

**Regulation of *E. coli* Genome Architecture and Transcription in Response to Changing Nutrient Levels**

by

Grace Mahony Kroner

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Biological Chemistry)  
in the University of Michigan  
2018

Doctoral Committee:

Assistant Professor Peter L. Freddolino, Chair  
Assistant Professor Uhn-Soo Cho  
Associate Professor Mark Saper  
Associate Professor Raymond Trievel  
Assistant Professor Kevin Wood

Grace M. Kroner

gmkroner@umich.edu

ORCID iD: 0000-0003-4568-1287

© Grace M. Kroner 2018

## **Dedication**

This work is dedicated to my parents for their endless support.

## **Acknowledgements**

This work could not have been accomplished without extensive support from a number of people. Michael Wolfe's collaboration with the data analysis on the Lrp project was invaluable beyond words- having another person with whom to discuss ideas whenever we invariably went back to the drawing-board was essential. Catherine Barnier's assistance on the extensive and complicated work of processing the 3C-sequencing data was critical in us trying to make sense of our data in the context of the field. A special thanks is due to Scott Scholz for his curation of several data sets from the literature and stimulating discussion about how bacterial architecture may be regulated. Other students and staff in the department, especially Elizabeth Abshire of the Trievel lab and Tom Jurkiw, Justin McNally and Michael Baldwin of the O'Brien lab deserve special thanks for patiently putting up with my protein purification endeavors and associated questions.

As an overall guide, my mentor Peter Freddolino provided me with critical suggestions for everything from for where to go to learn basic programming to how to deal with big analytical questions. With his amazing ability to answer questions at seemingly any hour and his full-hearted encouragement of my career plans, I cannot imagine a more supportive mentor.

The members of my thesis committee, Uhn-Soo Cho, Mark Saper, Ray Trievel and Kevin Wood have been equally supportive and kind in providing suggestions from all their respective areas of expertise. I would additionally like to thank Prof. Bob Blumenthal for being an early and continued outside advisor on the Lrp project- who fully sympathizes with all that protein's oddities.

Finally, the student administrative support in this department, especially Beth Goodwin, Amanda Howard, Prasanna Baragi, Lisa VanMeerbeek, Julie Woodruff, Mary Grapp, and Jamie Winkle have always been endlessly friendly and helpful.

I was supported by the Cellular Biotechnology Training Program (T32-GM008353) and a one-time Rackham Research Grant from the University of Michigan.

I'll end with a quote from one of my favorite (among many) authors to whom I owe a debt of gratitude for providing numerous mental breaks from bacteria over the years:

“It doesn't stop being magic just because you know how it works.”

-Sir Terry Pratchett, *The Wee Free Men*

## Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	ix
List of Figures	xi
Abbreviations	xiii
Abstract	xiv
Chapter 1 Introduction	1
Chromosome Organization	1
Cause and effect in chromosome organization and transcription	1
Models for Bacterial Genome Organization	3
Organizing Principles in Eukaryotic Chromosomes	4
Global Regulators	6
Importance of Global Regulators in Transcription Regulation	6
Lrp/AsnC Family of Global (and Local) Regulators	7
Lrp's Role in Bacterial Pathogenesis	8
Lrp Structure and Oligomerization	8
Lrp-DNA Binding	9
Leucine responsive: the coregulator effect	11
Lrp Expression Regulation	12
Methods of Regulation	13
Analogies to Eukaryotic Systems	13
Conclusions	14
Tables	16

Figures	17
Chapter 2 Chromosome Architecture in <i>E. coli</i>	19
Abstract	19
Introduction	19
Results	23
Interaction propensity of each region of DNA strongly influences the observed contact matrix	23
Limited global evidence for macrodomain-based structuring	24
CID boundaries are relatively consistent across conditions	24
rRNA clustering is recapitulated in our data	26
Patterns of interaction propensities are generally shared across samples	27
Regions of high and low interaction propensity are enriched for different classes of genes	28
Interaction propensity is globally correlated with several other features	30
Discussion	31
The <i>E. coli</i> chromosome does not form an ellipsoid or strong macrodomains	31
Interaction propensity-based model for chromosome organization	33
Materials and Methods	34
Genome Conformation Capture	34
Strains and media	35
Sequencing data analysis pipeline	37
Computational analysis	38
Tables	40
Figures	49
Chapter 3 Role of the Global Regulator Lrp	59
Abstract	59
Introduction	59
Results	62
ChIP-seq and RNA-seq identify hundreds of novel Lrp targets	62
Global analysis reveals that Lrp has condition-specific modes of binding and regulation	64
Lrp binding is enriched among regulatory regions of the genome	67
Direct Lrp targets explain the Lrp-dependent regulatory effect at some indirect targets	68

Direct and indirect Lrp targets have both shared and unique GO-term classifications	70
Lrp is poised at many targets to enable combinatorial regulation	72
Lrp connects with other regulatory factors	74
Lrp binding sites have a condition- and time-specific motif preference	78
Lrp binding peak length is relatively invariable	80
Discussion	81
Lrp regulates hundreds of genes in distinct categories by direct and indirect mechanisms	81
Primed Lrp binding argues for interaction with coregulatory factors	82
Lrp binding activity is partially predicted by known sequence motifs	84
Materials and Methods	85
Strains and media	85
ChIP-seq	87
RNA-seq	88
Preprocessing and alignment of ChIP-seq data	89
Calculation of ChIP-seq summary signal	90
Determination of high-confidence Lrp binding sites	90
Preprocessing of RNA-Seq data	92
Determination of Lrp-dependent changes in Transcription	92
Antibody development and testing	93
Filtering of genes into Lrp-dependent categories	93
Data Availability	94
Tables	95
Figures	103
Chapter 4 Conclusions	121
Summary of Thesis	121
<i>E. coli</i> chromosome architecture is dominated by interaction propensities	121
rRNA operon clustering represents an example of a secondary organizing feature	121
Macrodomains are not universally observed in our data	122
Regulatory patterns provide strong evidence for Lrp as a global regulator	122
Lrp binding exhibits changing specificity during later stages of growth	123
Lrp likely interacts with a variety of coregulators	124
Future Directions	124



## List of Tables

Table 1.1: Roles of global regulators .....	16
Table 2.1: Traditional macrodomain permutation test results .....	40
Table 2.2: Modified macrodomain permutation test results .....	41
Table 2.3: Results of permutation test for enrichment of highly transcribed regions at CID boundaries. ....	42
Table 2.4: Results of permutation test for enrichment of low interaction propensity at CID centers .....	42
Table 2.5: rRNA operon clustering permutation test results .....	43
Table 2.6: Highly expressed, related gene clustering permutation test results .....	44
Table 2.7: Genome-wide feature correlations .....	46
Table 2.8: Genotype of strains used in this study .....	47
Table 2.9: Primers used in this study .....	48
Table 3.1: Genes with significant Lrp-dependent changes in expression .....	95
Table 3.2: Ambiguous indirect targets .....	95
Table 3.3: Lrp preferentially binds regulatory regions .....	96
Table 3.4: Locations of non-regulatory region peaks. ....	97
Table 3.5: Indirect target annotation .....	98
Table 3.6: Results of permutation test for enrichment of direct Lrp targets relative to NAP-type targets within the known $\sigma$ factor regulons at each condition .....	99

Table 3.7: Results of permutation tests for enrichment of direct Lrp-activated targets relative to direct Lrp-repressed and NAP-type targets within the known $\sigma$ factor regulons at each condition .....	100
Table 3.8: Performance of Lrp binding site prediction models .....	101
Table 3.9: Primers used for <i>lrp::kanR</i> construction.....	101
Table 3.10: Media conditions for cell growth.....	102

## List of Figures

Figure 1.1: Proposed models for bacterial genome organization.....	17
Figure 1.2: Hierarchical regulatory structure in <i>E. coli</i> .....	17
Figure 1.3: Lrp monomer and octamer structure.....	18
Figure 2.1: Depiction of ring polymer model.....	49
Figure 2.2: Contact matrices are dominated by the interaction propensity signal.....	50
Figure 2.3: CID organization is consistent across samples and correlated with interaction propensity.....	51
Figure 2.4: rRNA operons spatially cluster in a time-dependent manner.....	53
Figure 2.5: Interaction propensity is strongly conserved.....	55
Figure 2.6: Select mutations result in perturbations of the interaction propensity.....	56
Figure 2.7: GO-term enrichment relative to interaction propensity.....	57
Figure 2.8: Modeling results.....	58
Figure 3.1: Depiction of experimental time points.....	103
Figure 3.2: ChIP-seq data shows agreement with previous data and reveals novel Lrp binding sites.....	104
Figure 3.3: Lrp regulates genes both directly and indirectly.....	107
Figure 3.4: Intragenic Lrp peaks do not systematically affect transcription.....	108
Figure 3.5: Known targets of direct Lrp targets explain the mechanism of indirect Lrp regulation at some genes.....	109

Figure 3.6: Enriched GO-terms differ for direct and indirect Lrp targets .....	110
Figure 3.7: Full GO-term enrichment results for general target classification and sub- classification by direction of Lrp regulatory change .....	111
Figure 3.8: Lrp sits at genes in poised position in preparation for regulatory activity .....	113
Figure 3.9: Lrp interacts with other regulatory factors to control some targets' expression .....	114
Figure 3.10: Characteristics and regulatory activities of potential Lrp partners.....	115
Figure 3.11: Lrp exhibits condition-dependent sequence-preference.....	116
Figure 3.12: Changes in BIC for leave-one-out logistic regression models.....	117
Figure 3.13: Lrp peaks are a consistent length. ....	118
Figure 3.14: Lrp ChIP-Seq data is highly reproducible.....	119
Figure 3.15: Lrp antibody does not interfere with DNA binding and is specific for Lrp.....	120

## Abbreviations

3C	→ Chromosome conformation capture
BIC	→ Bayesian information criterion
ChIP	→ Chromatin immunoprecipitation
CID	→ Chromosomal interaction domain
CRP	→ Cyclic AMP receptor protein
CTCF	→ CCCTC-binding factor
Dam	→ DNA adenine methyltransferase
DI	→ Directional index
EPOD	→ Extended protein occupancy domain
Fis	→ Factor for inversion stimulation
FNR	→ Fumarate and nitrate reductase
GO-term	→ Gene ontology term
H-NS	→ Histone-like nucleoid-structuring protein
HTH	→ Helix-turn-helix domain
IHF	→ Integration host factor
LIV	→ Minimal media plus leucine, isoleucine and valine
Log	→ Logarithmic phase
Lrp	→ Leucine responsive regulatory protein
MCC	→ Matthew's correlation coefficient
MD	→ Macrodomain
MIN	→ Minimal media
NAP	→ Nucleoid-associated protein
OriC	→ Origin of replication
PTM	→ Post-translational modification
qPCR	→ quantitative reverse-transcriptase PCR
RAM	→ Regulator of amino acid metabolism domain
RDM	→ Rich defined media
RNA-seq	→ RNA-sequencing
RNAP	→ RNA polymerase
rRNA	→ Ribosomal RNA
SMC	→ Structural maintenance of chromosome
Stat	→ Stationary phase
TAD	→ Topologically associated domain
Ter	→ Terminus
Trans	→ Transition point
TSS	→ Transcription start site

## Abstract

Organisms live in an incredible variety of conditions and to survive must have the ability to sense environmental conditions and respond accordingly. A fundamental level of response is the triggering of changes in gene expression. In bacteria, as in eukaryotes, a dizzying array of factors, including chromosome organization and transcription factor activities, can contribute to transcriptional regulatory effects. The goal of this work was twofold: 1) to monitor the global changes and potential contributing factors of *E. coli* chromosome architecture and 2) to elucidate the role of the global regulator Lrp.

Chromosome architecture has a subtle but fundamental influence on regulatory control. In order to analyze chromosome organization in *E. coli* at a global level, I utilized chromosome conformation capture (3C)-sequencing experiments in a variety of strains and conditions. We document chromosomal interaction domains as seen in earlier studies, but observe only weak evidence for macrodomains. Uniquely, we identify that a strong determinant of chromosome architecture is the propensity of each DNA region to interact with other DNA regions. The highly interacting and strongly isolated regions of DNA are often consistent across various strains and conditions, and we observe a close interplay between these interaction propensities and several other characteristics, such as transcription levels, DAM methylation sites, and known nucleoid associated protein binding locations. In addition, we observe enrichment of certain unique gene classes in regions with low or high interaction propensity. For example, genetic components of several core biosynthetic processes appear in regions of high interaction

propensity, while genes with transposase activity are present in low interacting regions. As previously seen in *E. coli*, we document clustering of the ribosomal RNA (rRNA) genes during logarithmic growth. The clustering is attenuated during both stationary and late-stationary growth points, suggesting that high requirements for ribosomes during rapid growth may contribute to rRNA genes being clustered. This pattern evokes the concept of the eukaryotic nucleolus and documented transcription factories.

Lrp (leucine responsive regulatory protein) is a global regulator in *E. coli* known to control at least 10% of the genome and is especially critical upon nutrient limitation and entrance to stationary phase. To document the full binding and regulatory activity of Lrp, I performed matched ChIP-seq and RNA sequencing under nine physiological conditions. These experiments not only allowed us to identify hundreds of novel direct and indirect targets, expanding the Lrp regulon to 35% of all genes, but also enabled us to propose several potential methods of Lrp regulation. At many promoters, we note that Lrp binding can occur without causing any regulatory activity (nucleoid associated protein (NAP)-type targets). This poised binding may indicate that Lrp regulatory activity requires cooperation with other factors in *E. coli*, such as the nitrogen-response sigma factor  $\sigma^{54}$ . In addition, we observe an increase in Lrp's DNA-binding specificity during later points of growth, potentially explaining why Lrp has generally been considered a regulator of stationary phase entry.

In summary, this work provides an enhanced conception of bacterial regulation on the global scale in terms of overall chromosome organization and the global regulator Lrp's full spectrum of activity. The possible methods posited for Lrp's regulatory behavior may inform study of other global regulators, and the investigation of chromosome architecture illuminates

interesting future avenues of research on the nature of causative factors that establish chromosome organization.

## **Chapter 1 Introduction**

Organisms live in an ever-changing milieu and to survive must have the ability to adapt to fluctuating environmental conditions. A fundamental level of response is the triggering of changes in gene expression. This can be seen in everything from bacteria responding to stressful environments to multicellular eukaryotic organisms following a meal. While there is another level of control for translation, variation in transcript levels accounts for half of the variability in protein abundance (Guimaraes, Rocha, & Arkin, 2014; Lu, Vogel, Wang, Yao, & Marcotte, 2006). Transcriptional regulation classically results from interventions by specific transcription factors, but other aspects of bacterial physiology can also modulate transcription levels.

### **Chromosome Organization**

#### *Cause and effect in chromosome organization and transcription*

Chromosome structure is a well-established method of regulation in eukaryotes; it facilitates coordinated regulation of related genes through modulating transcription machinery's access to DNA and orchestrating the formation or elimination of crucial enhancer-promoter interactions (Doyle, Fudenberg, Imakaev, & Mirny, 2014). Chromosome conformation capture studies in eukaryotes have provided high-resolution data about the precise interactions of folded chromosomes (Dekker, Rippe, Dekker, & Kleckner, 2002; Lieberman-Aiden et al., 2009). Recent studies have drawn attention to bacterial genome organization and how that identified organization might impact transcription or be influenced by transcription. At a one-dimensional

level, there is evidence that the location of genes along the chromosome may have some effect on transcription (Scholz, Diao, Fivenson, Lin & Freddolino, in preparation). Transcription of a reporter randomly inserted at tens of thousands of sites within the *Escherichia coli* (*E. coli*) chromosome resulted in a dramatic range of expression values, termed transcription propensity, which vary cyclically along the genome. Earlier studies with a limited set of sites also saw variation in transcription levels and gene insertion position (Bryant, Sellars, Busby, & Lee, 2014). Some of the locations with highest reporter expression occur near the genes encoding ribosomal RNA (rRNA). rRNA operons are often some of the most highly-expressed genes in *E. coli*, so it is not surprising that they might have an effect on linearly neighboring genes. However, it is well established that distal two-dimensional locations do not imply equivalently far three-dimensional positions. For example, there is evidence for spatial colocalization of six out of the seven rRNA operons (Gaal et al., 2016), which could point to the formation of ‘transcription factory’ like regions in bacteria.

In addition, the proteins canonically responsible for packing and condensing bacterial DNA, the nucleoid associated proteins (NAPs), have recently been implicated in the formation of extended protein occupancy domains (EPODs) on DNA (Goss & Freddolino in preparation). DNA regions within EPODs exhibit low transcription, and a combination of NAPs, such as Fis (factor for inversion stimulation), HU and HNS (Histone-like nucleoid-structuring protein), all appear to be important for binding, yet are all redundant to a certain degree. The presence of certain NAPs may facilitate or inhibit various three-dimensional structures of the DNA, and so these EPOD regions suggest another potential link between chromosome organization and transcription regulation.

## *Models for Bacterial Genome Organization*

Studies of chromosome organization in bacteria have suggested two organizational schemes: macrodomains and ellipsoid folding (Figure 1.1). Macrodomains are relatively large regions of the chromosome whose DNA preferentially interacts with other DNA in the same macrodomain, while making limited contacts to DNA in other macrodomains. Data suggests that the *E. coli* chromosome is organized into four macrodomains- *ter*, *ori*, left and right- and two unstructured regions (Cagliero, Grand, Jones, Jin, & O'Sullivan, 2013; Espeli, Mercier, & Boccard, 2008; Valens, Penaud, Rossignol, Cornet, & Boccard, 2004). However, recent Hi-C studies in *E. coli* documented weaker demarcations for the macrodomains besides *ter* (Lioy et al., 2018). This result suggests that the experimental method used may greatly influence the ability to detect macrodomains, or that not all macrodomains may be equivalent in terms of their DNA-interaction preferences.

Several proteins have been identified as being critical for individual macrodomain formation in *E. coli*. MatP binding to its recognition site, *matS*, facilitates isolation of the *ter* macrodomain (Mercier et al., 2008). In parallel, occupation of *maoS* sites by MaoP helps condense the *ori* macrodomain (Valens, Thiel, & Boccard, 2016).

Instead of forming macrodomains, the single circular chromosome of *Caulobacter crescentus* twists and folds in half like a coiled rubber band with some areas forming chromosomal interaction domains (Le, Imakaev, Mirny, & Laub, 2013; Umbarger et al., 2011). A similar pattern is apparent for the *Bacillus subtilis* chromosome (Marbouty et al., 2015). This elongated chromosome structure is usually aligned along the long-axis of the cell, and locations of regulatory sequences in the DNA play crucial roles in determining the poles of the folded, condensed circle (Umbarger et al., 2011).

NAPs are likely important in establishing both of these global forms of organization as well as the local interactions important for condensation. By oligomerizing, H-NS is known to be able to bridge two distant regions of DNA (Dame, 2005), in addition to causing transcriptional repression, especially in regions of horizontally acquired DNA (Browning, Grainger, & Busby, 2010). MukB in *E. coli* and other structural maintenance of chromosome (SMC) complexes can form restrictive rings around two DNA duplexes, facilitating close interactions between the bound regions of DNA and sometimes forming an extruded loop (Luijsterburg, Noom, Wuite, & Dame, 2006).

