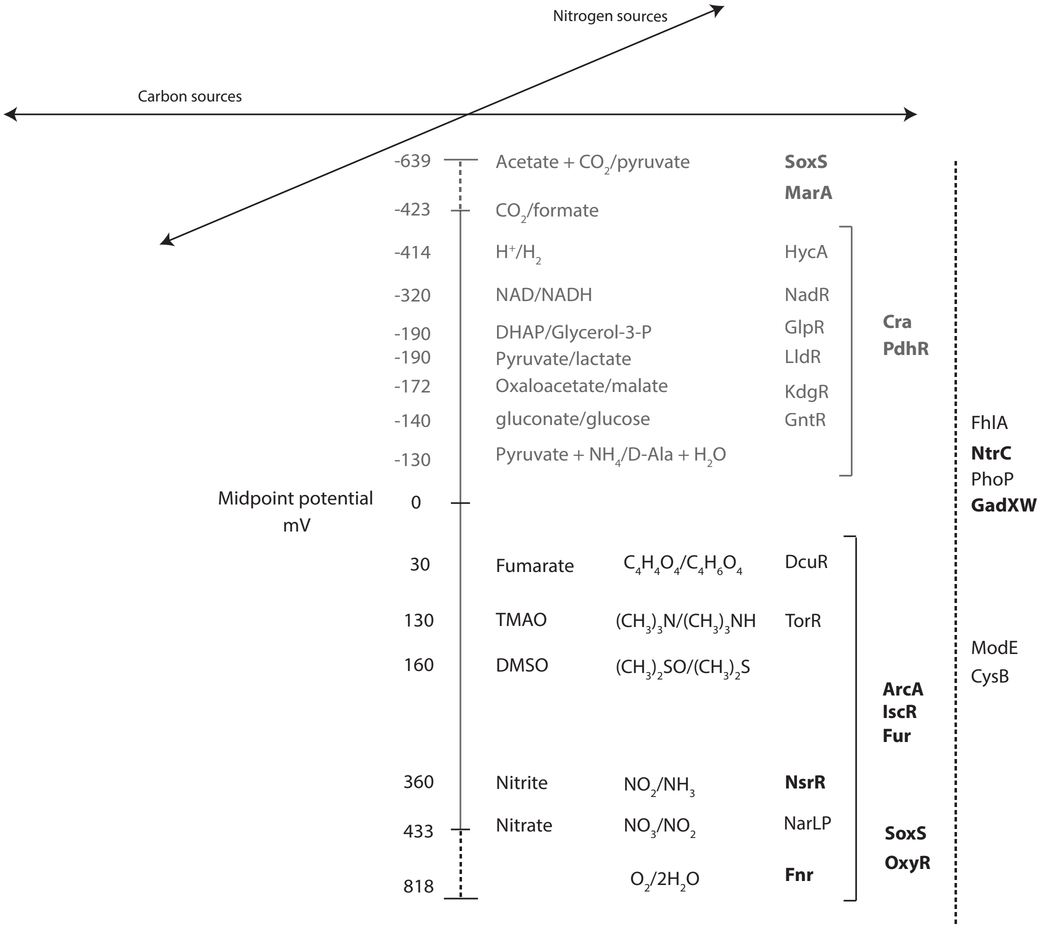


Figure 7.2: **Flow based model of the metabolic-regulatory network explains regulation throughout the anaerobic shift.** (A) Considering a mass balance around the ratio of reduced to oxidized molecules allows the unification of catabolism, anabolism, and chemiosmosis into a single process. The ratio of reduced to oxidized molecules is then sensed by ArcA and Fnr to elicit corresponding feedforward and feedback regulatory circuitry which allows the cell to maintain this critical ratio. (B) Mapping of the regulation of gene products (Figure 3) for each branch of the circuit reveals a broad trend of feedforward with feedback-trim regulation. Under fermentative conditions the redox ratio is high and the observed regulation works to lower the input and activate the output to bring the ratio down. Under nitrate respiration, the ratio drops and the circuit maintains a similar number of connections but is shown to decrease in gross activity levels.

hierarchical tree below a given branch point. This full hierarchical tree will thus represent the full hierarchy of the regulatory network of a cell. Once bowspans are calculated they can be used to create a process control model of the joint metabolic-regulatory network. This is done by developing a model in which feedforward or feedback regulatory terms or equations are added to a stoichiometric metabolic model such that the feedforward or feedback regulation acts on the reactions upstream or downstream of the metabolite which is being sensed and exists along the bowspan or set of bowspans. Often times the branch points will correspond directly to the metabolites being sensed and corresponding regulatory structure of the model. The key advance is that by knowing the hierarchy and the input/output relationships it will be possible to significantly reduce the dimensionality of the metabolic- regulatory model. This can result in a metabolic-regulatory model generated automatically based off of a stoichiometric metabolic reconstruction and parameterized on 20-30 terms corresponding to the global and intermediate regulators of a cell and their regulation response times following increase or decrease of the metabolite concentration which they sense along with the regulatory reconstruction of the genes or enzymes which they affect transcriptionally, translationally, or allosterically.

7.3 Using the linkage number of DNA as an equality constraint for transcription

The heightened resolution of ChIP-exo will allow for the observation of multiple binding sites occurring at the same promoter which reflect the 3-D topology of the genome. This occurs when dimers or tetramers of certain transcription factors form and result in multiple binding sites spread out by 100-200bp. Perhaps the best example comes from the E. coli malKp promoter in which two distally spaced Crp binding sites are replaced by one IHF binding site to recover transcriptional activation [189]. This is remarkable as crystal structures for both Crp and IHF show that a single Crp binding site will bend DNA at an angle of approximately 90 whereas an IHF site will bend DNA at an angle of 180. Thus two 90 bends can be replaced by a single 180 bend and the promoter is unaffected. We view this level of integration of structural information as an achievable and exciting avenue going forward. In particular we hope to be able to use the linkage number as an equality constraint in models of promoter or genome organization.

The linkage number (L) defined as L = T + W where T indicates the number of twists and W indicates the number of writhes that exist on a strand of DNA can be used as a global equality constraint. As a strict equality this constraint can be applied on top of genome-scale models of metabolism and macromolecular synthesis by constraining the solution space available for the rates of transcription initiation. The DNA double helix must be unwound or untwisted for transcription to initiate. Transcription factors are known to bind DNA, inducing bends or writhes in the strand which allow neighboring areas on the DNA strand to untwist and thus allow transcription to initiate. A global reconstruction of every possible transcription factor binding site on a strand of DNA coupled with structural knowledge indicating the degree of bending induced by each transcription factor can allow for a global constraint to be applied on top of joint models of metabolism, regulation, and macromolecular synthesis. This insight can also be applied to individual DNA plasmids or vectors in which transcription factor binding sites for transcription factors that have known structural bending profiles are engineered at precisely spaced distances to finely tune transcriptional rates.

Bibliography

[1] K. Kochanowski, U. Sauer, and V. Chubukov, “Somewhat in control—the role of transcription in regulating microbial metabolic fluxes,” *Current Opinion in Biotechnology*, vol. 24, pp. 987–993, Dec. 2013.

[2] N. B. Reppas, J. T. Wade, G. M. Church, and K. Struhl, “The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting.,” *Molecular Cell*, vol. 24, pp. 747–757, Dec. 2006.

[3] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp, “EcoCyc: fusing model organism databases with systems biology.,” *Nucleic acids research*, vol. 41, pp. D605–12, Jan. 2013.

[4] A. Bordbar, H. Nagarajan, N. E. Lewis, H. Latif, A. Ebrahim, S. Talon, J. Schellenberger, and B. Ø. Palsson, “Minimal metabolic pathway structure is consistent with associated biomolecular interactions,” *Molecular systems biology*, vol. 10, July 2014.

[5] N. E. Lewis, H. Nagarajan, and B. Ø. Palsson, “Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods,” *Nature Reviews Microbiology*, vol. 10, pp. 291–305, Apr. 2012.

[6] C. You, H. Okano, S. Hui, Z. Zhang, M. Kim, C. W. Gunderson, Y.-P. Wang, P. Lenz, D. Yan, and T. Hwa, “Coordination of bacterial proteome with metabolism by cyclic AMP signalling,” *Nature*, vol. 500, pp. 301–306, Aug. 2013.

[7] M. Kim, Z. Zhang, H. Okano, and D. Yan, “Need-based activation of ammonium uptake in Escherichia coli,” *Molecular systems . . .* , 2012.

[8] J. Monk, J. Nogales, and B. O. Palsson, “Optimizing genome-scale network reconstructions.,” *Nature biotechnology*, vol. 32, pp. 447–452, May 2014.

[9] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson, “Constraint-based models predict metabolic and associated cellular functions.,” *Nature reviews. Genetics*, vol. 15, pp. 107–120, Feb. 2014.

[10] M. A. Oberhardt, B. Ø. Palsson, and J. A. Papin, “Applications of genome-scale metabolic reconstructions.,” *Molecular systems biology*, vol. 5, p. 320, 2009.

[11] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson, “Reconstruction of biochemical networks in microorganisms.,” *Nature reviews. Microbiology*, vol. 7, pp. 129–143, Feb. 2009.

[12] J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, N. E. Lewis, J. D. Orth, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, K. Zengler, and B. Ø. Palsson, “In silico method for modelling metabolism and gene product expression at genome scale,” *Nature Communications*, vol. 3, pp. 929–, July 2012.

[13] M. W. Covert, C. H. Schilling, and B. Palsson, “Regulation of gene expression in flux balance models of metabolism.,” *Journal of theoretical biology*, vol. 213, pp. 73–88, Nov. 2001.

[14] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgård, and B. Ø. Palsson, “Integrating high-throughput and computational data elucidates bacterial networks.,” *Nature*, vol. 429, pp. 92–96, May 2004.

[15] M. J. Herrgård, M. W. Covert, and B. Ø. Palsson, “Reconstruction of microbial transcriptional regulatory networks.,” *Current opinion in biotechnology*, vol. 15, pp. 70–77, Feb. 2004.

[16] C. L. Barrett, C. D. Herring, J. L. Reed, and B. O. Palsson, “The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 19103–19108, Dec. 2005.

[17] C. L. Barrett and B. O. Palsson, “Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach.,” *PLoS computational biology*, vol. 2, p. e52, May 2006.

[18] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, “A genome-scale computational study of the interplay between transcriptional regulation and metabolism.,” in *Molecular systems biology*, p. 101, School of Computer Science, Tel Aviv University, Tel Aviv, Israel. shlomito@post.tau.ac.il, 2007.

[19] E. P. Gianchandani, A. R. Joyce, B. Ø. Palsson, and J. A. Papin, “Functional states of the genome-scale Escherichia coli transcriptional regulatory system.,” *PLoS computational biology*, vol. 5, p. e1000403, June 2009.

[20] S. Alexeeva, K. J. Hellingwerf, and M. J. Teixeira de Mattos, “Requirement of ArcA for redox regulation in Escherichia coli under microaerobic but not anaerobic or aerobic conditions.,” *Journal of bacteriology*, vol. 185, pp. 204–209, Jan. 2003.

[21] R. J. Rolfes and H. Zalkin, “Escherichia coli gene purR encoding a repressor protein for purine nucleotide synthesis. Cloning, nucleotide sequence, and interaction with the purF operator.,” *The Journal of biological chemistry*, vol. 263, pp. 19653–19661, Dec. 1988.

[22] U. Houlberg and K. F. Jensen, “Role of hypoxanthine and guanine in regulation of Salmonella typhimurium pur gene expression.,” *Journal of bacteriology*, vol. 153, pp. 837–845, Feb. 1983.

[23] H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and A.-P. Zeng, “An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs.,” *Nucleic acids research*, vol. 32, no. 22, pp. 6643–6649, 2004.

[24] A. S. N. Seshasayee, G. M. Fraser, M. M. Babu, and N. M. Luscombe, “Principles of transcriptional regulation and evolution of the metabolic system in E. coli.,” *Genome Research*, vol. 19, pp. 79–91, Jan. 2009.