NAPs that bend DNA include Fis, Lrp (leucine responsive regulatory protein), HU, and its related protein IHF (integration host factor) (Dame, 2005). Fis is highly prevalent during logarithmic growth and is able to either activate or repress genes by a variety of means (Browning et al., 2010). Lrp can form octamers around which DNA is wrapped, creating a nucleosome-like structure (Dame, 2005), and, like Fis, can activate or repress genes in addition to its DNA-packaging role (see Chapter 3). HU is implicated in controlling supercoiling to some degree through its non-specific binding activity (Luijsterburg et al., 2006), while the closely related protein IHF has a more regulatory role due to its sequence-specific DNA binding activity (Dame, 2005). The loops formed from DNA-bridging NAP-activity are often bound near the loop tip by DNA-bending proteins, and so it has been proposed that DNA-bending proteins, like their bridging counterparts, may contribute to loop location selection (Luijsterburg et al., 2006). How these proteins contribute to global organization either via macrodomains or ellipsoid formation is not entirely clear.

*Organizing Principles in Eukaryotic Chromosomes*

Extensive studies have tackled the challenge of deciphering chromosome organization in eukaryotic, especially mammalian, cells. Hi-C, an experimental method modified from 3C to improve its functionality with large genomes by isolating the junction containing DNA using a biotin labeled nucleotide, has enjoyed considerable success. The earliest Hi-C studies showed partitioning of the chromosomes into mostly non-overlapping chromosome territories (Lieberman-Aiden et al., 2009). In addition, normalization of the Hi-C data revealed an intriguing plaid pattern suggesting that there were two main compartments (A and B) of the genome that preferentially interacted with other DNA in the compartment rather than DNA in the other compartment. Upon investigation, compartment A was correlated to more accessible chromatin regions, annotated gene-coding regions, and higher levels of transcription, indicating that it might represent euchromatin as opposed to the more heterochromatin-like, closely packed, DNA in compartment B (Lieberman-Aiden et al., 2009). The assignment of DNA regions to A and B compartments is cell-type dependent (Dekker, Marti-Renom, & Mirny, 2013), in agreement with the phenomenon's proposed connection to transcription regulation.

In contrast, topologically associated domains (TADs) are relatively immobile in different cell types (Dekker et al., 2013). TADs are identified by the fact that DNA within a TAD is more likely to interact with other DNA in that TAD compared to DNA outside the TAD. They range in size from 400-500 kb, and their boundaries are often marked by genetic boundary elements, which sometimes exhibit CTCF (CCCTC-binding factor) binding (Dekker et al., 2013). In bacteria, the TAD equivalent is the chromosomal interaction domain (CID). CIDs range from 30-420 kb, and their boundary regions often encode highly expressed genes (Le et al., 2013). Among the previously studied bacterial species of *Bacillus subtilis*, *Caulobacter crescentus*, *E.*

*coli* and *Vibrio cholerae*, the number of CIDs range from 20 to 30 (Le et al., 2013; Liou et al., 2018; Marbouty et al., 2015; Val et al., 2016; X. Wang et al., 2015).

Studies employing fluorescent labeling have also identified the existence of transcription factories in eukaryotes. The most well-known is the nucleolus, which spatially localizes ribosome biogenesis by clustering the hundreds of copies of the 45S rRNA gene in an area with a high concentration of RNA polymerase I (Papantonis & Cook, 2013). In general, transcription factories are recognized as distinct foci with an increased local concentration of one of the RNA polymerases (RNAP). Unique factories containing RNAP II and RNAP III contribute to the vast majority of cellular transcription activity (Papantonis & Cook, 2013). There are indications that similar factories may occur in bacteria based on studies of rRNA colocalization (Gaal et al., 2016) and fluorescent RNAP visualization in several conditions (Cabrera Julio & Jin Ding, 2003; Endesfelder et al.).

## **Global Regulators**

### *Importance of Global Regulators in Transcription Regulation*

In bacteria and archaea, regulatory systems are often set up in a hierarchical manner to facilitate far-ranging responses to environmental changes (Figure 1.2). Thus, the full complement of transcription factors is regulated by a limited number of global regulators (Ma, Buer, & Zeng, 2004). In addition to activating or repressing genes, these global regulators can recruit alternative sigma factors or serve as direct effectors of RNA polymerase (Newman, D'Ari, & Lin, 1992). In *E. coli*, the seven global regulators, ArcA, FNR (fumarate and nitrate reductase), Fis, CRP (cyclic AMP receptor protein), IHF, H-NS and Lrp, control an astonishing

percentage of genes: 50% of genes are responsive to at least one of the seven global regulators (Table 1.1, Martínez-Antonio & Collado-Vides, 2003).

Understanding the role global regulators play in bacterial transcription control is an important step to better modelling of bacterial system networks. It can even potentially provide insight into regulation in eukaryotes (Kawashima et al., 2008).

### *Lrp/AsnC Family of Global (and Local) Regulators*

One important family of global regulators present throughout many bacteria and archaea is the Lrp/AsnC family (Brinkman, Ettema, De Vos, & Van Der Oost, 2003). Among all homologs, there are significant differences in gene targets and in the amount of protein produced (Lintner et al., 2008). This is in part due to the distinction between Lrp/AsnC family members with a global role (Lrp-like, with a large variety of gene targets and a relatively high cellular concentration) as opposed to those with a local role (AsnC-like, with limited gene targets and a relatively low cellular concentration) (Friedberg, Midkiff, & Calvo, 2001). For example, LrfB in *Haemophilus influenzae* is 75% identical to *E. coli* Lrp, but has a local regulatory role and an average concentration of 130 dimers per cell, as opposed to the 3000 dimers per cell for *E. coli* Lrp (Friedberg et al., 2001; D A Willins, Ryan, Platko, & Calvo, 1991). Generally, the global regulatory role for Lrp-AsnC family members is restricted to enteric bacteria and commonly mediates the feast-famine response (Friedberg et al., 2001).

*E. coli* Lrp is the eponymous member of the Lrp/AsnC protein family, and regulates 70% of the 215 genes with differential expression upon entrance to stationary phase (Tani, Khodursky et al. 2002). It influences a variety of cellular processes: amino acid synthesis, degradation and transport, porin expression, and pilus formation (Willins, Ryan et al. 1991, Haney, Platko et al.

1992). A recent study has also implicated Lrp in regulation of the *tos* operon, required for nonfimbrial adhesion in urinary tract infections caused by *E. coli* (Engstrom & Mobley, 2016). The latter finding adds to a growing list of Lrp homologues which have recently been tied to expression of virulence genes (see below).

### *Lrp's Role in Bacterial Pathogenesis*

Although Lrp in *E. coli* has been particularly well studied due to the fact that *E. coli* is a commonly used model organism, studies of homologs in other organisms also provide insight into the importance and functionality of Lrp/AsnC family members. Recent studies have provided additional evidence connecting Lrp homologs to bacterial virulence. For example, overexpression of the *Salmonella enterica* serovar Typhimurium Lrp homolog, with 99% sequence identity to the *E. coli* protein, decreases virulence, and deletion causes increased virulence (Baek, Wang, Roland, & Curtiss, 2009). In addition, LrpA in *Mycobacterium tuberculosis* has been implicated in long term bacterial persistence (Parti et al., 2008). In *Vibrio cholerae*, the Lrp homolog activates expression of the transcription factor AphA, which activates virulence associated genes (W. Lin, Kovacikova, & Skorupski, 2007). These studies in homologs further establish an interesting and important role for Lrp/AsnC family proteins in bacterial regulation of pathogenesis, supporting the need for further study.

### *Lrp Structure and Oligomerization*

*Escherichia coli* Lrp is a basic 164 amino acid protein with a mass of 18.8 kD and a pI of 9.2-9.4 (Shaolin Chen, Hao, Bieniek, & Calvo, 2001; Perona, 2007; D A Willins et al., 1991). Initial mutational analysis of Lrp revealed the presence of three hypothetical domains: residues 16-70 affected DNA binding, residues 76-125 affected transcriptional activation, and residues

108-149 affected leucine binding (Platko & Calvo, 1993). Upon further studies of homology and eventual crystallization of Lrp, two major domains were identified: an N-terminal helix-turn-helix domain (HTH) and a C-terminal regulator of amino acid metabolism domain (RAM) (Figure 1.3, Perona, 2007). The HTH is a well-established domain for DNA binding, and the RAM causes allosteric regulation of amino acid metabolizing enzymes in prokaryotes (Ettema, Brinkman, Tani, Rafferty, & van der Oost, 2002). The  $\alpha\beta$  sandwich structure within the RAM domain is a key part of the dimer interface (Perona, 2007). Differences in amino acids of the N-terminal tail cause alterations in DNA binding affinity, indicating there may be a secondary interaction with DNA, besides that of the HTH (Hart et al., 2011).

Lrp dimers are known to oligomerize into octamer or hexadecamer forms. At micromolar concentrations, the Lrp hexadecamer is prevalent, and only at nanomolar concentrations is the dimer the major species (Shaolin Chen, Rosner, & Calvo, 2001). Salt concentrations exert strong effects on the oligomerization state, indicating the importance of electrostatic interactions, especially for the hexadecamer (Hart et al., 2011). In addition, binding of one leucine per octamer is sufficient to trigger hexadecamer dissociation (Shaolin Chen & Calvo, 2002; Shaolin Chen, Rosner, et al., 2001). This leucine binding site is expected to be near a conserved amino acid region on a loop between  $\beta$ -strands at the dimer interface (Perona, 2007). Finally, eleven amino acids at the C-terminus of the protein are required for higher-level oligomerization; in a  $\Delta 11$  mutant, dimers, but no other oligomerization states, are present despite having similar secondary structure (Shaolin Chen, Rosner, et al., 2001).

### *Lrp-DNA Binding*

Lrp has a weak, 15 bp consensus binding site established by SELEX experiments (Yuhai Cui, 1995). Lrp only binds dsDNA (Q. Wang & Calvo, 1993), and binds DNA as a dimer (Cui, Midkiff, Wang, & Calvo, 1996). The apparent dissociation constant for Lrp binding is 60 nM (Azam & Ishihama, 1999), but Lrp levels in a cell are usually higher, so the mechanism of regulatory action is slightly unclear (Shaolin Chen, Rosner, et al., 2001). Experiments done in a mini-cell producing strain allow quantification of free Lrp and provide partial answers to this question. Rapidly growing cells have low free Lrp concentrations, while slowly growing cells have high free-Lrp concentrations (Shaolin Chen, Hao, et al., 2001). Leucine binding decreases the amount of free Lrp by increasing Lrp's ability to bind DNA non-specifically: in minimal media, 38% Lrp is free without leucine, but upon leucine addition, only 17% is unbound (Shaolin Chen, Hao, et al., 2001). Additionally, the larger ratio of global regulators to affected sites is appropriate since regulators will sometimes be free in solution, as examined above, or bound non-specifically, so the relative amount bound is still effective as a regulator (Friedberg et al., 2001). This is especially true for a regulator with relatively high non-specific DNA binding activity, such as Lrp.

Lrp binding has important effects on DNA, even aside from its obvious regulatory impact. Lrp induces DNA curvature, causing a 52° bend if bound at one site and a 135° bend when bound at two sites (Q. Wang & Calvo, 1993). This bending effect is thought to be important in gene regulation (Roesch & Blomfield, 1998). In addition, Lrp binding is a cooperative process, which may be explained by binding-induced bending that allows further Lrp dimer interactions (S. N. Peterson, Dahlquist, & Reich, 2007). This cooperativity depends on proper spacing of the Lrp binding sites (S. Chen, Iannolo, & Calvo, 2005). Lrp cooperativity is additionally increased by leucine binding *in vitro* (Shaolin Chen & Calvo, 2002).

### *Leucine responsive: the coregulator effect*

Leucine was the first amino acid identified as having an effect on Lrp (D A Willins et al., 1991). Recent studies have revealed that amino acids such as alanine, methionine, isoleucine, histidine, and threonine also have detectable effects on Lrp function (Hart & Blumenthal, 2011). Leucine as a coregulator for a feast-famine protein is especially logical because leucine is a metabolic dead-end in *E. coli*, so it is a good signal for rich conditions (Newman et al., 1992). Binding studies have suggested that Lrp may have both a high affinity (-7 kcal/mol) and a low affinity binding site (between -4.66 and -6.75 kcal/mol depending on the oligomerization state of Lrp) (Shaolin Chen & Calvo, 2002). The low affinity binding site appears to be linked with changes in oligomerization.

However, leucine binding does not only affect the structure of Lrp, but is also critical for Lrp regulatory activity. Depending on the target, Lrp either activates or represses transcription, and in turn, leucine binding to Lrp either potentiates or inhibits Lrp function (Cho, Barrett, Knight, Park, & Palsson, 2008). For example, the *livK* gene is repressed in the presence of Lrp plus leucine, but upregulated when Lrp alone binds (Hart & Blumenthal, 2011). Similarly, Lrp activates the *foo* operon, which encodes genes required for fimbriae production, a key mediator of virulence (Berthiaume et al., 2004). However, in the presence of leucine or alanine, *foo* expression is repressed (Berthiaume et al., 2004). Another well-studied Lrp target *ilvIH* responds to leucine but not alanine. Finally, the *dadAX* operon has three Lrp binding sites that each have different responses to leucine or alanine binding and result in either operon activation or repression depending on the Lrp occupancy (Zhi, Mathew, & Freundlich, 1999). The mechanism of selective regulator modulation still needs to be investigated.

### *Lrp Expression Regulation*

It is self-evident that to have an effective global regulator, the expression and activity of the regulator itself must be tightly controlled in turn. Lrp is no exception. In minimal media, Lrp levels remain relatively constant, while in rich media, Lrp levels are low during exponential growth, but then increase in stationary phase (Landgraf, Wu, & Calvo, 1996). Approximate quantification of Lrp levels yields a value of 15  $\mu\text{M}$  in minimal media, versus 3.75-5  $\mu\text{M}$  in rich media, linking to its role as a feast-famine responsive protein (Shaolin Chen & Calvo, 2002). One aspect of Lrp regulation is sRNA-mediated translational control. The *E. coli* sRNA MicF is upregulated during rich conditions and inhibits Lrp translation by blocking initiation complex formation (Holmqvist, Unoson, Reimegård, & Wagner, 2012).

At the level of transcriptional control, Lrp binding represses its own promoter (Lintner et al., 2008; Q. Wang, Wu, Friedberg, Plakto, & Calvo, 1994).  $\beta$ -galactosidase reporter assays reveal that in a Lrp knock-out strain, Lrp operon expression increased two to three fold (Q. Wang et al., 1994). Conversely, when Lrp is overexpressed (to about eight times haploid levels), operon expression decreases approximately ten-fold. The same study identified that Lrp binds in the -80 to -32 region of the Lrp promoter based on electrophoretic mobility shift assays. Interestingly, leucine does not affect Lrp's auto-regulatory function (Q. Wang et al., 1994). Despite this interaction, Lrp auto-regulation does not seem to play a critical role in producing the characteristic low levels of Lrp during conditions of rich growth or the upregulation in stationary phase (Landgraf et al., 1996; R. Lin, D'Ari, & Newman, 1992).

To augment Lrp self-regulation, H-NS represses Lrp expression by binding in the promoter region (Oshima, Ito, Kabayama, & Nakamura, 1995). In opposition, the nucleotide

ppGpp activates the Lrp promoter, perhaps by binding to RNA polymerase (Landgraf et al., 1996). Like Lrp, ppGpp is inversely correlated with growth rate, so it is logical that Lrp activation would be linked with high levels of ppGpp.

### *Methods of Regulation*

Protein-protein interactions are a major method of regulation. Lrp may affect transcription in some cases by directly interacting with RNA polymerase (U. Pul, Wurm, & Wagner, 2007). In other cases, Lrp has been documented to interact with H-NS to cause gene repression by both binding DNA and forming a nucleo-protein structure that blocks transcriptional machinery (U. Pul et al., 2007). Interestingly, this phenomenon is not dependent on the ability of H-NS to bind DNA (Ü. Pul et al., 2005). Lrp can also compete with other proteins to bind DNA. Such is the case with regulation of the *pap* operon: Lrp competes with DNA-adenine methyltransferase (Dam) for certain binding sites, and only when a certain site is methylated and Lrp is bound to other sites will transcription occur (Stacey N. Peterson & Reich, 2008).

### *Analogies to Eukaryotic Systems*

Perhaps unsurprisingly, eukaryotic systems exhibit parallels to many aspects of bacterial regulation. At a fundamental level, eukaryotic regulatory networks also display a hierarchical organization (Gerstein et al., 2012). In addition, regulation at many promoters requires combinatorial interactions from several proteins, and the combination of factors binding and the downstream regulatory effects vary in a condition-dependent manner (Gerstein et al., 2012). Both transcriptional repression and activation can be accomplished through a variety of means:

directly inhibiting or stabilizing the formation of the transcription complex, creating favorable or unfavorable DNA architecture or interfering with the actions of other regulatory proteins (Gaston & Jayaraman, 2003; Green, 2005). While the eukaryotic system is more complex than the system in bacteria due to the greater genome size and larger number of regulated targets, it is clear that many fundamental principles of regulation are conserved.

## **Conclusions**

Regulation of transcription in response to changing environmental conditions is vital for every organism, and organisms have evolved to employ a variety of mechanisms for this requirement. Genome organization can control access to large swathes of the genome or can impact regulation by disrupting one crucial enhancer-promoter interaction. Likewise, by regulating an amalgam of metabolic enzymes, transporters and transcription factors, Lrp results in global changes in cellular behavior that allow survival under limiting nutrient conditions. The interplay between Lrp as a DNA-bending protein that might contribute to chromosomal architecture and Lrp as a regulatory factor are interesting to consider.

*E. coli* is a tractable system in which to investigate global regulators, genome organization and their interactions due to its limited size compared to other model organisms. Even though it is considered a well-established model organism, much is still unknown about how it accomplishes transcriptional regulation and chromosomal organization. The mechanism of Lrp's multi-faceted regulatory activity has stymied researchers ever since its discovery nearly thirty years ago. Early genome conformation capture studies in *E. coli* have indicated a dramatically different architecture than that seen for other bacterial species, so it is important to investigate both of these methods of regulation further. Not only will they provide further

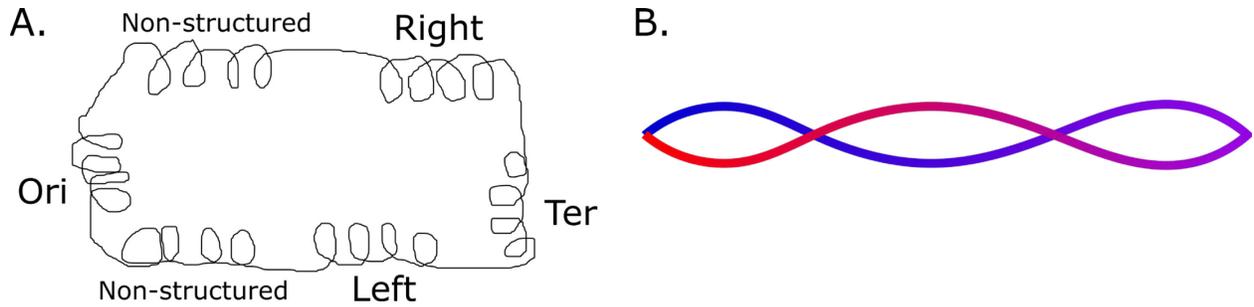
understanding of the *E. coli* system, but they may illuminate patterns that will aid in our comprehension of eukaryotic systems.