[25] A. Martínez-Antonio and J. Collado-Vides, “Identifying global regulators in transcriptional regulatory networks in bacteria.,” *Current opinion in microbiology*, vol. 6, pp. 482–489, Oct. 2003.

[26] H. T. C W Tabor, “Polyamines in microorganisms.,” *Microbiological reviews*, vol. 49, p. 81, Mar. 1985.

[27] Y. Jin, R. M. Watt, A. Danchin, and J.-d. Huang, “Small noncoding RNA GcvB is a novel regulator of acid resistance in Escherichia coli,” *BMC Genomics*, vol. 10, p. 165, Apr. 2009.

[28] R. Balbontín, F. Fiorini, N. Figueroa Bossi, J. Casadesús, and L. Bossi, “Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica,” *Molecular microbiology*, vol. 78, pp. 380–394, Oct. 2010.

[29] K. Y. Choi and H. Zalkin, “Regulation of Escherichia coli pyrC by the purine regulon repressor protein.,” *Journal of bacteriology*, vol. 172, pp. 3201–3207, June 1990.

[30] H. R. Wilson and C. L. Turnbough, “Role of the purine repressor in the regulation of pyrimidine gene expression in Escherichia coli K-12.,” *Journal of bacteriology*, vol. 172, pp. 3208–3213, June 1990.

[31] J. N. Larsen and K. F. Jensen, “Nucleotide sequence of the pyrD gene of Escherichia coli and characterization of the flavoprotein dihydroorotate dehydrogenase.,” *European journal of biochemistry / FEBS*, vol. 151, pp. 59–65, Aug. 1985.

[32] N. Devroede, T.-L. Thia-Toong, D. Gigot, D. Maes, and D. Charlier, “Purine and pyrimidine-specific repression of the Escherichia coli carAB operon are functionally and structurally coupled.,” *Journal of molecular biology*, vol. 336, pp. 25–42, Feb. 2004.

[33] J. Piette, H. Nyunoya, C. J. Lusty, R. Cunin, G. Weyens, M. Crabeel, D. Charlier, N. Glansdorff, and A. Piérard, “DNA sequence of the carA gene and the control region of carAB: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in Escherichia coli K-12,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, pp. 4134–4138, July 1984.

[34] K. Burton, “Adenine Transport in Escherichia coli,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 255, pp. 153–157, Feb. 1994.

[35] J. A. Opdyke, J.-G. Kang, and G. Storz, “GadY, a small-RNA regulator of acid response genes in Escherichia coli.,” *Journal of bacteriology*, vol. 186, pp. 6698–6705, Oct. 2004.

[36] B.-K. Cho, K. Zengler, Y. Qiu, Y.-S. Park, E. M. Knight, C. L. Barrett, Y. Gao, and B. Ø. Palsson, “The transcription unit architecture of the Escherichia coli genome.,” *Nature Biotechnology*, vol. 27, pp. 1043–1049, Nov. 2009.

[37] B. K. Cho, C. L. Barrett, E. M. Knight, Y. S. Park, and B. O. Palsson, “Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli,” *Proceedings of the National Academy of Sciences*, vol. 105, pp. 19462–19467, Dec. 2008.

[38] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, “MEME SUITE: tools for motif discovery and searching,” *Nucleic acids research*, vol. 37, pp. W202–W208, June 2009.

[39] B. He and H. Zalkin, “Repression of Escherichia coli purB is by a transcriptional roadblock mechanism.,” *Journal of bacteriology*, vol. 174, pp. 7121–7127, Nov. 1992.

[40] P. Karatza and S. Frillingos, “Cloning and functional characterization of two bacterial members of the NAT/NCS2 family in Escherichia coli,” *dx.doi.org*, vol. 22, pp. 251–261, July 2009.

[41] B.-K. Cho, E. M. Knight, C. L. Barrett, and B. Ø. Palsson, “Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts,” *Genome Research*, vol. 18, pp. 900–910, June 2008.

[42] T. Shimada, A. Ishihama, S. J. W. Busby, and D. C. Grainger, “The Escherichia coli RutR transcription factor binds at targets within genes as well as intergenic regions.,” *Nucleic acids research*, vol. 36, pp. 3950–3955, July 2008.

[43] S. Krishna, S. Semsey, and K. Sneppen, “Combinatorics of feedback in cellular uptake and metabolism of small molecules.,” *Proceedings of the National Academy of Sciences*, vol. 104, pp. 20815–20819, Dec. 2007.

[44] D. C. Grainger, H. Aiba, D. Hurd, D. F. Browning, and S. J. W. Busby, “Transcription factor distribution in Escherichia coli: studies with FNR protein.,” *Nucleic acids research*, vol. 35, no. 1, pp. 269–278, 2007.

[45] D. Grainger, D. Hurd, and M. Harrison, “Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome,” in *Proceedings of the ...*, 2005.

[46] B.-K. Cho, E. M. Knight, and B. Ø. Palsson, “PCR-based tandem epitope tagging system for Escherichia coli genome engineering.,” tech. rep., University of California at San Diego, La Jolla, CA 92093, USA., Jan. 2006.

[47] K. A. Datsenko and B. L. Wanner, “One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 6640–6645, June 2000.

[48] A. R. Joyce and B. Ø. Palsson, “The model organism as a system: integrating ’omics’ data sets.,” *Nature Reviews Molecular Cell Biology*, vol. 7, pp. 198–210, Mar. 2006.

[49] U. Alon, “Network motifs: theory and experimental approaches,” *Nature Reviews Genetics*, vol. 8, pp. 450–461, June 2007.

[50] D. Charlier, M. Roovers, F. Van Vliet, A. Boyen, R. Cunin, Y. Nakamura, N. Glansdorff, and A. Piérard, “Arginine regulon of Escherichia coli K-12. A study of repressor-operator interactions and of in vitro binding affinities versus in vivo repression.,” *Journal of molecular biology*, vol. 226, pp. 367–386, July 1992.

[51] J. Yang, A. Gunasekera, T. A. Lavoie, L. Jin, D. E. Lewis, and J. Carey, “In vivo and in vitro studies of TrpR-DNA interactions.,” *Journal of molecular biology*, vol. 258, pp. 37–52, Apr. 1996.