## Tables

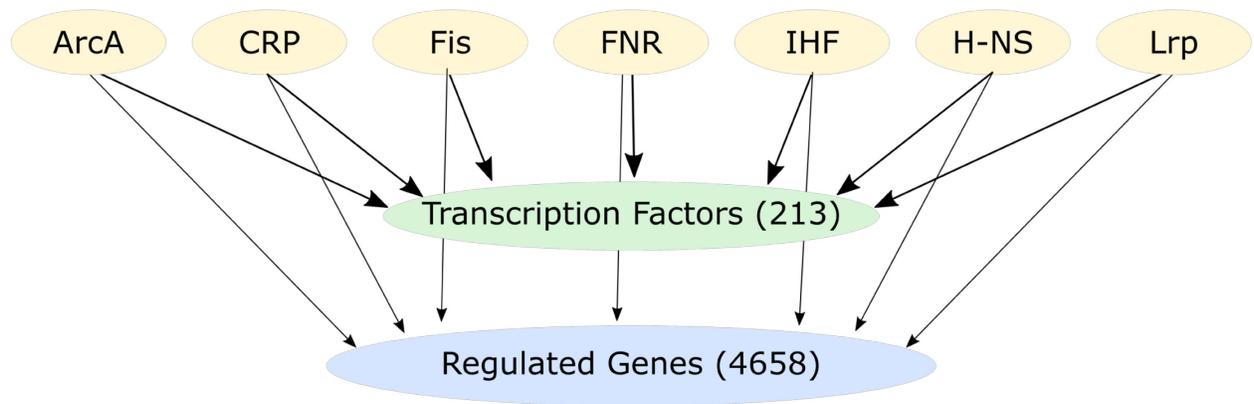
Global Regulator	Annotated Targets	Role	Abundance (molecules/cell)		
			Azam <i>et al.</i>	Schmidt <i>et al.</i>	Sutton <i>et al.</i>
ArcA	183	Energy/respiration mode		2863 (Log phase with glycerol) 2053 (Stationary phase)	
CRP	572	Energy balance/carbon source		2843 (Log phase with glycerol) 1229 (Stationary phase)	
FNR	306	Energy/respiration mode			4,100 (Log phase with glucose)
Fis	239	Effect DNA accessibility, energy dependent	60,000 (Log phase) undetectable (Stationary phase)		
H-NS	188	Effect DNA accessibility, energy dependent	20,000		
IHF	253	Effect DNA accessibility, energy dependent	12,000 (Log phase) 55,000 (Stationary phase)		
Lrp	109	Nutrient levels	6,400 (Log phase, minimal media)		

**Table 1.1: Roles of global regulators.** Number of annotated targets from Gama-Castro *et al.*, 2016. General functional role from Martínez-Antonio & Collado-Vides, 2003. Abundance estimates from Schmidt *et al.*, 2015, Ali Azam, Iwata, Nishimura, Ueda, & Ishihama, 1999 and Sutton, Mettert, Beinert, & Kiley, 2004.

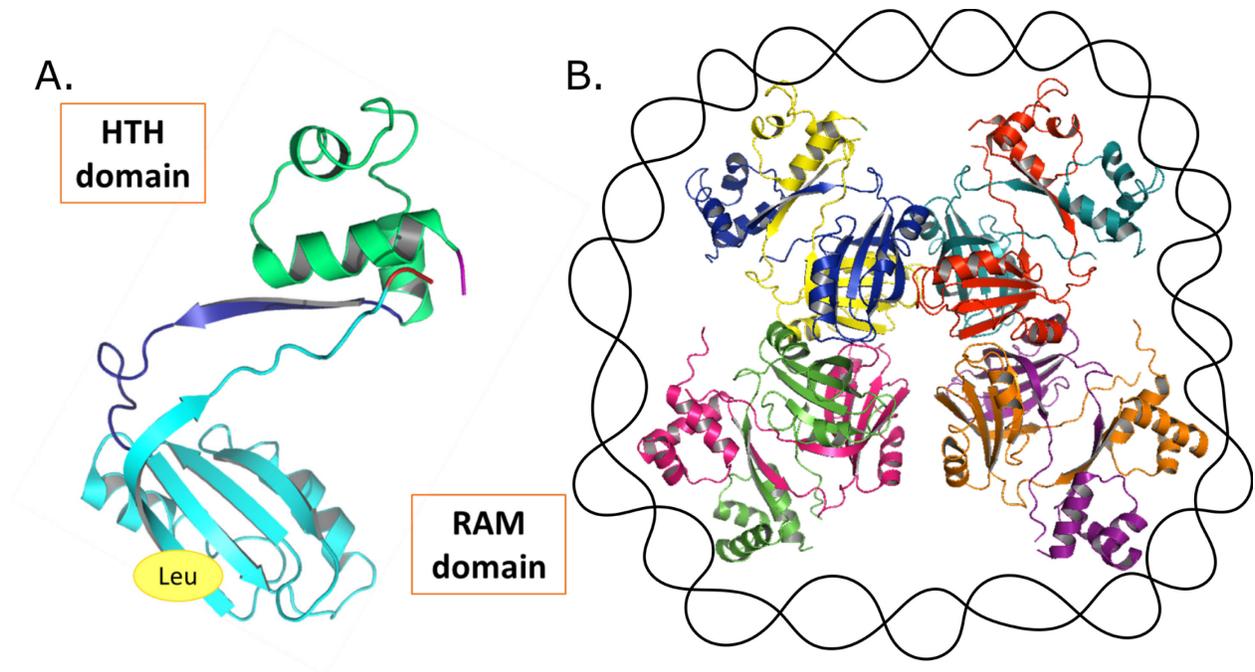
## Figures



**Figure 1.1: Proposed models for bacterial genome organization.** *A. Macrodomain model of *E. coli* chromosome organization. Figure based on Valens et al., 2004. B. Ellipsoid model of *Caulobacter crescentus* chromosome organization. Figure based on Umbarger et al., 2011.*



**Figure 1.2: Hierarchical regulatory structure in *E. coli*.** *The seven global regulators in *E. coli* control other transcription factors, as well as at least 50% of other genes in the genome directly. Figure adapted from Martínez-Antonio & Collado-Vides, 2003.*



**Figure 1.3: Lrp monomer and octamer structure.** *A. Structure of *E. coli* Lrp (PDB: 2GQQ). The N-terminal HTH domain, the C-terminal RAM domain, and the approximate leucine binding site location are indicated. B. Structure of *E. coli* Lrp octamer. Proposed location of DNA wrapping is shown by black lines.*

## Chapter 2 Chromosome Architecture in *E. coli*

### Abstract

Chromosome architecture plays an increasingly understood role in bacterial regulation. Previous work has proposed that the global architecture of the *E. coli* chromosome is organized into four large macrodomains which promote DNA-DNA interactions within, but not between, the macrodomains. Using 3C-sequencing from a variety of conditions and strains, we have assessed the strength and consistency of macrodomains as well as other genomic organizing features, such as the presence of CIDs and rRNA operon clustering. We find that while there is limited evidence for macrodomains in our data, we document strongly consistent CID boundaries and rRNA clustering. Critically, we identify a novel feature – each DNA region’s inherent propensity to form contacts with any other DNA, essentially its “stickiness” – that is a major determinant of chromosome organization. These interaction propensities are generally well-correlated across different conditions, with several interesting exceptions, and display correlations to several other genomic features. These results suggest that multiple organizing principles likely contribute to establishing bacterial chromatin architecture *in vivo*.

### Introduction

Given the increasingly recognized importance of chromosome architecture for regulation, there have been many investigations into the organization of the *E. coli* chromosome using a variety of different methods. One technique utilizes fluorescent proteins fused to specific DNA binding proteins (ParB) that bind to *parS* sites integrated at locations of interest on the

chromosome. Conveniently, ParB homologues from different species do not bind *parS* sites other than their own, allowing multiple fluorescent markers to be used to monitor the relative locations of two regions of the chromosome. Using this technique, Nielsen and colleagues showed that the left and right arms of the *E. coli* chromosome segregate to opposite sides of the cell along the longitudinal axis, with the *ori* region kept near the mid-point (Nielsen, Ottesen, Youngren, Austin, & Hansen, 2006). In their experiments, markers on the left and right arms of the chromosome appeared at opposite sides of the cells 70 to 90% of the time if the cells were not undergoing replication. Their observations agree with the ring polymer model for the *E. coli* chromosome with the *ori* region at the cell mid-point and the rest of the circular chromosome flattened along the longitudinal axis (Figure 2.1). Unlike the long fragments of the *C. crescentus* chromosome which wrap together like an old-fashioned telephone cord, the two halves of the *E. coli* chromosome in the ring polymer model do not physically interact (Youngren, Nielsen, Jun, & Austin, 2014). The model also suggests that regions of neighboring DNA condense to form 15-65 self-interacting domains along the *E. coli* chromosome (Youngren et al., 2014), which could be potentially analogous to the CIDs identified in other studies.

Another proposed model, not necessarily exclusive with the above, posits the existence of four macrodomains (MD) along the chromosome: *ori*, *ter*, right and left. This structure was initially proposed based on recombination-based assessments of loci proximity in which spatial localization was assumed to be equivalent to the efficiency of recombination between the two sites (Valens et al., 2004). Initial 3C-sequencing results in logarithmically growing cells in rich media supported the presence of the *ori* and *ter* MDs (Cagliero et al., 2013). However, in the analysis of that data, correct normalization for the varying abundance of DNA along the chromosome in rapidly growing cells was not performed; thus it is not surprising that we see

apparently higher interaction frequencies in the ori MD and depleted interactions in the ter MD since there are far more copies of DNA neighboring the origin compared to the terminus. Renormalization of this data revealed that operons which are coordinately regulated are kept in closer proximity than randomly chosen operons (Xie et al., 2015), suggesting a strong connection between regulation, transcription and chromosome architecture.

A more recent 3C-sequencing study showed relatively weak evidence for the presence of all MDs, but a strong argument that the ter MD region is kept isolated from the rest of the chromosome, such that it makes less long range contacts than other regions (Lioy et al., 2018). MatP and MukB appear to be critical for the ter region isolation, and NAPs such as Fis, H-NS and HU also appear to play a role in organization. Additionally, Lioy and colleagues note that the chromosome is divided into 31 40-300 kb CIDs in logarithmically growing cells; as in previous work (Le et al., 2013), the CID boundaries correlate with locations of highly expressed genes and sometimes (9 out of 31 cases) with genes carrying export signal sequences. Interestingly, while they still observe 30 CIDs in stationary phase cells, they state that the boundaries are non-overlapping with the boundaries identified from logarithmic cells.

Through direct observation of the DNA density in single *E. coli* cells chemically treated to have a wider cell diameter, Wu and colleagues note that the chromosome forms a torus, or donut shape, with low density regions corresponding to the ori and ter regions at mid-cell and more densely packed DNA in the right and left arms (Wu, Japaridze, Zheng, Kerssemakers, & Dekker, 2018). The low density ter and ori regions are eliminated in *matP* KO cells. They observe ~1 Mb “blobs” of density which rapidly move along the chromosome. These blobs are eliminated upon inhibition of transcription or supercoiling maintenance, suggesting some

potential mechanisms for controlling chromosome architecture (Wu et al., 2018). HU and Fis are implicated in establishing the flexible boundaries of the blobs.

Given that several of these studies have implicated transcription as being associated with chromosome architecture, Gaal and colleagues investigated whether the highly expressed rRNA operons are co-localized as in a eukaryotic nucleolus. Using the ParB/*parS* system described previously, they detect co-localization of six out of the seven rRNA copies during logarithmic growth in rich media (Gaal et al., 2016). However, transcription is not actually required for the observed co-localization, since co-localization occurs between operons in stationary phase, as well as in cells treated with the transcription inhibitor rifampicin (Gaal et al., 2016).

In this work, we employ 3C-sequencing assays to monitor the conservation of chromosome organization and assess potential contributing factors across a variety of conditions and strains, including lab strain *E. coli* and *E. coli* Crooks (ATCC8739). We document CIDs and colocalization of the rRNA operons as in previous studies. However, we make the novel observation that each region of DNA appears to have an inherent interaction propensity that influences its likelihood of interacting with any other DNA. Since this is an abundant and consistent pattern in our data, we investigated what factors, such as NAP binding, DNA-methylation sites, or overlap with certain classes of genes, might contribute to this interaction propensity variability. Regions of high interaction propensity are correlated with increased transcription and DNA-methylation, while regions of low interaction propensity are enriched for protein-binding. These observations suggest that chromosomal organization in *E. coli* may be analogous to the A/B (euchromatin/ heterochromatin) compartmentalization seen in eukaryotes (Lieberman-Aiden et al., 2009).

## Results

### *Interaction propensity of each region of DNA strongly influences the observed contact matrix*

We performed 3C-sequencing on WT *E. coli* during logarithmic growth, stationary phase, and late stationary phase, in addition to testing a number of mutants under logarithmic growth conditions. The genome-wide pattern of interaction frequencies is displayed in a contact matrix for each sample, and for the following analysis, the resolution of the contact matrices is 20 kb (i.e. each pixel represents a 20 kb region of DNA). The contact matrices allow a global assessment of genome organization.

Upon normalizing our contact frequency matrix for the abundance of DNA, we observed a strong and consistent plaid pattern across conditions (Figure 2.2A). This pattern appeared to occur because some regions of DNA have a higher inherent propensity to interact with any other fragment of DNA, while other regions are more isolated, leading to intersecting bright and dark bands on the contact matrix. We hypothesized that these strong patterns might be obscuring subtler patterns in the contact matrix. Accordingly, we added another level of normalization that removes each bin's inherent interaction propensity by dividing the signal at each position (x,y) in the contact matrix by the product of the interaction propensities for the x and y bins; the interaction propensity for each bin is determined by summing all the interaction frequencies of the bin with all other bins. The resulting contact matrix was essentially random noise and did not display any evidence of higher order structure, such as macrodomains or the off-diagonal seen in *C. crescentus* and *B. subtilis* (Figure 2.2B).

To confirm that the inherent interaction propensity is the major contributor to the observed contact matrices, we used the calculated interaction propensities of each bin to simulate a contact matrix (Figure 2.2C). As seen in our actual data, the simulated data displays a striking

plaid pattern, with bands of high and low interaction frequency. Many of the bins near the *ter* region have low interaction propensities.

#### *Limited global evidence for macrodomain-based structuring*

Given that our contact matrices did not visually resemble what we would expect if MDs were present (compare Figure 2.2A and D), we wanted to investigate if MDs were present but masked by the strong plaid interaction propensity signal. As originally defined, MDs are regions of DNA that interact preferentially with other DNA inside the MD rather than DNA outside it. We tested this by comparing the median interaction frequency of interactions within the MD to the median interaction frequency of interactions between DNA in the macrodomain and DNA outside the macrodomain (see Methods). Results of the permutation test querying every potential bin location at which a MD might start are shown in Table 2.1. Only the left MD has an enrichment of interactions within relative to without. The definition of macrodomain is now sometimes used to suggest something unique about the interaction frequency of that region of DNA compared to the rest of the chromosome (Lioy et al., 2018). Accordingly, we tested whether the median interaction frequency within a proposed MD was significantly different than the median interaction frequency for identically sized regions at every bin location in the genome. Again, the left MD was the only one which was statistically significant, though some individual conditions showed evidence for the presence of the *ter* MD (Table 2.2). While these results do not rule out MDs, they suggest that MD organization is potentially obscured or overwhelmed by other organizational factors depending on the measurement technique used.

#### *CID boundaries are relatively consistent across conditions*

CIDs have been identified in all bacteria whose genome structures have been studied to date. Interestingly, in *E. coli*, CID boundaries for logarithmically growing cells did not overlap well with boundaries for cells in stationary phase (Lioy et al., 2018). We performed directional index (DI) analysis on our data to identify the level of upstream or downstream interaction bias for each bin, and then used the resulting t-statistic to call CID boundaries (see Methods); a sample of the DI plot and the resultant CID boundaries is shown in Figure 2.3A. We observe between 14 and 19 CIDs in our data, ranging in size from 80 kb to 680 kb. Although the CID boundaries are not precisely aligned when comparing between conditions, we observe fairly minor shifting in location, especially in the first 75% of the chromosome (Figure 2.3B). This agrees with the observation in eukaryotic cells that TADs, the eukaryotic equivalent of CIDs, are fairly immobile in different cell-types. We observe limited agreement between the locations of CID boundaries in our WT log phase cells and the CID boundaries identified in Lioy et al., 2018. However, we do not have exact coordinates for their boundaries, and given the differences in strains used, that makes the comparison approximate at best.

Previously identified CID boundaries in *C. crescentus* and *E. coli* were enriched for genes that were highly expressed (Le et al., 2013; Lioy et al., 2018), and so we compared RNA-sequencing data from an identical strain grown in MOPS minimal conditions and harvested during logarithmic growth (see Chapter 3). Visually, the correlation did not appear very strong (Figure 2.3C), and that was borne out by a permutation test in which we randomly sampled RNA-seq coverage to determine if the average at actual CID boundaries was significantly higher than if the CIDs were randomly assigned (Table 2.3). We also compared our interaction propensity data to the CID locations. We noticed that locations near the center of CIDs often had a low interaction propensity (Figure 2.3D), suggesting the bottle-brush type model in which

extruded loops might have a region of low interactions at their apex. We summed the distances from the bins with the ten lowest interaction frequencies in each sample to the nearest CID center and determined that this was significantly lower than random arrangement of the interaction propensity relative to the CID boundaries (Table 2.4). A similar pattern of CIDs, and tight correlation between CID centers and low interaction propensity was apparent in data from the distantly related *E. coli* Crooks strain (Figure 2.3E).

#### *rRNA clustering is recapitulated in our data*

Previous fluorescence-based microscopy studies have identified clustering of all rRNA operons except *rrnC* (Gaal et al., 2016). To check the relevance of our data to the situation *in vivo*, we wanted to investigate if rRNA operon clustering was apparent in our contact matrices. Given the repetition of the rRNA operons seven times, aligning reads to one specific copy of the operon rather than another is challenging. If the rRNA operons are spatially clustered, we assumed that the DNA neighboring the rRNA operons would also be clustered. Therefore, we assessed whether the distribution of percentile scores for the interaction frequency of all combinations of bins that neighbor rRNA operons was non-uniform and right-shifted. The percentile scores of randomly selected locations on the contact matrix should form a uniform distribution. Visually, the distribution of many of our samples appeared to be right-shifted (Figure 2.4A). The same pattern, though slightly weaker, was clear in the Crooks data, indicating that the rRNA operon clustering is not specific to lab strains (Figure 2.4B). We performed a permutation test by comparing the true degree of right-shifting to the right-shifting among percentile scores of interaction frequencies at every other potential combination of bins that maintains the correct distance between rRNA neighbors. Results from the permutation tests are in Table 2.5. We performed a permutation test in which only the six operons identified as

clustering *in vivo* were included, as well as a permutation test in which the *rrnC* neighbors were included. Unlike the fluorescence data, our data do support co-clustering of *rrnC* with the other rRNA operons, and indeed clustering of the rRNA operons in our data is only statistically significant upon inclusion of *rrnC* neighbors. Gaal and colleagues propose that *rrnC* may not cluster with the other operons due to its proximity to the origin, whose own spatial organization may override the rRNA localization. Given the fact that their experiments are performed in rich media and ours are performed in minimal media, there may be enough differences in the frequency of origin firing that *rrnC* clustering with the other rRNA operons is apparent in our data but not in theirs.

In addition, unlike the Gaal experiments, in which rRNA clustering appears to be transcription independent, we see attenuation of the rRNA clustering at later time points (Figure 2.4C and Table 2.5), suggesting that transcription may play a role. We investigated this by considering the clustering of five randomly chosen, highly expressed genes as well as clusters of related highly expressed genes; clusters included outer membrane/secreted proteins, translation factors and metabolic enzymes. In all combinations, we do not have enough evidence to indicate that the genes are colocalized (Table 2.6). This suggests that while transcription may play a role in the clustering of the rRNA operons, it is not only the fact that those regions are highly transcribed that leads to their clustering.

#### *Patterns of interaction propensities are generally shared across samples*

Given that the inherent interaction propensities of various regions seemed to play a substantial role in chromosome organization, we compared the interaction propensities between conditions and strains. In considering the WT cells, we note that the replicates at log phase show more variation compared to the replicates at late stationary phase (Figure 2.5A, B). Given the

much higher level of transcription that occurs during log phase, this again suggests that transcription may play a role in chromosome architecture. Overall, there are few differences over the studied time course, though it appears that the range of interaction propensities may be enlarged during later stages of growth (Figure 2.5C). It is also important to note that the locations of the rRNA operons (marked by blue vertical lines in Figure 2.5C) are not regions of high interaction propensity; the interaction propensity percentile score of the bin to either side of the seven rRNA operons ranges from the 32<sup>nd</sup> to the 96<sup>th</sup> percentile, with a median at the 63<sup>rd</sup> percentile. Finally, despite several large genome rearrangements relative to the MG1655 strain used as our WT, the Crooks strain shows highly similar interaction propensities in a long region of alignment (Figure 2.5D).

However, we do see some significant changes in certain of the mutant strains. In the *lrp* KO, one location in the terminus exhibits dramatic deviation from the WT (Figure 2.6A). In the double KO of both DNA methyltransferases (*dam* and *dcm*), there is an extended region near the terminus that displays increased interactions in the KO relative to the WT (Figure 2.6B). This change is dependent on *dam* KO, since the *dcm* KO alone does not display that extended alteration (Figure 2.6B,C). Methylation of GATC sites by Dam is known to alter DNA structure and thus potentially affect protein-DNA binding (Polaczek, Kwan, & Campbell, 1998). Therefore, loss of Dam activity may alter the protein binding landscape in the terminus region that perhaps prevented a higher level of DNA-DNA interactions. Intriguingly, the *dcm* KO alone exhibits more genome-wide variation relative to WT, both in terms of increasing and decreasing interaction propensities.

*Regions of high and low interaction propensity are enriched for different classes of genes*

If the interaction propensity of a DNA region is an inherent property, one contributing factor might be the genes contained within that region. Therefore, we investigated if certain classes of genes were enriched in regions with low or high interaction propensity. To do this, we assigned each gene to a bin based on the center of the annotated coding region, and then assigned the bin's corresponding interaction propensity to each matched gene. Using iPAGE (Goodarzi, Elemento, & Tavazoie, 2009), we queried for enrichment or depletion of certain gene ontology (GO) terms upon splitting the range of interaction propensities into five levels. While many enrichments appear in only one or two conditions, several are well conserved across many, if not all, of our replicates (Figure 2.7). Enriched GO-terms for genes with high interaction propensity scores include histidine and thiamine biosynthetic processes and ethanolamine catabolic processes. For each of these GO-terms, most genes annotated to that GO-term are located in one location along the genome, so it is not surprising that they would all follow the same pattern for high or low interaction propensity. However, we are not able to determine whether something about that gene class leads to high interaction propensities.

Likewise, there are unique enriched GO-terms for genes with low interaction propensity, including self-proteolysis (in all conditions but late stationary phase) and transposase activity. The self-proteolysis GO-term cluster includes a limited number of genes, many of which are putative or computationally predicted, so it is challenging to ascribe a mechanistic hypothesis to this example. However, genes with transposase activity represent an interesting example. Since transposases are responsible for moving transposons to new locations, it may be logical for regions encoding those genes to be kept isolated from interacting with much other DNA in order to limit their potential activity.

### *Interaction propensity is globally correlated with several other features*

In order to identify potential causative factors for chromosome organization, we compared the interaction propensity from our 3C-seq experiments to a variety of other genome-wide characteristics or data-sets: methylation sites, AT content, prophage content (Keseler et al., 2017), degree of supercoiling (Lal et al., 2016), ChIP of several NAPs (Kahramanoglou et al., 2011; Prieto et al., 2012), RNA-seq, transcription propensity (Scholz, Diao, Fivenson, Lin & Freddolino, in preparation), and total protein occupancy (Goss & Freddolino, in preparation). In order to compare these data sets to our interaction propensity statistic, which is at the 20 kb resolution of our bin sizes, we employed a custom sliding window script (developed by Michael Wolfe). Given a set of input data and the set of window coordinates that we want to compare the input data against (here, our non-overlapping 20 kb bins), this program generates an output for each window; the output being a median value if the input is a continuous signal, the number of sites for methylation site regular expressions, the percent AT-content, or the percent of the window containing annotated prophage genes. We assessed the relatedness of the interaction propensity to each of these factors using a Spearman correlation, implemented with the `scipy.stats` module.

We observe strong positive correlations with Dam-specific methylation sites, transcription propensity, and RNA-seq, and weaker correlations with Dcm-specific methylation sites (Table 2.7). On the other hand, we observe strong negative correlations with AT-content, H-NS binding sites (as identified by ChIP), prophage density and total protein occupancy. Transcription propensity (see Chapter 1) represents the likelihood of transcription of a reporter upon random insertion, so, like the positive correlation with RNA-seq, suggests that frequently transcribed regions generally have more interactions with other DNA. This again presents

support for active transcription playing a major role in chromosome architecture. Since H-NS is a known repressive protein, a negative correlation there also suggests that transcription contributes to DNA organization. The negative correlation between prophage density and interaction propensity agrees with the iPAGE result showing enrichment for genes with transposase activity in regions of low interaction propensity. Bacterial cell survival could be aided by isolating prophage genes, especially if they are thus kept away from areas of active transcription. Thus, it seems that many of these correlations could be explained by identifying a connection between high DNA-DNA interaction levels and high transcription. The case of the Dam sites is somewhat unclear since a correlation between sites and interaction propensity still appears in the *dam* KO. If Dam site methylation was contributing to a certain level of interaction propensity, we would expect the correlation to disappear upon *dam* KO. The pattern that we observe suggests that some other unknown factor may act at GATC sites and thus contribute to higher interaction propensities. While we are only able to compare sequence-based factors for the Crooks data, we again document a positive correlation between interaction propensity and Dam sites and a negative correlation with AT-content.

## **Discussion**

### *The E. coli chromosome does not form an ellipsoid or strong macrodomains*

Previous global organizations proposed for bacterial chromosomes include the formation of an ellipsoid (Umbarger et al., 2011) and the presence of macrodomains (Valens et al., 2004). The results of our studies clearly show the lack of the off-diagonal signal that is characteristic of ellipsoid organization. The absence of this organization agrees with earlier microscopy work in *E. coli* (Nielsen et al., 2006; Wu et al., 2018). In addition, a previous study analyzed bacterial

genomes for regions of similarity and found that some bacteria, in addition to showing alignments corresponding to tandem duplications, had regions of homology at equivalent distances from the origin (Eisen, Heidelberg, White, & Salzberg, 2000). In the ellipsoid model, regions at equivalent distances from the origin would be spatially close to each other due to the twisting of the chromosomal arms together. Thus, an ellipsoid organization would favor formation of gene duplications on the opposite arm of the chromosome but at an equivalent distance from the origin, just what is seen in the alignment study for some species such as *Vibrio cholera* and *Streptococcus pyogenes*. Evidence for such diagonal alignments is not seen in *E. coli*; rather, the distribution of alignments is quite random, potentially suggesting a more fluid genomic organization.

We also do not document strong evidence for the presence of macrodomains, only having evidence to support the presence of the left MD. The initial study documenting macrodomains in *E. coli* measured recombination efficiency between locations on the genome and from there established the locations for MDs, within which recombination (contact) was possible compared to regions outside the MD, for which recombination levels were negligible (Valens et al., 2004). 3C-sequencing experiments performed by the same group (Lioy et al., 2018) showed weak support for four distinct MDs, but did document considerable isolation of the *ter* region. It may be that the varying techniques contribute to different results. If MDs exist but can be packed variably, both within the MD, and for all regions together, it is possible that pieces of DNA on the outside of the MDs in individual cells would be able to interact with DNA in other MDs (as seen in our data). Therefore, MD signals would be obscured on a population level. Basic coarse-grained simulations (performed by Peter Freddolino) of a polymer model of DNA based

on parameters from our interaction propensity data showed the colocalization of four distinct MD regions, even though we do not see evidence for it in the contact matrices (Figure 2.8).

### *Interaction propensity-based model for chromosome organization*

We observe strong consistency in the interaction propensity at different time points and even when compared across different strains. This suggests that the interaction propensity plays a fundamental role in genome organization. Strains bearing knockouts of either *dam* or *lrp* exhibit some significant differences near the *ter* region, indicating that some of the isolation of the *ter* region may be due to processes involving those proteins. For example, protein binding that is selective for methylated DNA may prevent interactions that become possible in the *dam* knockout. Although most Dam sites are thought to be methylated, there is evidence for epigenetic regulation mechanisms (Casadesús & Low, 2013; Stacey N. Peterson & Reich, 2008) and specific un-methylated sites (Cohen et al., 2016) that would allow for some *ter*-specific role of methylation.

The correlation between CID-centers and low interaction propensity regions suggests support for a ‘bottle-brush’ type model in which each CID is a loop of DNA with a more isolated region in the center of the loop (Le et al., 2013). Highly transcribed regions are also more likely to have a higher interaction propensity, and total protein occupancy is generally correlated with low interaction propensity. This suggests a model in which actively transcribed regions are more likely to make interactions with other regions of DNA, while transcriptionally silent regions, such as prophages or genes with transposase activity, are isolated from other regions of DNA, evoking parallels to eukaryotic euchromatin and heterochromatin. Interestingly, the rRNA

operons are not in areas of high interaction propensity, indicating that there are other levels of organization that control their clustering.

## **Materials and Methods**

### *Genome Conformation Capture*

Cells were cross-linked by adding formaldehyde (37% Sigma-Aldrich; St. Louis, MO) to 1% (vol/vol) and incubated with shaking for 30 minutes at room temperature. Formaldehyde cross-linking was neutralized by addition of Tris (pH 8) to a final concentration of 280 mM and incubation with shaking at room temperature for 10 minutes. The culture was then immediately centrifuged for 5 minutes at 5500xg at 4°C. The pellet was washed twice with 30 mL ice cold PBS before being resuspended in 1 mL PBS. Following a 3 minute centrifugation at 10,000xg at 4°C, the pellet was flash-frozen in a dry ice/ethanol bath and then stored at -80°C. Two biological replicates were performed for each condition.

The sample was thawed on ice and resuspended in lysis buffer (50% (vol/vol) Bacterial Protein Extraction Reagent II (Thermo Fisher; Waltham, MA), 1x Complete Mini EDTA-free Protease Inhibitors (Roche; Basel, Switzerland), 53 kilounits Ready-Lyse Lysozyme Solution (Epicentre; Madison, WI) and incubated for 15 minutes at 37°C. The resulting lysate was mixed with 3 mL 10 mM TEE containing 0.003% Triton X-100, applied to an Amicon Ultra-4 Centrifugal Filter Unit (EMD Millipore; Billerica, MA) and centrifuged for 30 min at 3124xg in a hanging bucket rotor. After addition of 3 mL TEE, the sample was centrifuged for an additional 25-30 minutes at 3214xg. The resulting concentrate was digested with 80 units HhaI in CutSmart Buffer (NEB; Ipswich, MA) in a total volume of 500 µL for 60 minutes at 37°C. A digest control plasmid spike in of 100 ng was also included. Following heat inactivation by

incubating at 65°C for 20 min, the sample was ligated using 1.25x T4 DNA ligase buffer and 800 units T4 DNA ligase (NEB; Ipswich, MA) at 16°C for 60 minutes. Then the sample was incubated overnight at 65°C.

The sample was treated with 0.1 mg RNase A (Thermo Fisher; Waltham, MA) for 2 hours at 37°C, then 0.4 mg Proteinase K (Thermo Fisher; Waltham, MA) for 2 hours at 50°C before the DNA was isolated by phenol-chloroform extraction and isopropanol precipitation. The DNA was further purified by using the Oligo Clean & Concentrator Kit (Zymo Research; Irvine, CA). Then, 1 µg of DNA was digested with 5 units AluI in CutSmart Buffer (NEB; Ipswich, MA) for 10 minutes at 37°C and cleaned up with DNA Clean & Concentrator 5 (Zymo Research; Irvine, CA). The samples were quantified (Quant-iT PicoGreen dsDNA Assay Kit, Thermo Fisher; Waltham, MA) and prepared for sequencing using the NEBNext DNA Library Prep Kit for Illumina (NEB; Ipswich, MA). The library was checked for quality by 2% agarose gel electrophoresis using GelRed stain (Biotium; Fremont, CA). Samples were pooled and the sequencing performed on an Illumina NextSeq -500, with paired end reads.

### *Strains and media*

All mutant strains (Table 2.8) were derived from *E. coli* K-12 MG1655 (ATCC 47076). Routine growth for cloning was done in LB medium (10 g/liter tryptone, 5 g/liter yeast extract, 5 g/liter NaCl) or on LB plates (LB medium plus 15 g/liter agar) supplemented with 50 µg/mL kanamycin or 100 µg/mL ampicillin (both from USBiological; Salem, MA) as required. Gene knock outs for *dam* and *dcm* were produced using P1 vir phage transduction from Keio collection strains (Baba et al., 2006). First, the donor strain containing the mutation of interest, either *dam::kanR* or *dcm::kanR*, was grown overnight and diluted 100-fold into 5 mL fresh LB

medium supplemented with 0.2% glucose and 5 mM CaCl<sub>2</sub>. After sixty minutes of growth, 70 μL of a previous P1 vir lysate was added before returning the culture to 37°C. The culture was grown until it appeared cloudy and then cleared, indicating that we had a population of phage containing fragments from our donor strain. Second, the recipient cells (PLF 308) were grown overnight in LB medium, and 2 mL culture was pelleted. The cells were resuspended in 1 mL LB supplemented with 10 mM MgSO<sub>4</sub> and 5 mM CaCl<sub>2</sub>, and 200 μL of the resulting solution was added to 100 μL of the prepared donor phage lysate. After 30 minutes of growth at 30°C with gentle shaking, a further 1 mL LB, supplemented with 7.7 mM sodium citrate was added followed by 30 minutes of growth at 37°C without shaking. The cells were pelleted by centrifugation, and then resuspended in 100 μL 1 M sodium citrate before being plated on appropriate selective media and grown at 37°C. To remove the residual phage, the resulting colonies were replica plated onto another selective plate and harvested for validation and storage from there. Marker removal proceeded as described below. To create the double knock out, GMK052 was used as the recipient strain for transduction with phage containing the *dcm::kanR* marker. Marker removal was then performed. The *lrp* and *maoP* deletion strains were constructed by homologous recombination resulting in the insertion of a kanamycin resistance cassette (Datsenko & Wanner, 2000). To remove the kan cassette marker, the pcp20 plasmid was transformed into the marker containing strain, and the transformants were grown on LB/Ampicillin plates at 30°C. A liquid culture from a single colony was grown at 42°C, and then plated at 37°C on LB. Scar formation was tested by plating candidate deletion strains on LB/Ampicillin at 30°C, LB/Kanamycin at 37°C, and LB at 37°C. Strains with a proper scar and plasmid elimination only grown on LB. All primers used for mutant strain construction and

validation are listed in Table 2.9. Mutant strains were confirmed by PCR product sizing and Sanger sequencing.

For each experimental replicate, an overnight culture grown from a single colony in m9 media was diluted to OD600=0.003 in 100 mL of fresh, pre-warmed m9 medium (1x m9 salts, 0.2% glucose, 2 mM MgSO<sub>4</sub>, 0.1 mM CaCl<sub>2</sub>, 1x MOPS micronutrients, 0.01 mM ferric citrate). The cells were grown at 37°C with shaking (200 rpm) until OD600 was between 0.2 and 0.3 (for log phase samples) or for 24 hours (long stationary phase samples). For stationary phase samples, upon reaching OD600 of 0.2-0.3, cells were pelleted by centrifugation for 3 minutes at 17,000xg. The pellet was resuspended in 100 mL 1x m9 salts and incubated at room temperature for 3 hours before continuing with the cross-linking procedure.

#### *Sequencing data analysis pipeline*

Sequencing analysis was performed by Peter Freddolino and Catherine Barnier. The forward and reverse reads for each paired end read were independently aligned to the U00096.3 MG1655 or ATCC8739 *E. coli* (Crooks) genome as appropriate. If the independent alignments for the reads in each pair were greater than 2 kb apart, we defined the read as an indirect fusion and calculated a score as follows:

$$Score = \frac{1}{P_{Fwd} \times P_{Rvs}}$$

(where P is the number of possible alignments made by the forward or reverse read)

The score was added to the contact matrix at the locations (x,y) and (y,x), where x is the 20 kb bin number in which the forward read aligns, and y is the 20kb bin number in which the reverse read aligns. Scores were cumulatively added to the contact matrix for all indirect fusions to yield a final interaction frequency score at each position.

The resulting contact matrix was normalized for copy number variation as follows. A smoothing function was fit to the abundance of any reads that were not counted as indirect fusions across the length of the genome, and this was assumed to be an estimate of genomic abundance. The interaction frequency score at each location was normalized for abundance as follows:

$$\text{Normalized interaction frequency } (x, y) = \frac{\text{interaction frequency}(x, y)}{A_x \times A_y}$$

(where  $A_x$  and  $A_y$  represent the normalized abundance for bin  $x$  and bin  $y$ , respectively)

### *Computational analysis*

Where noted in the text, we used permutation tests, which were implemented using custom python scripts and 1000 permutations, or the greatest possible combinations given our data ( $r=232$  or  $r=237$ ). We corrected for multiple hypothesis testing using the `statsmodels.sandbox.stats.multicomp.multipletests` module using the Benjamini-Hochberg method (Hochberg & Benjamini, 1990; Seabold & Perktold, 2010).

In order to test for variations in interaction frequency that would mark the presence of MDs, we slid a square identical in size to the MD along the diagonal of the 3C-seq interaction frequency matrices. For a starting position at each bin position along the genome (i.e. every 20 kb), we determined the median interaction frequency. We took the median of all 232 potential squares and calculated the magnitude of the difference between the square's median and the median of the medians (absolute distance). We eliminated any squares starting within 10 bins of the proposed MD location, and then calculated how many of the remaining random squares had an absolute distance greater than the proposed MD location. From there we obtained a p-value. We also tested if the difference between the median interaction frequency within a potential MD

square and the median interaction frequency of interactions between the MD and regions outside the MD was significantly larger than random locations.

The directional index (DI) analysis was performed as in (Le et al., 2013). In summary, for each 20 kb bin, we pulled vectors of interaction frequencies between that bin and neighboring bins either upstream or downstream for a total of 100 kb. The upstream and downstream vectors were compared by a paired t-test to determine if the bin showed a bias towards interacting with DNA upstream or downstream. T-statistics greater than a magnitude of 2 were capped at 2 or -2.

To identify the locations of CID boundaries, we smoothed the t-statistic obtained in the DI analysis by taking a rolling median at every 5 bins. CID boundaries should occur where there is a switch from a negative t-statistic (indicating upstream bias) to a positive t-statistic (indicating downstream bias). A bin was marked as a CID boundary if its t-statistic was less than 0 and the three subsequent bins had t-statistics greater than 0. This requirement is slightly more stringent than just requiring a negative to positive switch. Based on visual inspection of the DI analyses, we determined that the more restrictive cut-off eliminated some CID boundaries that only occurred because of a one or two bin switch in the t-statistic and which did not appear to represent true CIDs.

All plots were created using Matplotlib (Hunter, 2007).