[52] J. Pittard, H. Camakaris, and J. Yang, “The TyrR regulon,” *Molecular microbiology*, vol. 55, pp. 16–26, Jan. 2005.

[53] T. H. Tani, A. Khodursky, R. M. Blumenthal, P. O. Brown, and R. G. Matthews, “Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis.,” vol. 99, pp. 13471–13476, Oct. 2002.

[54] J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev, “Comprehensive comparative analysis of strand-specific RNA sequencing methods.,” *Nature methods*, vol. 7, pp. 709–715, Sept. 2010.

[55] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.,” *PLoS Biology*, vol. 5, p. e8, Jan. 2007.

[56] J. M. Calvo and R. G. Matthews, “The leucine-responsive regulatory protein, a global regulator of metabolism in Escherichia coli.,” *Microbiological reviews*, vol. 58, pp. 466–490, Sept. 1994.

[57] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides, “RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units),” *Nucleic acids research*, vol. 39, pp. D98–D105, Dec. 2010.

[58] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp, “EcoCyc: a comprehensive database of Escherichia coli biology,” *Nucleic acids research*, vol. 39, pp. D583–D590, Dec. 2010.

[59] M. Caldara, D. Charlier, and R. Cunin, “The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation.,” *Microbiology (Reading, England)*, vol. 152, pp. 3343–3354, Nov. 2006.

[60] M. Caldara, P. N. L. Minh, S. Bostoen, J. Massant, and D. Charlier, “ArgR-dependent repression of arginine and histidine transport genes in Escherichia coli K-12.,” *Journal of molecular biology*, vol. 373, pp. 251–267, Oct. 2007.

[61] M. Jeeves, P. D. Evans, R. A. Parslow, M. Jaseja, and E. I. Hyde, “Studies of the Escherichia coli Trp repressor binding to its five operators and to variant operator

sequences.," *European journal of biochemistry / FEBS*, vol. 265, pp. 919–928, Nov. 1999.

[62] R. G. Zhang, A. Joachimiak, C. L. Lawson, R. W. Schevitz, Z. Otwinowski, and P. B. Sigler, "The crystal structure of trp aporepressor at 1.8 A shows how binding tryptophan enhances DNA affinity.," *Nature*, vol. 327, pp. 591–597, June 1987.

[63] W. K. MAAS, "STUDIES ON THE MECHANISM OF REPRESSION OF ARGININE BIOSYNTHESIS IN ESCHERICHIA COLI. II. DOMINANCE OF REPRESSIBILITY IN DIPLOIDS.," *Journal of molecular biology*, vol. 8, pp. 365–370, Mar. 1964.

[64] A. K. Kiupakis and L. Reitzer, "ArgR-independent induction and ArgR-dependent superinduction of the astCADBE operon in Escherichia coli.," *Journal of bacteriology*, vol. 184, pp. 2940–2950, June 2002.

[65] A. B. Khodursky, B. J. Peter, N. R. Cozzarelli, D. Botstein, P. O. Brown, and C. Yanofsky, "DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli.," vol. 97, pp. 12170–12175, Oct. 2000.

[66] B. D. Bennett, E. H. Kimball, M. Gao, R. Osterhout, S. J. Van Dien, and J. D. Rabinowitz, "Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli," *Nature chemical biology*, vol. 5, pp. 593–599, June 2009.

[67] B.-K. Cho, S. A. Federowicz, M. Embree, Y.-S. Park, D. Kim, and B. Ø. Palsson, "The PurR regulon in Escherichia coli K-12 MG1655.," *Nucleic acids research*, vol. 39, pp. 6456–6464, Aug. 2011.

[68] S. Semsey, A. M. C. Andersson, S. Krishna, M. H. Jensen, E. Massé, and K. Sneppen, "Genetic regulation of fluxes: iron homeostasis of Escherichia coli.," *Nucleic acids research*, vol. 34, no. 17, pp. 4960–4967, 2006.

[69] S. Semsey, S. Krishna, K. Sneppen, and S. Adhya, "Signal integration in the galactose network of Escherichia coli.," *Molecular microbiology*, vol. 65, pp. 465–476, July 2007.

[70] L. S. Klig, D. L. Oxender, and C. Yanofsky, "Second-site revertants of Escherichia coli trp repressor mutants.," *Genetics*, vol. 120, pp. 651–655, Nov. 1988.

[71] B.-K. Cho, E. M. Knight, and B. Ø. Palsson, "Genomewide identification of protein binding locations using chromatin immunoprecipitation coupled with microarray.," vol. 439, no. Chapter 9, pp. 131–145, 2008.

[72] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society Series B . . .* , 1995.

[73] J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.,” *Genome Biol*, vol. 11, no. 8, p. R86, 2010.

[74] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, “Cytoscape 2.8: new features for data integration and network visualization.,” *Bioinformatics*, vol. 27, pp. 431–432, Feb. 2011.

[75] B. D. O’Connor, B. Merriman, and S. F. Nelson, “SeqWare Query Engine: storing and searching sequence data in the cloud,” *BMC Bioinformatics*, vol. 11, no. Suppl 12, p. S2, 2010.

[76] B. Palsson and K. Zengler, “The challenges of integrating multi-omic data sets,” *Nature chemical biology*, vol. 6, pp. 787–789, Nov. 2010.

[77] S. G. Thorleifsson and I. Thiele, “rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks.,” *Bioinformatics*, vol. 27, pp. 2009–2010, July 2011.

[78] M. Ganter, T. Bernard, S. Moretti, J. Stelling, and M. Pagni, “MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks.,” *Bioinformatics*, vol. 29, pp. 815–816, Mar. 2013.

[79] A. Kumar, P. F. Suthers, and C. D. Maranas, “MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases,” *BMC Bioinformatics*, vol. 13, no. 1, p. 6, 2012.

[80] S. Pabinger, R. Rader, R. Agren, J. Nielsen, and Z. Trajanoski, “MEMOSys: Bioinformatics platform for genome-scale metabolic models,” *BMC Systems Biology*, vol. 5, p. 20, Jan. 2011.

[81] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, “BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions,” *BMC Bioinformatics*, vol. 11, p. 213, Apr. 2010.

[82] A. Ebrahim, J. A. Lerman, and B. O. Palsson, “COBRApy: COnstraints-Based Reconstruction and Analysis for Python.,” *BMC systems . . .* , 2013.

[83] H. Rohn, A. Junker, A. Hartmann, E. Grafahrend-Belau, H. Treutler, M. Klapperstück, T. Czauderna, C. Klukas, and F. Schreiber, “VANTED v2: a framework for systems biology applications.,” *BMC Systems Biology*, vol. 6, p. 139, 2012.

[84] J. R. Karr, J. C. Sanghvi, D. N. Macklin, A. Arora, and M. W. Covert, "WholeCellKB: model organism databases for comprehensive whole-cell models.," *Nucleic acids research*, vol. 41, pp. D787–92, Jan. 2013.

[85] E. J. O middle dot Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. O. Palsson, "Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction," *Molecular systems biology*, vol. 9, pp. 693–693, Jan. 2013.

[86] S. Federowicz, D. Kim, A. Ebrahim, and J. Lerman, "Determining the Control Circuitry of Redox Metabolism at the Genome-Scale," *PLoS genetics*, 2014.

[87] B.-K. Cho, S. Talon, Y.-S. Park, K. Zengler, and B. Ø. Palsson, "Deciphering the transcriptional regulatory logic of amino acid metabolism," *Nature chemical biology*, vol. 8, pp. 65–71, Jan. 2012.

[88] S. A. Federowicz, M. Embree, and Y. S. Park, "The PurR regulon in Escherichia coli K-12 MG1655," *Nucleic acids . . .* , 2011.

[89] M. Hucka and F. Bergmann, "The systems biology markup language (SBML): language specification for level 3 version," *Nature*, 2010.

[90] A. Funahashi, Y. Matsuoka, and A. Jouraku, "CellDesigner 3.5: a versatile modeling tool for biochemical networks," in *Proceedings of the . . .* , 2008.

[91] A. Dräger, N. Rodriguez, M. Dumousseau, A. Dörr, C. Wrzodek, N. Le Novère, A. Zell, and M. Hucka, "JSBML: a flexible Java library for working with SBML.," *Bioinformatics*, vol. 27, pp. 2167–2168, Aug. 2011.

[92] M. Kanehisa, S. Goto, Y. Sato, and M. Furumichi, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic acids . . .* , 2011.

[93] D. Croft, A. F. Mundo, R. Haw, and M. Milacic, "The Reactome pathway knowledgebase," *Nucleic acids . . .* , 2014.

[94] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.," *Nucleic acids research*, vol. 42, pp. D459–71, Jan. 2014.

[95] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S.-A. Sansone, "ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.," *Bioinformatics*, vol. 26, pp. 2354–2356, Sept. 2010.

[96] M. L. Mo, B. Ø. Palsson, and M. J. Herrgård, "Connecting extracellular metabolomic measurements to intracellular flux states in yeast," *BMC Systems Biology*, vol. 3, no. 1, p. 37, 2009.

[97] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson, "A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011," *Molecular systems biology*, vol. 7, Oct. 2011.

[98] H. Hegyi and M. Gerstein, "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.," *Journal of molecular biology*, vol. 288, pp. 147–164, Apr. 1999.

[99] J. S. Fraser, J. D. Gross, and N. J. Krogan, "From systems to structure: bridging networks and mechanism," *Molecular cell*, 2013.

[100] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, and M. A. Singer, "A high-resolution map of active promoters in the human genome," *Nature*, 2005.

[101] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, and E. Lieberman, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, 2007.

[102] H. S. Rhee and B. F. Pugh, "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.," *Cell*, vol. 147, pp. 1408–1419, Dec. 2011.

[103] S. Busby and R. H. Ebright, "Transcription activation by catabolite activator protein (CAP).," *Journal of molecular biology*, vol. 293, pp. 199–213, Oct. 1999.

[104] G. Gosset, Z. Zhang, S. Nayyar, and W. Cuevas, "Transcriptome analysis of Crp-dependent catabolite control of gene expression in Escherichia coli," *Journal of* . . . , 2004.

[105] K. C. Kao, L. M. Tran, and J. C. Liao, "A global regulatory role of gluconeogenic genes in Escherichia coli revealed by transcriptome network analysis.," *The Journal of biological chemistry*, vol. 280, pp. 36079–36087, Oct. 2005.

[106] D. Zheng, C. Constantinidou, J. L. Hobman, and S. D. Minchin, "Identification of the CRP regulon using in vitro and in vivo transcriptional profiling," *Nucleic acids research*, vol. 32, pp. 5874–5893, Jan. 2004.

[107] T. Shimada, N. Fujita, K. Yamamoto, and A. Ishihama, "Novel Roles of cAMP Receptor Protein (CRP) in Regulation of Transport and Metabolism of Carbon Sources," *PLoS One*, vol. 6, p. e20081, June 2011.

[108] J. Germer, G. Becker, M. Metzner, and R. Hengge-Aronis, "Role of activator site position and a distal UP-element half-site for sigma factor selectivity at a CRP/H-NS-activated sigma(s)-dependent promoter in Escherichia coli.," *Molecular microbiology*, vol. 41, pp. 705–716, Aug. 2001.

[109] H. H. Kristensen, P. Valentin-Hansen, and L. Søgaard-Andersen, "Design of CytR regulated, cAMP-CRP dependent class II promoters in Escherichia coli: RNA polymerase-promoter interactions modulate the efficiency of CytR repression.," *Journal of molecular biology*, vol. 266, pp. 866–876, Mar. 1997.

[110] C. L. Lawson, D. Swigon, K. S. Murakami, and S. A. Darst, "Catabolite activator protein: DNA binding and transcription activation," *Current opinion in . . .* , 2004.

[111] H. Pedersen, J. Dall, and G. Dandanell, "Gene-regulatory modules In Escherichia coli: nucleoprotein complexes formed by cAMP-CRP and CytR at the nupG promoter," *Molecular . . .* , 1995.