## Tables

Condition	Ter		Ori		Right		Left	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Log1	1	1	0.769	0.939	0.175	0.313	0.009	0.012
Log2	1	1	0.656	0.939	0.170	0.313	0.005	0.008
Log3	0.467	1	0.797	0.939	0.344	0.366	0.005	0.008
Stat1	0.854	1	0.557	0.939	0.189	0.313	0.042	0.048
Stat2	0.972	1	0.675	0.939	0.127	0.313	0.005	0.008
LongStat1	1	1	0.288	0.939	0.142	0.313	0.052	0.055
LongStat2	1	1	0.250	0.939	0.075	0.313	0.005	0.008
<i>dam</i> KO 1	0.358	1	0.778	0.939	0.778	0.778	0.061	0.061
<i>dam</i> KO 2	0.627	1	0.920	0.939	0.156	0.313	0.005	0.008
<i>dcm</i> KO 1	0.5	1	0.717	0.939	0.203	0.313	0.005	0.008
<i>dcm</i> KO 2	0.675	1	0.830	0.939	0.108	0.313	0.005	0.008
<i>dam/dcm</i> KO 1	0.693	1	0.887	0.939	0.330	0.366	0.005	0.008
<i>dam/dcm</i> KO 2	0.703	1	0.844	0.939	0.259	0.315	0.005	0.008
<i>lrp</i> KO 1	0.759	1	0.675	0.939	0.198	0.313	0.038	0.046
<i>lrp</i> KO 2	0.509	1	0.939	0.939	0.099	0.313	0.009	0.012
<i>maoP</i> KO 1	0.613	1	0.854	0.939	0.259	0.315	0.009	0.012
<i>maoP</i> KO 2	0.594	1	0.840	0.939	0.222	0.314	0.005	0.008

**Table 2.1: Traditional macrodomain permutation test results.** Results of permutation test for enrichment of higher median interaction frequency within a macrodomain as compared to the median interaction frequency for interactions to DNA outside the macrodomain. Separate permutation tests were performed for each proposed macrodomain.

Condition	Ter		Ori		Right		Left	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Log1	0.004	0.024	0.682	0.773	0.352	0.499	0.103	0.117
Log2	0.052	0.219	0.884	0.884	0.481	0.545	0.004	0.008
Log3	0.794	0.837	0.498	0.773	0.232	0.499	0.009	0.013
Stat1	0.176	0.499	0.854	0.884	0.472	0.545	0.021	0.030
Stat2	0.133	0.452	0.670	0.773	0.318	0.499	0.004	0.008
LongStat1	0.004	0.024	0.644	0.773	0.343	0.499	0.137	0.137
LongStat2	0.004	0.024	0.567	0.773	0.330	0.499	0.116	0.123
<i>dam</i> KO 1	0.837	0.837	0.451	0.773	0.631	0.631	0.030	0.039
<i>dam</i> KO 2	0.498	0.837	0.343	0.773	0.240	0.499	0.004	0.008
<i>dcm</i> KO 1	0.811	0.837	0.618	0.773	0.202	0.499	0.004	0.008
<i>dcm</i> KO 2	0.545	0.837	0.455	0.773	0.167	0.499	0.004	0.008
<i>dam/dcm</i> KO 1	0.639	0.837	0.403	0.773	0.528	0.561	0.004	0.008
<i>dam/dcm</i> KO 2	0.382	0.837	0.429	0.773	0.395	0.516	0.004	0.008
<i>lrp</i> KO 1	0.403	0.837	0.678	0.773	0.275	0.499	0.034	0.042
<i>lrp</i> KO 2	0.725	0.837	0.227	0.773	0.245	0.499	0.004	0.008
<i>maoP</i> KO 1	0.687	0.837	0.506	0.773	0.330	0.499	0.009	0.013
<i>maoP</i> KO 2	0.721	0.837	0.403	0.773	0.283	0.499	0.004	0.008

**Table 2.2: Modified macrodomain permutation test results.** Results of permutation test for enrichment of a median interaction frequency within a macrodomain that is significantly different than the median interaction frequency of all potential equivalently sized regions along the chromosome. Separate permutation tests were performed for each proposed macrodomain.

Condition	p-value	q-value
Log1	0.154	0.477
Log2	0.706	0.800
Log3	0.117	0.477
Stat1	0.246	0.477
Stat2	0.181	0.477
LongStat1	0.912	0.912
LongStat2	0.274	0.477
<i>dam</i> KO 1	0.587	0.800
<i>dam</i> KO 2	0.370	0.571
<i>dcm</i> KO 1	0.073	0.477
<i>dcm</i> KO 2	0.239	0.477
<i>dam/dcm</i> KO 1	0.281	0.477
<i>dam/dcm</i> KO 2	0.074	0.477
<i>lrp</i> KO 1	0.643	0.800
<i>lrp</i> KO 2	0.670	0.800
<i>maoP</i> KO 1	0.833	0.885
<i>maoP</i> KO 2	0.165	0.477

**Table 2.3: Results of permutation test for enrichment of highly transcribed regions at CID boundaries.**

Condition	p-value	q-value
Log1	0.024	0.041
Log2	0.005	0.011
Log3	0.005	0.011
Stat1	0.061	0.069
Stat2	0.005	0.011
LongStat1	0.005	0.011
LongStat2	0.005	0.011
<i>dam</i> KO 1	0.042	0.062
<i>dam</i> KO 2	0.019	0.036
<i>dcm</i> KO 1	0.354	0.354
<i>dcm</i> KO 2	0.005	0.011
<i>dam/dcm</i> KO 1	0.061	0.069
<i>dam/dcm</i> KO 2	0.052	0.066
<i>lrp</i> KO 1	0.005	0.011
<i>lrp</i> KO 2	0.250	0.264
<i>maoP</i> KO 1	0.052	0.066
<i>maoP</i> KO 2	0.005	0.011
Crooks1	0.018	0.036
Crooks2	0.037	0.058

**Table 2.4: Results of permutation test for enrichment of low interaction propensity at CID centers.**

Condition	All 7 rRNA operons		Excluding <i>rrnC</i>	
	p-value	q-value	p-value	q-value
Log1	0.009	0.052	0.021	0.135
Log2	0.318	0.342	0.365	0.388
Log3	0.039	0.073	0.129	0.199
Stat1	0.017	0.052	0.030	0.135
Stat2	0.232	0.281	0.296	0.360
LongStat1	0.378	0.378	0.455	0.455
LongStat2	0.322	0.342	0.339	0.384
<i>dam</i> KO 1	0.017	0.052	0.064	0.137
<i>dam</i> KO 2	0.116	0.152	0.240	0.314
<i>dcm</i> KO 1	0.009	0.052	0.013	0.135
<i>dcm</i> KO 2	0.013	0.052	0.034	0.135
<i>dam/dcm</i> KO 1	0.021	0.052	0.052	0.135
<i>dam/dcm</i> KO 2	0.073	0.103	0.047	0.135
<i>lrp</i> KO 1	0.021	0.052	0.073	0.138
<i>lrp</i> KO 2	0.026	0.055	0.056	0.135
<i>maoP</i> KO 1	0.060	0.093	0.180	0.255
<i>maoP</i> KO 2	0.043	0.073	0.086	0.146
Crooks1	0.050	0.101	NA	NA
Crooks2	0.113	0.113	NA	NA

**Table 2.5: rRNA operon clustering permutation test results.** Testing either included all seven (left) or excluded *rrnC* (right).

Condition	Translation factors		Secreted/membrane proteins		Metabolic enzymes	
	p-value	q-value	p-value	q-value	p-value	q-value
Log1	1	1	0.854	0.974	0.923	1
Log2	1	1	0.974	0.974	0.948	1
Log3	0.498	1	0.421	0.974	0.777	1
Stat1	0.867	1	0.039	0.657	0.923	1
Stat2	1	1	0.914	0.974	0.996	1
LongStat1	1	1	0.927	0.974	0.871	1
LongStat2	1	1	0.936	0.974	0.888	1
<i>dam</i> KO 1	0.815	1	0.395	0.974	0.944	1
<i>dam</i> KO 2	1	1	0.815	0.974	0.751	1
<i>dcm</i> KO 1	0.554	1	0.266	0.974	0.948	1
<i>dcm</i> KO 2	0.828	1	0.442	0.974	0.760	1
<i>dam/dcm</i> KO 1	0.803	1	0.798	0.974	0.923	1
<i>dam/dcm</i> KO 2	1	1	0.725	0.974	0.987	1
<i>lrp</i> KO 1	1	1	0.807	0.974	0.773	1
<i>lrp</i> KO 2	0.841	1	0.858	0.974	0.996	1
<i>maoP</i> KO 1	1	1	0.751	0.974	0.464	1
<i>maoP</i> KO 2	1	1	0.502	0.974	1	1

**Table 2.6: Highly expressed, related gene clustering permutation test results.**

Condition		Dam Sites	Dcm Sites	AT-content	Supercoiling	HU ChIP
Log1	Spearman $\rho$	0.303	0.143	-0.559	-0.107	0.022
	p-value	2.5E-06	2.9E-02	1.9E-20	1.0E-01	7.3E-01
Log2	Spearman $\rho$	0.295	0.129	-0.705	-0.226	-0.170
	p-value	5.0E-06	4.9E-02	4.1E-36	5.3E-04	9.6E-03
Log3	Spearman $\rho$	0.217	0.106	-0.318	-0.055	0.122
	p-value	8.9E-04	1.1E-01	7.7E-07	4.0E-01	6.4E-02
Stat1	Spearman $\rho$	0.295	0.140	-0.451	-0.076	0.052
	p-value	4.9E-06	3.3E-02	4.8E-13	2.5E-01	4.3E-01
Stat2	Spearman $\rho$	0.324	0.130	-0.723	-0.237	-0.221
	p-value	4.4E-07	4.8E-02	8.6E-39	2.8E-04	6.9E-04
LongStat1	Spearman $\rho$	0.302	0.136	-0.729	-0.241	-0.262
	p-value	2.8E-06	3.9E-02	1.0E-39	2.1E-04	5.4E-05
LongStat2	Spearman $\rho$	0.295	0.127	-0.728	-0.255	-0.279
	p-value	4.7E-06	5.4E-02	1.5E-39	8.8E-05	1.6E-05
<i>dam</i> KO 1	Spearman $\rho$	0.221	0.119	-0.381	-0.089	0.045
	p-value	6.9E-04	6.9E-02	1.9E-09	1.8E-01	4.9E-01
<i>dam</i> KO 2	Spearman $\rho$	0.273	0.135	-0.602	-0.203	-0.152
	p-value	2.5E-05	4.0E-02	3.1E-24	1.9E-03	2.0E-02
<i>dcm</i> KO 1	Spearman $\rho$	0.196	0.081	-0.275	-0.027	0.153
	p-value	2.7E-03	2.2E-01	2.1E-05	6.8E-01	2.0E-02
<i>dcm</i> KO 2	Spearman $\rho$	0.243	0.107	-0.407	-0.074	0.076
	p-value	1.8E-04	1.0E-01	1.2E-10	2.6E-01	2.5E-01
<i>dam/dcm</i> KO 1	Spearman $\rho$	0.214	0.104	-0.376	-0.083	0.026
	p-value	1.0E-03	1.2E-01	3.4E-09	2.1E-01	6.9E-01
<i>dam/dcm</i> KO 2	Spearman $\rho$	0.279	0.106	-0.563	-0.176	-0.108
	p-value	1.6E-05	1.1E-01	9.0E-21	7.1E-03	1.0E-01
<i>lrp</i> KO 1	Spearman $\rho$	0.268	0.114	-0.519	-0.125	-0.035
	p-value	3.5E-05	8.3E-02	2.1E-17	5.8E-02	5.9E-01
<i>lrp</i> KO 2	Spearman $\rho$	0.246	0.104	-0.436	-0.120	0.009
	p-value	1.6E-04	1.1E-01	3.4E-12	6.9E-02	8.9E-01
<i>maoP</i> KO 1	Spearman $\rho$	0.221	0.113	-0.364	-0.079	0.093
	p-value	7.0E-04	8.6E-02	1.1E-08	2.3E-01	1.6E-01
<i>maoP</i> KO 2	Spearman $\rho$	0.261	0.146	-0.421	-0.093	0.065
	p-value	5.6E-05	2.6E-02	2.3E-11	1.6E-01	3.3E-01
Crooks1	Spearman $\rho$	0.404	0.109	-0.506	NA	NA
	p-value	1.0E-10	9.5E-02	8.6E-17	NA	NA
Crooks2	Spearman $\rho$	0.390	0.122	-0.449	NA	NA
	p-value	4.9E-10	6.0E-02	3.9E-13	NA	NA

Condition		H-NS ChIP	Fis ChIP	Prophage Density	Transcription Propensity	Total Protein Binding
Log1	Spearman $\rho$	-0.400	0.194	-0.177	0.267	-0.479
	p-value	2.6E-10	3.1E-03	6.8E-03	3.8E-05	9.8E-15
Log2	Spearman $\rho$	-0.435	0.138	-0.233	0.277	-0.592
	p-value	3.9E-12	3.5E-02	3.5E-04	1.9E-05	2.6E-23
Log3	Spearman $\rho$	-0.363	-0.021	-0.067	0.142	-0.228
	p-value	1.3E-08	7.5E-01	3.1E-01	3.1E-02	4.7E-04
Stat1	Spearman $\rho$	-0.384	0.113	-0.116	0.220	-0.378
	p-value	1.5E-09	8.5E-02	7.8E-02	7.6E-04	2.7E-09
Stat2	Spearman $\rho$	-0.485	0.087	-0.229	0.266	-0.589
	p-value	4.2E-15	1.8E-01	4.5E-04	4.1E-05	4.9E-23
LongStat1	Spearman $\rho$	-0.425	0.139	-0.259	0.275	-0.627
	p-value	1.4E-11	3.4E-02	6.7E-05	2.1E-05	8.5E-27
LongStat2	Spearman $\rho$	-0.410	0.137	-0.257	0.267	-0.624
	p-value	8.2E-11	3.7E-02	7.3E-05	3.7E-05	1.9E-26
<i>dam</i> KO 1	Spearman $\rho$	-0.364	0.028	-0.085	0.157	-0.292
	p-value	1.2E-08	6.7E-01	2.0E-01	1.7E-02	6.0E-06
<i>dam</i> KO 2	Spearman $\rho$	-0.423	0.037	-0.214	0.202	-0.488
	p-value	1.8E-11	5.7E-01	1.0E-03	2.0E-03	2.9E-15
<i>dcm</i> KO 1	Spearman $\rho$	-0.337	-0.001	-0.044	0.152	-0.192
	p-value	1.4E-07	9.9E-01	5.0E-01	2.1E-02	3.4E-03
<i>dcm</i> KO 2	Spearman $\rho$	-0.392	0.029	-0.096	0.193	-0.310
	p-value	6.0E-10	6.6E-01	1.5E-01	3.2E-03	1.5E-06
<i>dam/dcm</i> KO 1	Spearman $\rho$	-0.347	0.048	-0.081	0.162	-0.299
	p-value	5.7E-08	4.7E-01	2.2E-01	1.3E-02	3.5E-06
<i>dam/dcm</i> KO 2	Spearman $\rho$	-0.422	0.055	-0.167	0.202	-0.453
	p-value	1.9E-11	4.0E-01	1.1E-02	2.0E-03	4.1E-13
<i>lrp</i> KO 1	Spearman $\rho$	-0.431	0.031	-0.128	0.216	-0.404
	p-value	6.4E-12	6.3E-01	5.1E-02	9.0E-04	1.5E-10
<i>lrp</i> KO 2	Spearman $\rho$	-0.421	-0.011	-0.067	0.181	-0.323
	p-value	2.3E-11	8.7E-01	3.1E-01	5.8E-03	5.1E-07
<i>maoP</i> KO 1	Spearman $\rho$	-0.351	0.009	-0.087	0.159	-0.273
	p-value	4.0E-08	8.9E-01	1.9E-01	1.5E-02	2.4E-05
<i>maoP</i> KO 2	Spearman $\rho$	-0.390	0.007	-0.126	0.197	-0.331
	p-value	7.4E-10	9.2E-01	5.5E-02	2.6E-03	2.4E-07

**Table 2.7: Genome-wide feature correlations.** Results of testing for correlation between interaction propensity scores and other genome-wide features.

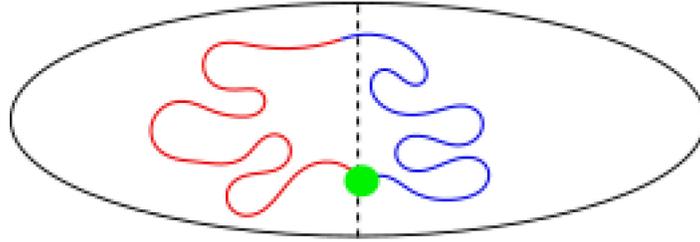
Strain name	Parental Strain	Genomic Modification	Origins
PLF308	MG1655		ATCC 47076
PLF007	Crooks		ATCC 8739
GMK052	MG1655	$\Delta dam$	This study
GMK056	MG1655	$\Delta dcm$	This study
GMK058	MG1655	$\Delta dam, \Delta dcm$	This study
GMK009	MG1655	<i>lrp::kanR</i>	This study
GMK057	MG1655	$\Delta maoP$	This study

**Table 2.8: Genotype of strains used in this study.**

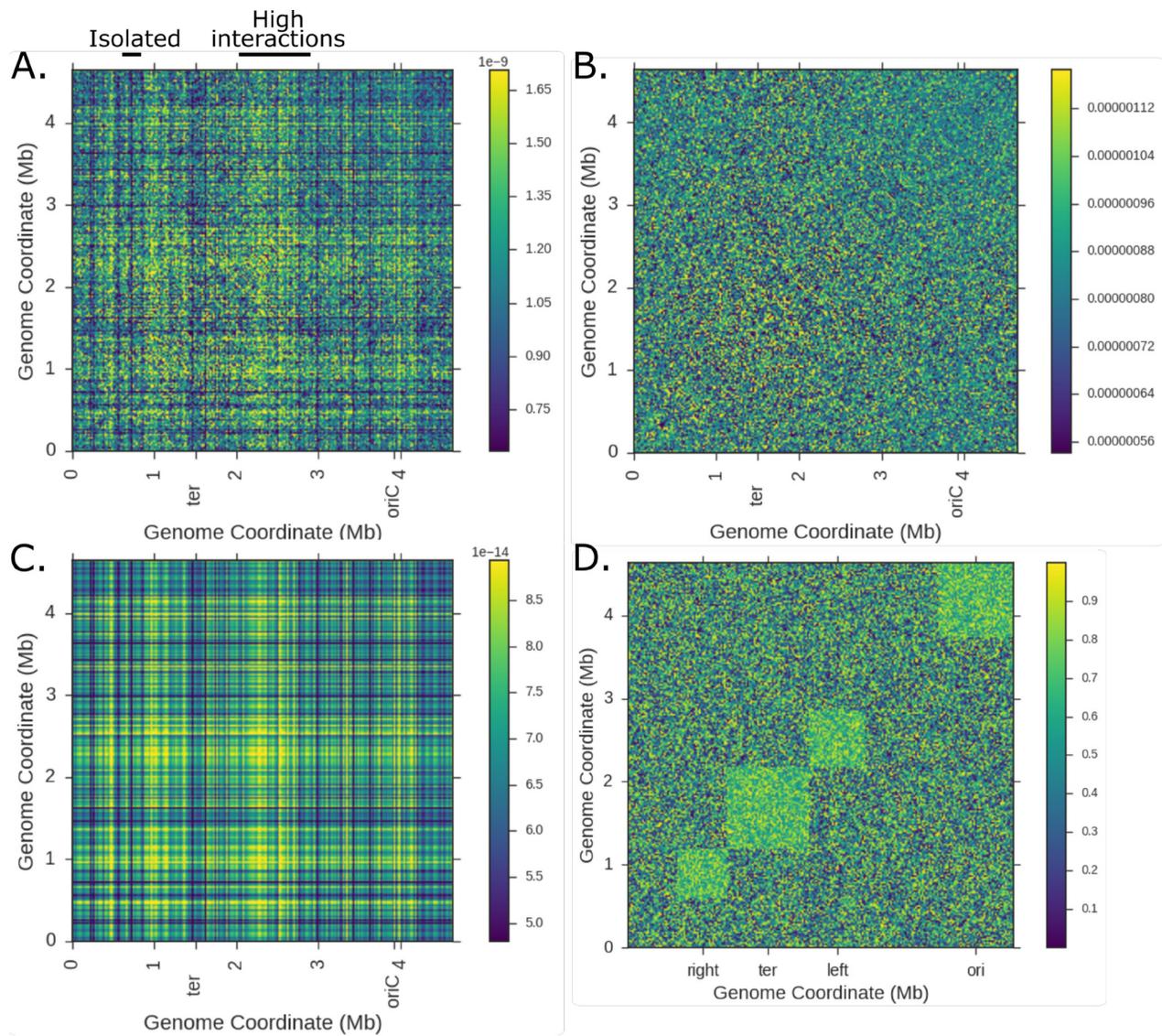
Identifier	Sequence	Notes
P1845	CGCCATAACTAGCTCGGTCAAAGAATTAG GAGCGTGCAGGTGTGTAGGCTGGAGCTGC TTC	Generate Kan cassette to delete <i>maoP</i>
P1846	GAAACTACTGACATAAAAAAAGGGCATT CGCCCTTTTACATATGAATATCCTCCTTA	
P1847	TACTCCGCGCCATAACTAGC	Test <i>maoP::kanR</i> deletion
P1848	GACCGTTTGCTCATCCATCT	
P1582	TCAGACAGGAGTAGGGAAGGAATACAGAG AGACAATAATATGTGTAGGCTGGAGCTGCT TC	Generate Kan cassette to delete <i>lrp</i>
P1583	GAGTGTAATCAAATAACGCCGATTTTGCAC CTGTTCCGTGCATATGAATATCCTCCTTA	
P965	GAACTTCGAAGCAGCTCCAG	Test <i>lrp::kanR</i> deletion
P1568	CAAGGCAACGGTCTTCTCAC	
P1569	CCTGGCTCAAGAAAGGCTCT	
P1838	GCAAGGATTCAGCACCATT	Test <i>dam::kanR</i> deletion
P1839	TCGAAAGAAGAGGCGAAAAA	
P1836	AGTTCCTGCAAGCGACTGAT	Test <i>dcm::kanR</i> deletion
P1837	CGCTGGATCATTTCAGACT	