[112] R. M. Williams, V. A. Rhodius, A. I. Bell, and A. Kolb, "Orientation of functional activating regions in the Escherichia coli CRP protein during transcription activation at class II promoters," *Nucleic acids . . .* , 1996.

[113] B.-K. Cho, D. Kim, E. M. Knight, K. Zengler, and B. Ø. Palsson, "Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states.," *BMC biology*, vol. 12, no. 1, p. 4, 2014.

[114] C. D. Herring, M. Raffaelle, T. E. Allen, E. I. Kanin, R. Landick, A. Z. Ansari, and B. Ø. Palsson, "Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays.," *Journal of bacteriology*, vol. 187, pp. 6166–6174, Sept. 2005.

[115] C. Kröger, S. C. Dillon, A. D. S. Cameron, K. Papenfort, S. K. Sivasankaran, K. Hokamp, Y. Chao, A. Sittka, M. Hébrard, K. Händler, A. Colgan, P. Leekitcharoenphon, G. C. Langridge, A. J. Lohan, B. Loftus, S. Lucchini, D. W. Ussery, C. J. Dorman, N. R. Thomson, J. Vogel, and J. C. D. Hinton, "The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium.," *Proceedings of the National Academy of Sciences*, vol. 109, pp. E1277–86, May 2012.

[116] Y. Qiu, B.-K. Cho, Y.-S. Park, D. Lovley, B. Ø. Palsson, and K. Zengler, "Structural and operational complexity of the Geobacter sulfurreducens genome.," *Genome Research*, vol. 20, pp. 1304–1311, Sept. 2010.

[117] J. T. Wade, K. Struhl, S. J. W. Busby, and D. C. Grainger, "Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization.," *Molecular microbiology*, vol. 65, pp. 21–26, July 2007.

[118] J. S. Mymryk and T. K. Archer, "Detection of transcription factor binding in vivo using lambda exonuclease.," *Nucleic acids research*, vol. 22, pp. 4344–4345, Oct. 1994.

[119] I. G. Hook-Barnard and D. M. Hinton, “Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters.,” *Gene regulation and systems biology*, vol. 1, pp. 275–293, 2007.

[120] A. Rogozina, E. Zaychikov, M. Buckle, H. Heumann, and B. Sclavi, “DNA melting by RNA polymerase at the T7A1 promoter precedes the rate-limiting step at 37 degrees C and results in the accumulation of an off-pathway intermediate.,” *Nucleic acids research*, vol. 37, pp. 5390–5404, Sept. 2009.

[121] P. Schickor, W. Metzger, W. Werel, H. Lederer, and H. Heumann, “Topography of intermediates in transcription initiation of E.coli.,” *The EMBO journal*, vol. 9, pp. 2215–2220, July 1990.

[122] B. Sclavi, E. Zaychikov, A. Rogozina, F. Walther, M. Buckle, and H. Heumann, “Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 4706–4711, Mar. 2005.

[123] R. T. Kovacic, “The 0 degree C closed complexes between Escherichia coli RNA polymerase and two promoters, T7-A3 and lacUV5.,” *The Journal of biological chemistry*, vol. 262, pp. 13654–13661, Oct. 1987.

[124] C. A. Davis, C. A. Bingman, R. Landick, M. T. Record, and R. M. Saecker, “Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by Escherichia coli RNA polymerase.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 7833–7838, May 2007.

[125] V. M. Cook and P. L. Dehaseth, “Strand opening-deficient Escherichia coli RNA polymerase facilitates investigation of closed complexes with promoter DNA: effects of DNA sequence and temperature.,” *The Journal of biological chemistry*, vol. 282, pp. 21319–21326, July 2007.

[126] S. Borukhov, V. Sagitov, C. A. Josaitis, R. L. Gourse, and A. Goldfarb, “Two modes of transcription initiation in vitro at the rrnB P1 promoter of Escherichia coli.,” *The Journal of biological chemistry*, vol. 268, pp. 23477–23482, Nov. 1993.

[127] R. L. Gourse, “Visualization and quantitative analysis of complex formation between E. coli RNA polymerase and an rRNA promoter in vitro.,” *Nucleic acids research*, vol. 16, pp. 9789–9809, Oct. 1988.

[128] J. T. Newlands, W. Ross, K. K. Gosink, and R. L. Gourse, “Factor-independent activation of Escherichia coli rRNA transcription. II. characterization of complexes of rrnB P1 promoters containing or lacking the upstream activator region with Escherichia coli RNA polymerase.,” *Journal of molecular biology*, vol. 220, pp. 569–583, Aug. 1991.

[129] A. Spassky, K. Kirkegaard, and H. Buc, “Changes in the DNA structure of the lac UV5 promoter during formation of an open complex with Escherichia coli RNA polymerase.,” *Biochemistry*, vol. 24, pp. 2723–2731, May 1985.

[130] B. Krummel and M. J. Chamberlin, “RNA chain initiation by Escherichia coli RNA polymerase. Structural transitions of the enzyme in early ternary complexes.,” *Biochemistry*, vol. 28, pp. 7829–7842, Sept. 1989.

[131] B. Krummel and M. J. Chamberlin, “Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes.,” *Journal of molecular biology*, vol. 225, pp. 239–250, May 1992.

[132] D. Crothers, “Catabolite activator protein-induced DNA bending in transcription initiation 10.1016/0022-2836(91)90562-K : Journal of Molecular Biology — ScienceDirect.com,” *Journal of molecular biology*, 1991.

[133] D. C. Straney and D. M. Crothers, “A stressed intermediate in the formation of stably initiated RNA chains at the Escherichia coli lac UV5 promoter.,” *Journal of molecular biology*, vol. 193, pp. 267–278, Jan. 1987.

[134] E. Zhilina, D. Esyunina, K. Brodolin, and A. Kulbachinskiy, “Structural transitions in the transcription elongation complexes of bacterial RNA polymerase during σ-dependent pausing.,” *Nucleic acids research*, vol. 40, pp. 3078–3091, Apr. 2012.