**Table 2.9: Primers used in this study**

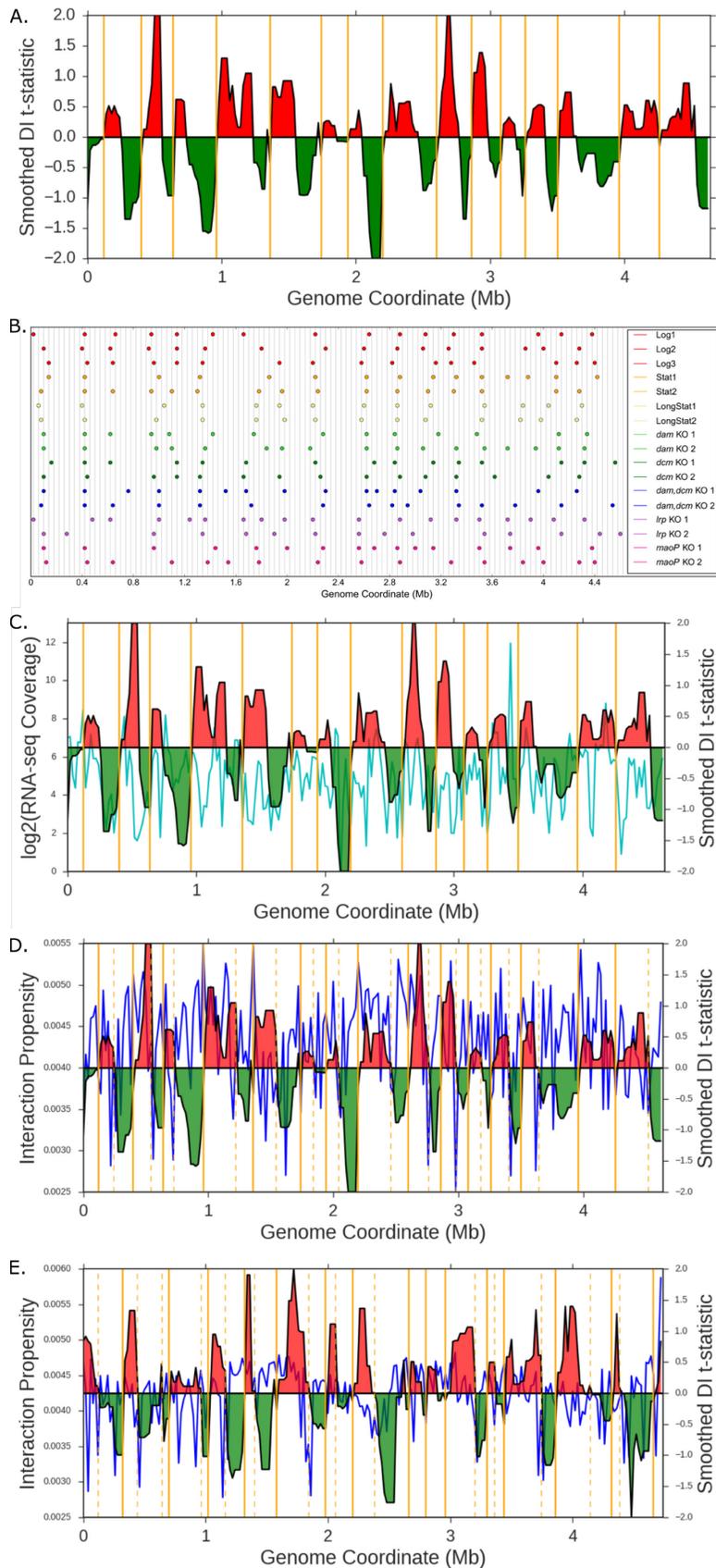
## Figures



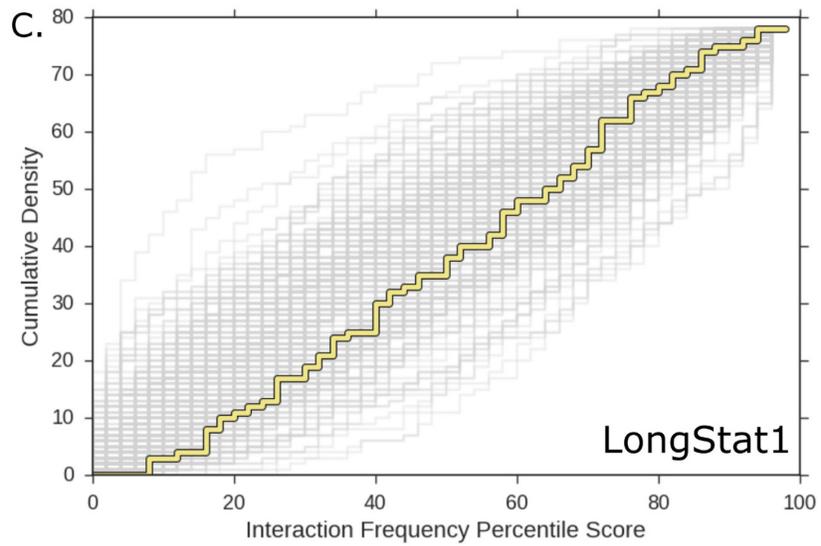
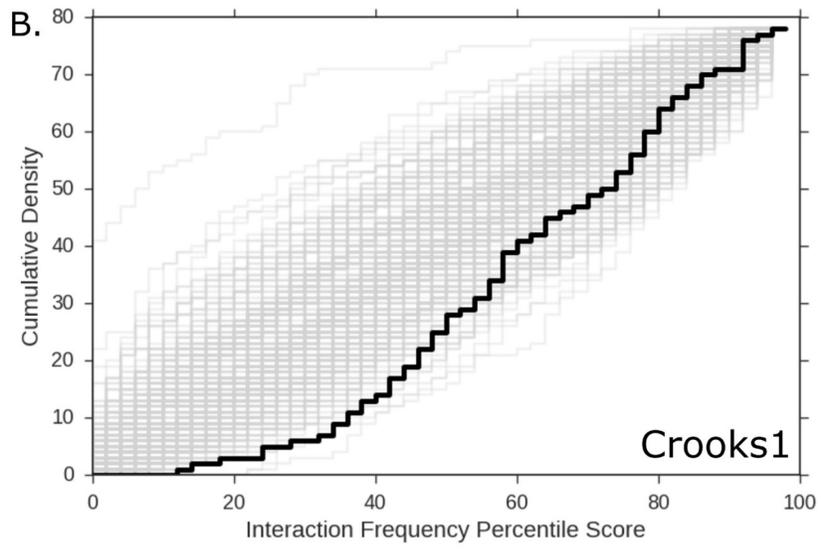
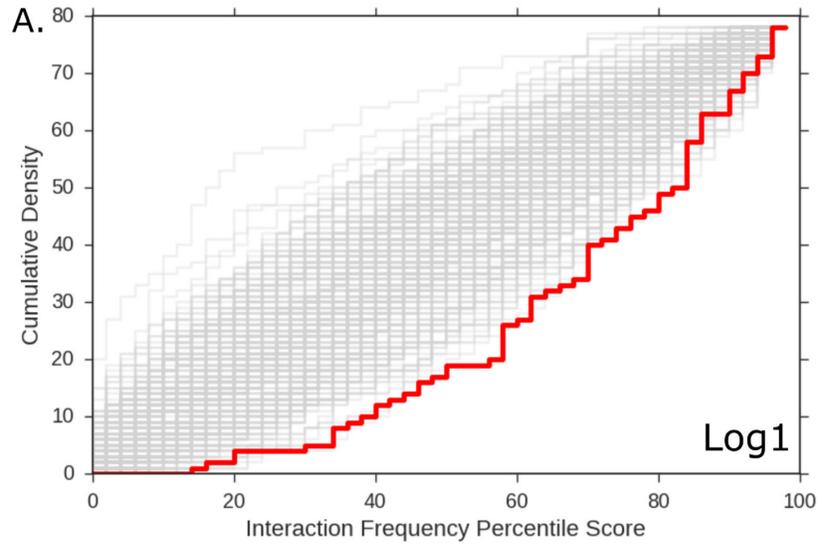
*Figure 2.1: Depiction of ring polymer model. Left and right chromosome arms are shown in red and blue, and the origin of replication is marked by a green circle.*



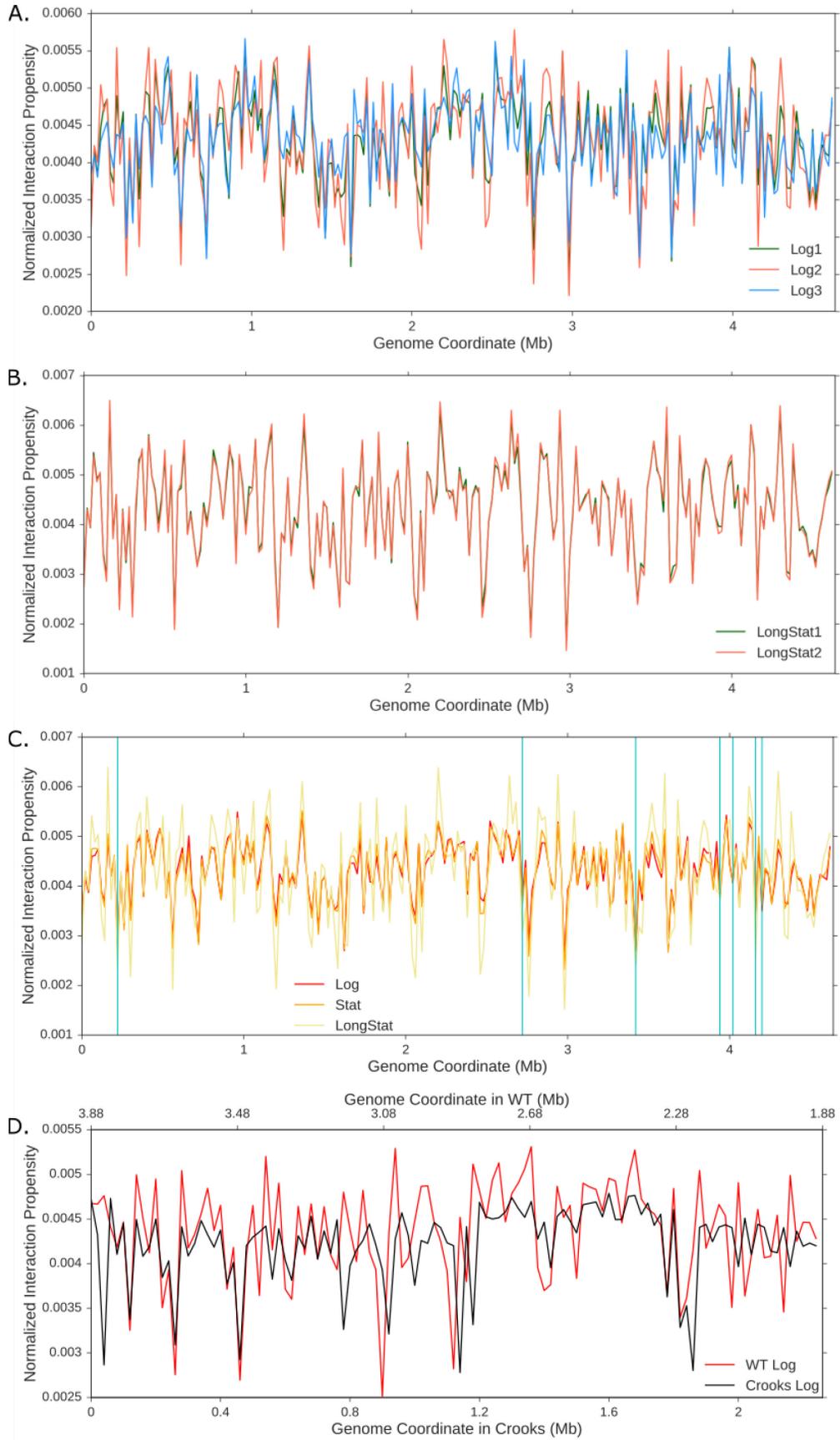
**Figure 2.2: Contact matrices are dominated by the interaction propensity signal.** *A.* Normalized contact matrix of WT cells in logarithmic growth. Regions with high and low interaction propensity are marked. *B.* Contact matrix from *A* normalized by the interaction propensity signal. *C.* Results of simulation of a contact matrix by using only the interaction propensity signal. *D.* Model of expected results for a macromodel-based genome structure. Positions of macromodels are based on Valens et al., 2004.



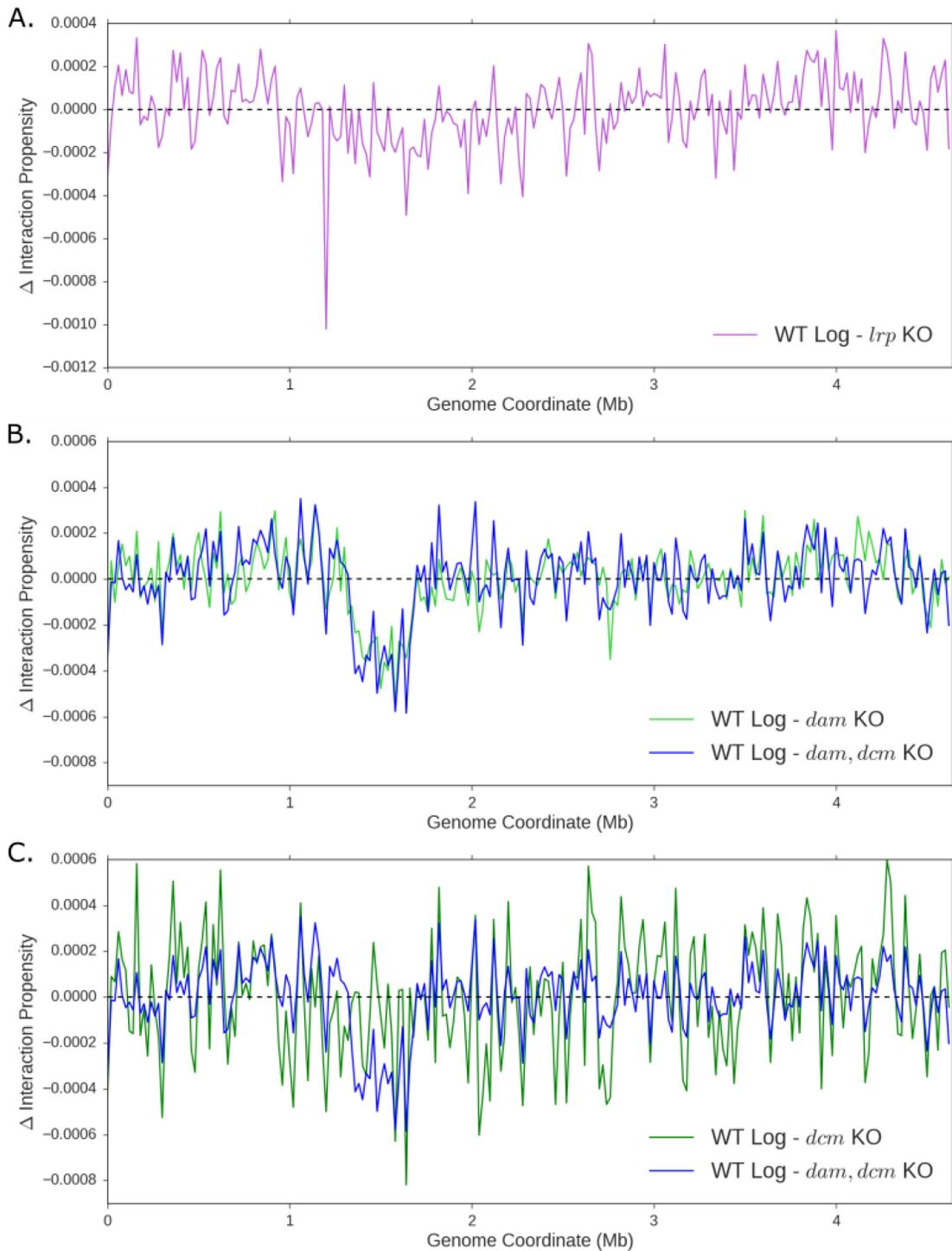
**Figure 2.3: CID organization is consistent across samples and correlated with interaction propensity.** *A.* Directional index (DI) analysis of WT cells in logarithmic growth. Positive t-statistic indicates downstream interaction bias and negative t-statistic indicates upstream interaction bias. CID boundaries identified by a switch from upstream to downstream bias are marked by solid orange lines. *B.* Comparison of CID boundary locations identified in each of the experimental replicates. Grey lines are marked every two units of boundary location assignment. *C.* DI analysis as in *A* overlaid with genome-wide RNA-seq coverage (turquoise). *D.* DI analysis as in *A* overlaid with interaction propensity score (blue). CID centers identified by a switch from downstream to upstream bias are marked by dashed orange lines. *E.* DI analysis of Crooks cells in logarithmic growth overlaid with interaction propensity score as in *D*.



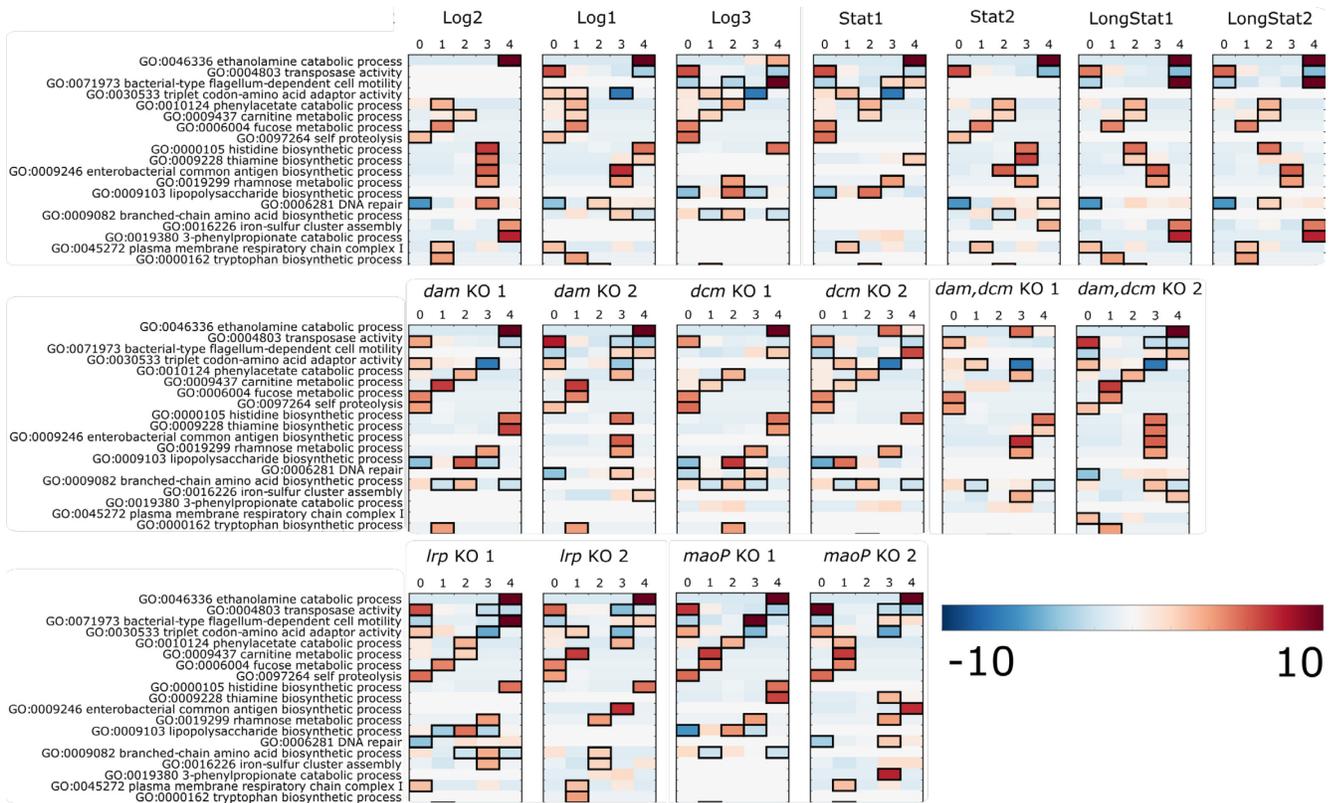
**Figure 2.4: rRNA operons spatially cluster in a time-dependent manner.** Cumulative density histograms for the interaction frequency percentile scores between regions neighboring rRNA operons (red, black or yellow) and every potential permutation of those positions while maintaining inter-operon distance (grey). **A.** WT cells in logarithmic growth. **B.** Crooks cells in logarithmic growth. **C.** WT cells in late stationary phase.



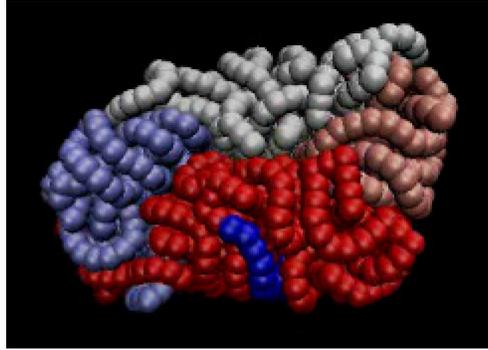
**Figure 2.5: Interaction propensity is strongly conserved.** *A. Normalized interaction propensity for replicates of WT cells in logarithmic growth (n=3). B. Normalized interaction propensity for replicates of WT cells in late stationary phase (n=2). C. Mean normalized interaction propensity for WT cells in logarithmic (n=3), stationary (n=2) and late stationary (n=2) phase. rRNA operon locations are marked in turquoise. D. Mean normalized interaction propensity for WT (n=3) and Crooks (n=2) cells in logarithmic growth. The largest region of alignment between the genomes is shown, with genomic coordinates for Crooks (lower axis) and WT (upper axis) indicated.*



**Figure 2.6: Select mutations result in perturbations of the interaction propensity.** Difference between mean normalized interaction propensity in WT cells ( $n=3$ ) and **A)** *lrp* KO cells ( $n=2$ ), **B)** *dam* KO ( $n=2$ ) and *dam,dcm* KO cells ( $n=2$ ), and **C)** *dcm* KO ( $n=2$ ) and *dam,dcm* KO cells ( $n=2$ ). All samples are from cells in logarithmic growth. Positive differences indicate the interaction propensity is higher in WT cells, and negative differences indicate the interaction propensity is lower in WT cells.



**Figure 2.7: GO-term enrichment relative to interaction propensity.** Enriched GO-terms differ for genes in regions of low or high interaction propensity. A subset of GO-terms enriched or depleted within various conditions are listed to the left. Interaction propensity scores were divided into 5 bins, with 0 indicating low interaction propensity and 4 indicating high interaction propensity. Boxes around a specific GO-term/condition/interaction propensity level indicate a significant enrichment or depletion as indicated by a hypergeometric test ( $p$ -value  $< 0.01$ ). Color inside the box specifies the magnitude of enrichment (red) or depletion (blue) as indicated by the color bar.



***Figure 2.8: Modeling results.** Representative end point from polymer simulation of the chromosome with interaction strengths between beads parameterized based on the interaction propensities. Beads are colored according to macrodomain boundaries. Macrodomains spontaneously self-segregate along the surface of the nucleoid, despite a lack of specific internal interactions, due to low interaction propensities at their boundaries.*

## Chapter 3 Role of the Global Regulator Lrp

### Abstract

The global regulator Lrp plays a crucial role in regulating metabolism, virulence and motility in response to environmental conditions. Lrp has previously been shown to activate or repress approximately 10% of genes in *Escherichia coli*. However, the full spectrum of targets, and how Lrp acts to regulate them, has stymied earlier study. We have combined matched ChIP-seq and RNA sequencing under nine physiological conditions to map the binding and regulatory activity of Lrp as it directs responses to nutrient abundance. In addition to identifying hundreds of novel Lrp targets, we observe two new global trends: first, that Lrp will often bind to promoters in a poised position under conditions when it has no regulatory activity, and second, that nutrient levels induce a global shift in the equilibrium between non-specific and sequence-specific DNA binding. The overall regulatory behavior of Lrp, which as we now show regulates 35% of *E. coli* genes directly or indirectly under at least one condition, thus arises from the interaction between changes in Lrp binding specificity and cooperative action with other regulators.

### Introduction

Over 50% of *E. coli* genes respond to at least one of seven global regulators, of which Lrp is one. Depending on the target, Lrp either activates or represses transcription, and in turn, leucine binding to Lrp either potentiates or inhibits Lrp function (Cho et al., 2008). Overall, Lrp is implicating in regulating about 10% of genes in *E. coli* (Shimada, Saito, Maeda, Tanaka, & Ishihama, 2015). Known targets of Lrp include proteins involved in nitrogen metabolism,

catabolic and anabolic amino acid processes, nutrient transport and pili biogenesis (Calvo & Matthews, 1994; Ishihama, Shimada, & Yamazaki, 2016). Many of these genes have important roles at the entrance to stationary phase; in fact, 70% of the genes that are known to be regulated upon entrance to stationary phase are affected by Lrp (Tani, Khodursky, Blumenthal, Brown, & Matthews, 2002). Lrp expression is highest during minimal conditions, especially when cells begin to enter stationary phase, in agreement with this point being subject to regulation by Lrp (Landgraf et al., 1996).

Cho *et al.* performed chromatin-immunoprecipitation (ChIP) using epitope-tagged Lrp under three conditions, resulting in some expansion of the known Lrp regulon (Cho et al., 2008). However, in comparison to other global regulators, the Lrp regulon as currently known is relatively small, suggesting that all targets have not been identified. In addition, although the concentration of Lrp is not as high as some nucleoid-associated proteins like H-NS and HU, it is expressed to a similar degree as CRP (Newman et al., 1992). Thus one might expect that their regulons would be of similar size; however, there is currently a dramatic discrepancy between these proteins with CRP annotated as regulating 572 genes, while Lrp only regulates 109 (RegulonDB 9.0). Based on estimates about the levels of Lrp and the percentage found free of the nucleoid (Shaolin Chen, Hao, et al., 2001), we estimate that there should be between 400 and 500 Lrp octamers bound and capable of modulating transcription levels under logarithmic growth in both rich and minimal media conditions.

Additionally, we still lack a mechanistic understanding of how Lrp regulation occurs. Not only can Lrp regulation be repressive or activating, its activity can be inhibited, strengthened or not affected by the presence of leucine. Variations in Lrp octamer and hexadecamer affinity for specific target sequences has been suggested as a potential reason for leucine influence

(Shaolin Chen, Rosner, et al., 2001). Interaction between Lrp and a number of coregulators (alternative sigma factors, DNA methyltransferases, and other transcription regulators such as H-NS) has been proposed as a reason for activation or repression on a gene by gene basis, but these patterns are not clear on a global scale.

Making use of a carefully refined ChIP-grade antibody for Lrp, we employed chromatin-immunoprecipitation followed by DNA sequencing (ChIP-seq) of native Lrp in a variety of media conditions and growth phases to assess the full spectrum of Lrp binding sites. Coupled RNA-seq experiments on both wild type (WT) and Lrp knockout (*lrp::kanR*) cells enabled us to distinguish between productive and apparently non-functional binding events, and between direct and indirect Lrp regulatory targets. This rich, high-confidence data set has allowed us to categorize hundreds of novel direct and indirect Lrp targets, expanding Lrp's regulon to 35% of genes in *E. coli* (roughly one-fifth of which are direct targets of Lrp), compared to the 10% previously documented. In addition, we identify a surprising but highly prevalent mode of Lrp binding in which Lrp binds to a site under many physiological conditions, but only alters transcription under certain conditions, similar to poised transcription factor binding in eukaryotes (Graunke, Fornace, & Pieper, 1999; Xiao, White, & Bargonetti, 1998). We show that some of Lrp's poised regulation may be explained by interactions with other regulatory factors such as the nitrogen-response sigma factor,  $\sigma^{54}$ . Despite extensive efforts, we were unable to identify systematically enriched sequence determinants sufficient to either explain transitions from poised to active regulation, or predict Lrp activation from Lrp repression. However, we did observe a shift in Lrp's DNA binding specificity in response to varying nutrient conditions. The conservation of Lrp across many species of bacteria and archaea (Brinkman et al., 2003) argues for its critical role in organismal survival, and this work provides the most comprehensive

picture of the Lrp regulon in *E. coli* to date, establishing rules for Lrp behavior that will likely illuminate study of the protein in many species. The general principles of Lrp's behavior across conditions may also serve as a template for other bacterial global regulators.

## Results

### *ChIP-seq and RNA-seq identify hundreds of novel Lrp targets*

We performed both ChIP-seq and RNA-seq on WT and Lrp knockout (*lrp::kanR*) cells to establish a global picture of Lrp binding and regulatory effects in nine physiological conditions. Conditions and time points will be referenced as follows: the time points are denoted X\_Log (logarithmic phase), X\_Trans (transition point), and X\_Stat (stationary phase), where the X may be MIN (minimal media), LIV (minimal media supplemented with branched-chain amino acids), or RDM (rich defined media, Figure 3.1). Overall, the combination of Lrp binding data from the ChIP-seq experiments and the expression data from the RNA-seq experiments resulted in identification of hundreds of novel Lrp targets. We document a ten-fold range (between 62 and 666) in the number of Lrp peaks identified across the nine physiological conditions examined here. Fewer Lrp binding sites are identified in media with higher nutrient conditions (either LIV or RDM) relative to the MIN (Figure 3.2A), in agreement with previously published Lrp ChIP data (Cho et al., 2008) and with Lrp's known role as a regulator which responds to decreasing nutrient levels. However, our data identifies between two- and five-fold more binding sites overall than previous studies. In general, we document more Lrp binding sites at later time points (Trans and Stat) relative to Log (Figure 3.2A); again in agreement with previously published Lrp ChIP data (Cho et al., 2008) and with the known role of Lrp as being a critical regulator at the transition to stationary phase. Comparing our data to previously published ChIP-

ChIP studies (Cho et al., 2008), we identify extensive overlap in binding locations: 96% of sites in prior ChIP-ChIP data are reproduced in our data at MIN\_Log (27.7 fold enrichment;  $p < 0.001$ , permutation test,  $r=1000$ ; here and throughout the manuscript we use  $r$  to refer to the number of replicates used for resampling tests), 44% at LIV\_Log (123.1 fold enrichment,  $p < 0.001$ , permutation test,  $r=1000$ ) and 84% at MIN\_Stat (15.5 fold enrichment,  $p < 0.001$ , permutation test,  $r=1000$ ). Comparing at the level of genes which are identified as having a Lrp-dependent change in expression as measured by RNA-seq, our data set overlaps with 78% of the known targets in RegulonDB (1.50 fold enrichment,  $p < 0.001$ , permutation test,  $r=1000$ ), 77% of the previously identified ChIP-ChIP targets (1.49 fold enrichment,  $p < 0.001$ , permutation test,  $r=1000$ ) (Cho et al., 2008), and 89% of the previously identified microarray targets (1.72 fold enrichment,  $p < 0.001$ , permutation test,  $r=1000$ ) (Tani et al., 2002), showing good agreement across the variety of strains and media conditions present in the compared studies, despite some variations in precise experimental conditions. We also identify over 900 novel Lrp binding sites, and 2104 genes with previously undocumented Lrp-dependent expression.

Many well-studied Lrp targets are reproduced in our data. IlvI (b0077) is an enzyme critical for valine and isoleucine biosynthesis that is known to be activated by Lrp (D. A. Willins & Calvo, 1992). Consistent with prior work, we see a strong Lrp binding signal at the *ilvI* transcription start site (Figure 3.2B, left panel), and a Lrp-dependent activation of *ilvI* transcription in several media conditions (Figure 3.2B, right panel). The extent of activation is weakened or eliminated completely in LIV or RDM conditions, in agreement with previous studies showing that leucine inhibits the Lrp-mediated activation of *ilvI* (Platko, Willins, & Calvo, 1990).

A strong Lrp binding signal under MIN conditions is also evident at the promoter region for *OppA*, a protein critical for oligopeptide transport (Klepsch et al., 2011). Lrp is known to repress expression of the *oppABCDF* operon in the absence of leucine (Calvo & Matthews, 1994). Accordingly, we see Lrp-dependent repression of *oppA* under MIN conditions (Figure 3.2C, right panel). The Lrp binding signal is strongly attenuated, and there are no Lrp-dependent expression effects, during LIV and RDM conditions (Figure 3.2C).

#### *Global analysis reveals that Lrp has condition-specific modes of binding and regulation*

Global regulators are known to act both directly, by binding target sites and modulating transcription levels, and indirectly, by modulating the expression of transcription factors which have their own targets (Martínez-Antonio & Collado-Vides, 2003). Previously, most focus on Lrp regulation has been at the direct target level. By comparing the binding data from our ChIP-seq experiments and the corresponding expression data provided by our RNA-seq experiments, we are able to identify and categorize both direct and indirect targets under a variety of physiological conditions. Direct and indirect targets are both characterized by Lrp-dependent changes in expression, but only direct targets have a Lrp binding signal in their regulatory regions, defined as 500 bp upstream and downstream of the annotated transcription start site (Figure 3.3A). In addition, our data shows many examples of a converse mode of Lrp activity, in which binding of Lrp is apparent at a particular promoter, but there are no Lrp-dependent changes in expression (these sites comprise 64-94% of all instances of Lrp binding across the conditions that we studied). We refer to such cases as instances of nucleoid-associated protein (NAP) activity of Lrp, thus described due to the similarity of Lrp's behavior at these locations with highly abundant, low-specificity NAPs such as H-NS and HU (Luijsterburg et al., 2006). In

our data set, neither NAP-type Lrp targets nor genes unconnected to Lrp show Lrp-dependent RNA expression changes by definition, but the NAP targets have a Lrp binding signal (Figure 3.3A). This is apparent, for example, at the *ybjN* gene; its promoter region is always bound by Lrp, but never exhibits a significant change in expression, thus making it a NAP-type target under all conditions (Figure 3.3B). Interestingly, YbjN is proposed to play a role in stress response and motility (D. Wang et al., 2011), areas to which many of Lrp's targets are known to belong. The consistent binding of *ybjN*'s promoter by Lrp coupled with the similarity of its role to other Lrp targets suggests that Lrp may always be poised to regulate YbjN, and that it may be a direct target of Lrp under conditions not tested here.

Based on our RNA-seq data, we find that 1.7% to 29% of all *E. coli* genes are regulated by Lrp in each condition (Table 3.1); in all, 2320 genes (50% of all genes in *E. coli*) show Lrp-dependent changes in transcript levels under at least one condition. However, due to the presence of operons in *E. coli*, in the analysis below we only categorized genes (as direct, indirect or NAP-type targets) if a transcription start site exists in the PromoterSet dataset in RegulonDB version 9.4 within 500 base pairs upstream of the start of the coding region, resulting in categorization of 2875 genes out of the 4658 present in *E. coli* MG1655. From our analysis of that categorizable subset, we note that 35% of all *E. coli* genes are regulated by Lrp, either directly or indirectly, in at least one condition. Out of those, about 13% are regulated directly, 81% are regulated indirectly, and 6% are labelled as indirect and direct targets in different conditions. Due to the restriction on categorizing genes noted above, the counts given here are an underestimate. If we assume each transcription unit is fully transcribed and therefore assign the Lrp categorization of the first gene to each subsequent gene in the transcription unit, that increases the Lrp regulon to 49% of all *E. coli* genes (2289 genes/4658 total genes), with

16% of that total being direct targets, 78% being indirect targets and 6% being categorized as both in different conditions. This estimate at the level of transcriptional units is nearly identical to the fraction of genes that show Lrp-dependent changes in expression in the gene-level data discussed above. In addition, since some later genes within a transcriptional unit have independent transcription start sites, there exists the possibility of those genes being categorized as indirect targets because the operon itself is Lrp-controlled even if there is no Lrp peak near the later gene in the operon. Given that we do not know how often the overall operon start site is used compared to the internal start sites, we are not able to place those genes unambiguously as indirect targets. We calculated how many of these ambiguous indirect targets were present per condition and noted that it is a fairly small percentage of total indirect targets, ranging from 1.2% to 11.8% (with a median of 5.7%). These ambiguous targets comprise 44 total genes in *E. coli* (Table 3.2). Given the small size of this gene subset, we proceeded with the original categories and included the ambiguous indirect targets with the indirect targets in the analysis below.