[135] W. Metzger, P. Schickor, and H. Heumann, “A cinematographic view of Escherichia coli RNA polymerase translocation.,” *The EMBO journal*, vol. 8, pp. 2745–2754, Sept. 1989.

[136] A. J. Carpousis and J. D. Gralla, “Interaction of RNA polymerase with lacUV5 promoter DNA during mRNA initiation and elongation. Footprinting, methylation, and rifampicin-sensitivity changes accompanying transcription initiation.,” *Journal of molecular biology*, vol. 183, pp. 165–177, May 1985.

[137] D. F. Browning and S. J. W. Busby, “The regulation of bacterial transcription initiation,” *Nature Reviews Microbiology*, vol. 2, pp. 57–65, Jan. 2004.

[138] P. Valentin-Hansen, “Tandem CRP binding sites in the deo operon of Escherichia coli K-12.,” *The EMBO journal*, vol. 1, no. 9, pp. 1049–1054, 1982.

[139] E. A. Campbell, N. Korzheva, A. Mustaev, K. Murakami, S. Nair, A. Goldfarb, and S. A. Darst, “Structural mechanism for rifampicin inhibition of bacterial rna polymerase.,” *Cell*, vol. 104, pp. 901–912, Mar. 2001.

[140] A. A. Serandour, G. D. Brown, J. D. Cohen, and J. S. Carroll, “Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties.,” *Genome Biol*, vol. 14, no. 12, p. R147, 2013.

[141] R. Gourse and W. Ross, “UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition,” *Molecular microbiology*, 2000.

[142] D. West, R. Williams, V. Rhodius, A. Bell, N. Sharma, C. Zou, N. Fujita, A. Ishihama, and S. Busby, “Interactions between the Escherichia coli cyclic AMP receptor protein and RNA polymerase at class II promoters.,” *Molecular microbiology*, vol. 10, pp. 789–797, Nov. 1993.

[143] V. A. Rhodius, D. M. West, C. L. Webster, S. J. Busby, and N. J. Savery, “Transcription activation at class II CRP-dependent promoters: the role of different activating regions.,” *Nucleic acids research*, vol. 25, pp. 326–332, Jan. 1997.

[144] H. Salgado, M. Peralta-Gil, and S. Gama-Castro, “RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more,” *Nucleic acids . . .* , 2013.

[145] R. Carlson and F. Srienc, “Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: Identification of reactions,” *Biotechnology and bioengineering*, vol. 85, no. 1, pp. 1–19, 2003.

[146] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer, “Multidimensional Optimality of Microbial Metabolism,” *Science*, vol. 336, pp. 597–601, May 2012.

[147] R. Rabus, J. Reizer, I. Paulsen, and M. H. Saier, “Enzyme I(Ntr) from Escherichia coli. A novel enzyme of the phosphoenolpyruvate-dependent phosphotransferase system exhibiting strict specificity for its phosphoryl acceptor, NPr.,” *The Journal of biological chemistry*, vol. 274, pp. 26185–26191, Sept. 1999.

[148] A. Weijland, K. Harmark, R. H. Cool, P. H. Anborgh, and A. Parmeggiani, “Elongation factor Tu: a molecular switch in protein biosynthesis.,” *Molecular microbiology*, vol. 6, pp. 683–688, Mar. 1992.

[149] Y. Guo, S. Mahony, and D. K. Gifford, “High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints.,” *PLoS Computational Biology*, vol. 8, no. 8, p. e1002638, 2012.

[150] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nature methods*, vol. 9, pp. 357–359, Mar. 2012.

[151] J. Green and M. S. Paget, “Bacterial redox sensors,” *Nature Reviews Microbiology*, vol. 2, pp. 954–966, Dec. 2004.

[152] S. Iuchi and E. C. Lin, “Adaptation of Escherichia coli to respiratory conditions: regulation of gene expression.,” *Cell*, vol. 66, no. 1, pp. 5–7, 1991.

[153] M. D. Rolfe, A. T. Beek, A. I. Graham, E. W. Trotter, H. M. S. Asif, G. Sanguinetti, J. T. de Mattos, R. K. Poole, and J. Green, "Transcript Profiling and Inference of *Escherichia coli* K-12 ArcA Activity across the Range of Physiologically Relevant Oxygen Concentrations," *Journal of Biological Chemistry*, vol. 286, pp. 10147–10154, Mar. 2011.

[154] G. Unden and P. Dünnwald, "The aerobic and anaerobic respiratory chain of Escherichia coli and Salmonella enterica: enzymes and energetics," *EcoSal-Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2008.

[155] C. Constantinidou, J. Hobman, L. Griffiths, M. Patel, C. Penn, J. Cole, and T. Overton, "A reassessment of the FNR regulon and transcriptomic analysis of the effects of nitrate, nitrite, NarXL, and NarQP as *Escherichia coli* K12 adapts from aerobic to . . . ," *The Journal of biological chemistry*, vol. 281, pp. 4802–4815, Feb. 2006.

[156] E. W. Trotter, M. D. Rolfe, A. M. Hounslow, C. J. Craven, M. P. Williamson, G. Sanguinetti, R. K. Poole, and J. Green, "Reprogramming of Escherichia coli K-12 Metabolism during the Initial Phase of Transition from an Anaerobic to a Micro-Aerobic Environment," vol. 6, p. e25501, Sept. 2011.

[157] S. Iuchi and E. C. C. Lin, "Adaptation of Escherichia coli to redox environments by gene expression," *Molecular microbiology*, vol. 9, pp. 9–15, July 1993.

[158] S. Iuchi and L. Weiner, "Cellular and molecular physiology of *Escherichia coli* in the adaptation to aerobic environments.," *Journal of biochemistry*, vol. 120, no. 6, pp. 1055–1063, 1996.

[159] S. Shalel Levanon, K.-Y. San, and G. N. Bennett, "Effect of oxygen on the *Escherichia coli* ArcA and FNR regulation systems and metabolic responses," *Biotechnology and bioengineering*, vol. 89, no. 5, pp. 556–564, 2005.

[160] J. D. Partridge, C. Scott, Y. Tang, R. K. Poole, and J. Green, "Escherichia coli transcriptome dynamics during the transition from anaerobic to aerobic conditions.," *The Journal of biological chemistry*, vol. 281, pp. 27806–27815, Sept. 2006.

[161] J. Green and M. S. Paget, "Bacterial redox sensors," *Nat Rev Micro*, vol. 2, pp. 954–966, Dec. 2004.

[162] R. Malpica, B. Franco, C. Rodriguez, O. Kwon, and D. Georgellis, "Identification of a quinone-sensitive redox switch in the ArcB sensor kinase," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 13318–13323, Aug. 2004.

[163] P. Kiley and H. Beinert, "Oxygen sensing by the global regulator, FNR: the role of the iron-sulfur cluster,"

[164] G. Unden, S. Achebach, and G. Holighaus, "Control of FNR Function of Escherichia coli by O˜ 2 and Reducing Conditions," *Journal of molecular . . .*, 2002.

[165] E. Noor, E. Eden, R. Milo, and U. Alon, "Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy," *Molecular Cell*, vol. 39, pp. 809–820, Sept. 2010.

[166] R. Carlson and F. Srienc, "Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: creation of overall flux states," *Biotechnology and bioengineering*, vol. 86, no. 2, pp. 149–162, 2004.

[167] K. Robison, A. M. McGuire, and G. M. Church, "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome," *Journal of molecular biology*, vol. 284, pp. 241–254, Nov. 1998.

[168] A. M. McGuire, P. De Wulf, G. M. Church, and E. C. C. Lin, "A weight matrix for binding recognition by the redox-response regulator ArcA-P of Escherichia coli," *Molecular microbiology*, vol. 32, pp. 219–221, Apr. 1999.

[169] S. Estrem, W. Ross, T. Gaal, Z. Chen, W. Niu, R. Ebright, and R. Gourse, "Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase α subunit," *Genes & Development*, vol. 13, pp. 2134–2147, 1999.

[170] O. Chumsakul, H. Takahashi, T. Oshima, T. Hishimoto, S. Kanaya, N. Ogasawara, and S. Ishikawa, "Genome-wide binding profiles of the Bacillus subtilis transition state regulator AbrB and its homolog Abh reveals their interactive role in transcriptional regulation.," *Nucleic acids research*, vol. 39, pp. 414–428, Jan. 2011.

[171] D. P. Clark and J. E. Cronan, Jr, "Two-carbon compounds and fatty acids as carbon sources," *Escherichia coli and Salmonella: cellular and . . .*, 2005.

[172] U. Sauer, "The Soluble and Membrane-bound Transhydrogenases UdhA and PntAB Have Divergent Functions in NADPH Metabolism of *Escherichia coli*," *Journal of Biological Chemistry*, vol. 279, pp. 6613–6619, Nov. 2003.

[173] G. Unden and P. Dünnwald, "The Aerobic and Anaerobic Respiratory Chain of *Escherichia coli* and *Salmonella enterica*: Enzymes and Energetics," *Ecosal*, Mar. 2008.

[174] K. Kochanowski, B. Volkmer, L. Gerosa, B. R. H. van Rijsewijk, A. Schmidt, and M. Heinemann, "Functioning of a metabolic flux sensor in *Escherichia coli*," *Proceedings of the . . .*, vol. 110, pp. 1130–1135, 2013.

[175] J. D. Partridge, G. Sanguinetti, D. P. Dibden, R. E. Roberts, R. K. Poole, and J. Green, “Transition of *Escherichia coli* from Aerobic to Micro-aerobic Conditions Involves Fast and Slow Reacting Regulatory Components,” *Journal of Biological Chemistry*, vol. 282, pp. 11230–11237, Feb. 2007.

[176] H. van Rijsewijk Bart R B, A. Nanchen, S. Nallet, R. J. Kleijn, and U. Sauer, “Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli,” *Molecular systems biology*, vol. 7, pp. 1–12, Mar. 2011.

[177] T. A. Krulwich, G. Sachs, and E. Padan, “Molecular aspects of bacterial pH sensing and homeostasis,” *Nature Reviews Microbiology*, vol. 9, pp. 330–343, Apr. 2011.

[178] M. Cosentino Lagomarsino, P. Jona, B. Bassetti, and H. Isambert, “Hierarchy and feedback in the evolution of the Escherichia coli transcription network.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 5516–5520, Mar. 2007.

[179] D. E. Seborg, D. A. Mellichamp, T. F. Edgar, and I. Francis J Doyle, *Process Dynamics and Control*. Wiley, Apr. 2010.

[180] K. Patil and H. Nielsen, “Uncovering transcriptional regulation of metabolism by using metabolic network topology,” *Proceedings of the National Academy of Sciences*, vol. 102, pp. 2685–2689, Feb. 2005.

[181] N.-M. Gruning, H. Lehrach, and M. Ralser, “Regulatory crosstalk of the metabolic network,” *Trends in Biochemical Sciences*, vol. 35, pp. 220–227, Apr. 2010.

[182] G. Semenza, “Hypoxia-Inducible Factors in Physiology and Medicine,” *Cell*, vol. 148, pp. 399–408, Feb. 2012.

[183] M. Bostock, V. Ogievetsky, and J. Heer, “D3: Data-Driven Documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[184] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal, “Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters,” *Nature Biotechnology*, vol. 30, pp. 521–530, May 2012.

[185] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. Ø. Palsson, “Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models,” *Molecular systems biology*, vol. 6, July 2010.

[186] J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. Ø. Palsson, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.," *Nature protocols*, vol. 6, pp. 1290–1307, Sept. 2011.

[187] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 21–29, 2007.

[188] A. M. Feist, H. Nagarajan, A.-E. Rotaru, P.-L. Tremblay, T. Zhang, K. P. Nevin, D. R. Lovley, and K. Zengler, "Constraint-based modeling of carbon fixation and the energetics of electron transfer in Geobacter metallireducens.," *PLoS Computational Biology*, vol. 10, p. e1003575, Apr. 2014.

[189] L. S.-A. E Richet, "CRP induces the repositioning of MalT at the Escherichia coli malKp promoter primarily through DNA bending.," *The EMBO journal*, vol. 13, p. 4558, Oct. 1994.