We next used hierarchical clustering to order the categorized genes (those 2875 genes with an annotated transcription start site) by assessing how similar their categorization assignments were over the nine sampled physiological conditions (Figure 3.3C). We immediately noted that genes can transition between labels, e.g. from direct target to NAP, depending on the media condition and time point. As seen with the case of YbjN above, these findings suggest that Lrp is often poised at a particular gene under many conditions, but must act combinatorially with some other factor or environmental stimulus in order to actually alter expression. In addition, we see evidence for leucine-independent and leucine-dependent binding; some genes are always NAP-type or direct targets (i.e. Lrp bound) regardless of condition (the leucine-independent group) and some are only bound during MIN conditions (the

leucine-dependent group). There is no obvious cluster that is bound only under conditions of high leucine levels. We also observe a dramatic increase in the number of indirect targets at MIN\_Trans and RDM\_Stat, going from 152 to 509 indirect targets between MIN\_Log and MIN\_Trans, and from 34 to 919 indirect targets from RDM\_Trans to RDM\_Stat. These conditions likely represent points in growth at which Lrp's regulatory activity is particularly important for fitness.

### *Lrp binding is enriched among regulatory regions of the genome*

As detailed in the Methods section, our process for categorizing genes as Lrp targets involved testing whether there was a called Lrp peak within 500 bp upstream or downstream of each annotated transcription start site (TSS) in the *E. coli* genome. If there were multiple annotated transcription start sites, we took 500 bp upstream of the most distal TSS (relative to the start of the gene itself) and 500 bp downstream of the most proximal TSS. We classified those approximately 1000 bp windows as regulatory regions, and tested whether Lrp binding was significantly enriched in those regions. Overall, 48% of the *E. coli* genome falls into these regulatory regions. However, we observe between 63% and 89% of Lrp peaks appearing in regulatory regions. A permutation test in which the same size and number of peaks were randomly shuffled across the genome indicated that there is significant enrichment for Lrp binding in regulatory regions (Table 3.3). This strongly supports Lrp's role as a specific regulatory protein.

The Lrp peaks not in regulatory regions were distributed in gene coding regions, between genes in a transcription unit, or in truly intergenic regions at relative ratios similar to the proportion of those regions on a genome-wide scale (Table 3.4). We investigated whether any of

those peaks might affect full transcription of an operon, hypothesizing that Lrp binding in the middle of an operon might block RNA polymerase. From the RNA-seq data, we identified any genes that showed a Lrp dependent change in expression, were not classified (and so did not have their own transcription start site), had a Lrp response that was different from the first gene in the corresponding transcriptional unit, and had a Lrp peak within 1000 bp upstream or downstream of the gene coding region. Due to incomplete annotation of the *E. coli* genome, some of these genes appear to be ones that should have a unique transcription start site based on visual inspection of the genomic context. However, for the remaining examples, we compared the RNA-seq coverage to the location of the peak as identified by the Lrp ChIP signal. As seen for the binding at *ilvI*, we again note that Lrp binding does not guarantee a regulatory effect. Genes that have a strong internal Lrp binding site under all conditions do not evince a Lrp dependent change in expression at all times, and Lrp binding sites within an operon do not, in general, appear to hamper transcription (Figure 3.4). These findings again suggest that Lrp regulation is often dependent on cooperative interaction with other regulatory factors, and that Lrp binding alone within operons does not have a constant, systematic effect.

#### *Direct Lrp targets explain the Lrp-dependent regulatory effect at some indirect targets*

Given the high proportion of indirect Lrp targets, and especially the dramatic increase in the number of indirect targets at MIN\_Trans and RDM\_Stat, we investigated whether some of the expression changes of those indirect targets can be explained by the activity of direct Lrp targets at those time points. As Lrp is a global regulator, we expected to find that some percentage of the indirect targets could be explained by considering the known targets of the transcriptional regulators categorized as direct targets under that condition. We would expect

that in such cases, we should observe an enrichment among Lrp indirect targets of genes known to be regulated by Lrp direct targets. We observe significant, albeit small, enrichment of explainable indirect targets across all conditions except MIN\_Stat, LIV\_Trans, LIV\_Stat and RDM\_Trans; a maximum of 8% of indirect targets can be explained by the currently known targets of direct Lrp targets (Table 3.5). Direct Lrp targets that are not currently identified as transcriptional regulators or regulators with incompletely documented regulons could account for why we are not able to explain more instances of indirect regulation, as could transcriptional units regulated by aspects of cellular state that are themselves Lrp-dependent. Several key transcription factors that are direct Lrp targets are responsible for explaining the identified indirect Lrp targets across conditions: Nac, GadW, PurR, LeuO, ArgR, QseB, CysB, NagC, SlyA, SoxS, and LrhA. Several of these transcription factors have also previously been identified as Lrp targets (Shimada et al., 2015).

Investigating at a local as opposed to global scale provides several informative examples. At LIV\_Log, LIV\_Trans and RDM\_Log, the dual regulator LrhA is a direct Lrp-activated target gene (Figure 3.5A). LrhA activates *fimE* and represses *flhC* and *flhD* (Figure 3.5B). At LIV\_Log, *fimE* is indirectly activated; at LIV\_Trans, *flhC* is indirectly repressed, and at RDM\_Log, both *flhC* and *flhD* are indirectly repressed (Figure 3.5A). While this pattern does not show activity at every LrhA target in each condition, overall it suggests that indirect regulation of *fimE* and *flhCD* by Lrp may be explained in some cases by direct LrhA activation by Lrp. All three target genes are also known to be regulated by other transcription factors, potentially explaining the incomplete activity from LrhA. Similarly, at MIN\_Trans, the transcriptional regulator CysB is a direct Lrp-repressed target gene (Figure 3.5C). CysB is known to activate *tcyP* and *cysI*, among other genes (Figure 3.5D). Both *tcyP* and *cysI* were

categorized as indirect Lrp-repressed targets, supporting the fact that Lrp repression of *cysB* is what leads to repression of *tcyP* and *cysI*. The transcription factor GadW is an interesting example in that it is a direct Lrp-repressed target at LIV\_Log and a direct Lrp-activated target at RDM\_Stat. At both conditions, more than 75% of GadW's annotated targets are indirect Lrp targets, all repressed at LIV\_Log and all activated at RDM\_Stat, as would be expected if GadW activates them. Thus, this illustrates another case where indirect Lrp-mediated regulation is explained by identifying a transcription factor which is a direct Lrp target.

*Direct and indirect Lrp targets have both shared and unique GO-term classifications*

After grouping direct and indirect targets, we used iPAGE (Goodarzi et al., 2009) to search for enrichment of gene ontology (GO) terms that share mutual information with our categorization scheme. We observe the general trend that pathways involved in direct synthesis or acquisition of nutrients (e.g. amino acid transport and L-serine biosynthetic processes) tend to be direct targets or NAP-type targets, whereas those involved in regulation of cellular behavior and foraging strategies (e.g. flagellum and motility) tend to be indirect targets, particularly under the richer media conditions LIV\_Log and RDM\_Log (Figure 3.6A, Figure 3.7A). Interestingly, in testing for enrichment among the large block of indirect targets at RDM\_Stat, we observe that it is depleted for flagellum-related genes. Under minimal conditions, indirect targets overlap with some of the transport pathways otherwise mainly observed to be enriched among direct targets.

We also see overlapping enriched GO-terms at direct and NAP-type targets, suggesting that Lrp may preemptively bind some target genes before conditions occur at which regulatory action is required (discussed in more detail below). A particularly clear example of such poised

regulation comes in identifying significant GO-terms among the genes that have Lrp binding activity in at least eight of the nine conditions tested, but only become direct targets during certain conditions (e.g. many of the genes in the leucine independent cluster (Figure 3.3C) qualify). Strikingly, enriched GO-terms include leucine transport, serine biosynthetic processes and general amino acid transport (Figure 3.6B). This indicates that regardless of the level of leucine, Lrp's traditionally recognized small molecular partner, Lrp remains bound to and poised to regulate critical genes if conditions change. Furthermore, the key signal causing a transition between NAP-type activity and direct transcriptional regulation is unlikely to be leucine levels themselves, as NAP to direct changes occur at certain genes across the time course of growth under Minimal conditions (see Figure 3.3C). Overall, the dynamic nature of what constitutes a Lrp-regulated target under different media conditions and points of growth demonstrates the complexity of the Lrp regulon.

In order to illuminate what distinguishes the various effects of Lrp binding on gene regulation, we tested whether splitting direct and indirect targets into sub-classes that are activated or repressed by Lrp revealed a different pattern of GO-terms. From this analysis, it is evident that the flagellar genes are enriched among indirect Lrp-repressed targets specifically (Figure 3.7B). In addition, many of the genes involved in transport processes appear to be direct Lrp-repressed target genes, whereas the genes involved in biosynthetic processes are often direct Lrp-activated target genes. Interestingly, at MIN\_Trans, a condition in which we see a spike of indirect targets, there is a specific class of GO-terms which are enriched for either indirect Lrp-activated (e.g. ferrous iron binding and N-terminal protein acetylation) or indirect Lrp-repressed targets (e.g. NAD binding and histidine biosynthetic processes). Those GO-terms do not have

enrichment among indirect targets at RDM\_Stat, reinforcing the notion that Lrp acts on unique sub-clusters of its targets under different conditions (as seen in Figure 3.3C above).

*Lrp is poised at many targets to enable combinatorial regulation*

Upon filtering and categorizing genes, we noticed immediately that many genes shift between being a NAP-type target and a direct Lrp-activated or repressed target under different conditions (see Figure 3.3C above). In fact, 91% of direct Lrp targets are NAP-type targets in at least one condition, and thus have Lrp bound to their promoter even though it has no impact on transcription. For example, the MIN\_Trans cluster consists of genes that are bound by Lrp in all Minimal media time points, but only show Lrp-dependent changes in transcript level during the Trans time point. This suggests that Lrp binds some promoters in a poised position under a broad range of conditions, but only regulates when certain additional criteria are met, perhaps by coordinating with a second regulatory factor to enable combinatorial logic. Among genes that undergo a transition between being a NAP-type target and a direct target, 38.9% become activated, 47.4% become repressed and 13.7% become both activated and repressed in different conditions.

For example, *potF*, a component of the putrescine ABC transporter (Vassylyev, Tomitori, Kashiwagi, Morikawa, & Igarashi, 1998), shows Lrp binding in its promoter region under all nine conditions measured in our data, but is only activated by Lrp during MIN, LIV\_Stat, RDM\_Log and RDM\_Stat conditions (Figure 3.8A). In contrast with the variable Lrp-dependent RNA expression levels, Lrp binding at *potF* is very similar across conditions, spanning a similar length of DNA, and showing maximal signal at the same point. *potF* was previously identified as a Lrp regulated target which is repressed by Lrp alone, and activated when leucine is present

(Cho et al., 2008). However, those experiments employed glucose rather than glycerol as a carbon source, and monitored response to the addition of 10 mM leucine alone versus 0.2% (w/v) isoleucine, valine and leucine (equivalent to 15.25 mM leucine) which could explain the differences in observations of Lrp's regulatory action at *potF*.

*lrhA*, a transcriptional regulator involved in fimbriae synthesis (Blumer et al., 2005), also has Lrp binding signal under all conditions. Interestingly, it is activated only at the high-leucine conditions LIV\_Log, LIV\_Trans and RDM\_Log (Figure 3.8B), a different pattern from many other activated genes, which generally are activated in later time points in the growth curve or under MIN conditions. Again, the Lrp binding signal at *lrhA* is very similar across conditions, with only slight variation in the signal magnitude, in contrast to the sharp differences in the Lrp-dependent RNA expression changes. Thus, the changes in regulatory activity cannot be due to changes in the location of Lrp binding.

*dadA*, which encodes a critical enzyme in amino acid degradation (Franklin & Venables, 1976), is one of the interesting class of examples that we see transition from a NAP-type target to being a direct Lrp-activated or repressed target in different conditions. *dadA* expression is strongly Lrp-repressed at MIN\_Log, whereas it is activated during LIV\_Log, LIV\_Trans or RDM\_Trans (Figure 3.8C). Lrp is known to repress *dadA* in the absence of leucine (Mathew, Zhi, & Freundlich, 1996), a fact strongly supported by our data in which we see Lrp-mediated repression in minimal media and alleviation of repression during growth with higher levels of leucine. This variability in regulatory effect is in sharp contrast to the almost identical Lrp binding signal present in all nine conditions. Another gene which transitions between being a NAP-type target and being a direct Lrp-repressed target while having a similar Lrp binding profile is *pepD*, which is repressed at MIN\_Trans and RDM\_Stat (Figure 3.8D). *pepD* encodes

Peptidase D, which cleaves a variety of dipeptides (Schroeder, Henrich, Fink, & Plapp, 1994). It is important to note that the location of the Lrp binding peak relative to the transcription start site does not systematically affect the direction of Lrp regulation (for example, compare *ilvI* and *potF*).

### *Lrp connects with other regulatory factors*

The phenomenon outlined above -- of Lrp frequently binding to a promoter under many conditions but only showing regulatory activity under a few -- suggests that other regulatory factors, such as  $\sigma$  factors or transcription factors, may be important in triggering an activating or repressive effect secondary to Lrp binding. If a  $\sigma$  factor and Lrp coregulate some set of targets, we expect to see enrichment for direct targets relative to NAP-type targets within the  $\sigma$  factor's regulon, especially at conditions when the  $\sigma$  factor is most active. To establish relative  $\sigma$  factor activity, we determined the average expression of all known  $\sigma$  factor target genes at each of our nine experimental conditions (Figure 3.9A)(Gama-Castro et al., 2016). One caveat of our analysis is that some data is missing since we do not classify all genes in relation to Lrp, as outlined above, and, likewise, it is not known by which  $\sigma$  factor all genes that are classified are regulated. Subject to these constraints, our analysis in this section included 1534 genes. In addition, in some cases, overlap between other factors and Lrp may not indicate a direct interaction but may indicate that the other factor and Lrp have independent roles or functions at shared targets, here termed convergent regulation. However, if Lrp does interact directly with certain  $\sigma$  factors to activate target genes at specific conditions, there are a few possible explanations for why the NAP-type to direct target transition occurs at those points: 1) the transition only occurs when the genes' controlling  $\sigma$  factor is active; 2) the nature or extent of

Lrp binding itself changes at that condition; or 3) an accessory factor needed for Lrp- $\sigma$  factor interaction is only present at that condition.

We applied a permutation test to identify any  $\sigma$  factors with a significant enrichment of overlap between their targets and all direct Lrp targets or specifically direct Lrp-activated targets. All q-values and enrichment levels for the permutation test with all direct targets are listed in Table 3.6; results from the permutation test with only direct-activated targets are in Table 3.7 ( $r=10000$  for both). Only  $\sigma^{54}$  at MIN\_Trans had significant overlaps ( $q<0.05$ ); in addition,  $\sigma^{38}$ , which has previously been implicated in Lrp-mediated regulation at two genes, *osmY* and *osmC* (Bouvier et al., 1998; Colland, Barth, Hengge-Aronis, & Kolb, 2000), only narrowly missed our threshold at RDM\_Stat (2.3 fold enrichment of direct activated targets, q-value: 0.059). A role for  $\sigma^{38}$  at stationary phase is logical since it coordinates general stress responses in *E. coli* (Battesti, Majdalani, & Gottesman, 2011). However, direct  $\sigma^{38}$ /Lrp interaction is not likely since many of the Lrp- $\sigma^{38}$  shared regulated genes at RDM\_Stat are NAP-type Lrp targets in other conditions (MIN\_Stat, LIV\_Stat and RDM\_Trans) when  $\sigma^{38}$  is more active (Figure 3.9A,B). Therefore, this overlap is likely a result of convergent regulation between Lrp, a “feast-famine” regulatory protein, and  $\sigma^{38}$ , a stress-response  $\sigma$  factor.

In contrast with  $\sigma^{38}$ , for  $\sigma^{54}$  we observe marked enrichment, especially of direct Lrp-activated targets, under the condition when  $\sigma^{54}$  is most active. Specifically, we document enrichment for direct Lrp targets with  $\sigma^{54}$ ( $\sigma^N$ ) at MIN\_Trans (1.8-fold enrichment, q-value: 0.076). At MIN\_Trans, 39% of Lrp binding sites overall are direct targets, whereas 70% of  $\sigma^{54}$  targets with Lrp binding sites are direct targets. Furthermore, as we would expect for the case where Lrp acts as a co-activator for a given  $\sigma$  factor, there is enrichment specifically for direct Lrp-activated target genes among  $\sigma^{54}$  targets at MIN\_Trans (3.0-fold enrichment, q-value:

0.016). Overall, 20% of Lrp binding sites are direct activated targets at MIN\_Trans, whereas Lrp-bound targets in the  $\sigma^{54}$  regulon are direct Lrp-activated targets 61% of the time, a 3.0-fold increase.  $\sigma^{54}$  regulates many genes involved in nitrogen assimilation (Larry Reitzer & Schneider, 2001), and these results indicate that Lrp is likely involved in co-activating some  $\sigma^{54}$  dependent genes, in agreement with Lrp's role in sensing and responding to nutrient levels. At MIN\_Trans, Lrp actually also weakly represses  $\sigma^{54}$  itself directly;  $\sigma^{54}$  is not a direct or indirect target under any other conditions (Figure 3.10B).

Average expression of  $\sigma^{54}$  targets peaks at MIN\_Trans (Figure 3.9A), in agreement with when we see overlap between its targets and direct Lrp-activated targets (13.3% of the direct Lrp-activated targets at MIN\_Trans are known  $\sigma^{54}$  targets, and conversely 22.6% of the classified  $\sigma^{54}$  targets are direct Lrp-activated targets at MIN\_Trans). Nine out of the fourteen overlapping target genes only become a direct Lrp-activated target at MIN\_Trans. The remaining four genes (*astC*, *hisJ*, *potF*, *yhdW*) are affected at other conditions when there is a slight peak in  $\sigma^{54}$  activity, as measured by the overall expression of known target genes (Figure 3.9A), and could be subject to other regulatory control. For example, *astC* and *hisJ* are also regulated by ArgR in some conditions (Caldara, Charlier, & Cunin, 2006; Kiupakis & Reitzer, 2002). The fact that the shared regulated genes are only direct Lrp-activated targets when  $\sigma^{54}$  itself is most active supports the notion that  $\sigma^{54}$  may require Lrp binding to activate transcription of certain genes. At a molecular level, this suggests that while expression of  $\sigma^{54}$  itself during MIN\_Trans does not require Lrp (and in fact, is slightly repressed by Lrp), its transcriptional activity is enhanced by the presence of Lrp (also see Figure 3.10A).

To investigate the possibility that Lrp binding itself changes to facilitate interaction with  $\sigma^{54}$ , we visualized the Lrp-ChIP binding signal at shared direct Lrp/ $\sigma^{54}$  targets. Changes in Lrp

binding, either complete reversals of binding between conditions or changes in peak length, are evident in the cases of some genes (*glnH*, *yeaG* and *yhdW*), while others, such as *ibpB* and *potF* have almost identical binding regardless of condition (see Figure 3.8A for *potF* Lrp-binding signal); thus, it is unlikely that changes in Lrp binding itself are in general responsible for the regulatory interaction with  $\sigma^{54}$ . Given that  $\sigma^{54}$  is known to require activating factors, it is likely that an accessory factor may facilitate Lrp/ $\sigma^{54}$  coregulation.

To identify other candidates for coregulators acting with Lrp, just as we tested for Lrp coregulation with  $\sigma$  factors, we investigated whether Lrp has particular correlations with any of the other annotated transcription factors in *E. coli*, including the other six global regulators. We compared the average expression of all annotated targets of individual transcription factors in WT and Lrp KO conditions to identify those transcription factor regulons that show Lrp-dependent changes. Several transcription factors were identified as significant ( $q < 0.05$ ) based on a permutation test ( $r = 10000$ ): EvgA, FlhDC, LeuO, ModE, NtrC, PhoP and TorR. We then applied the additional threshold of requiring an average four-fold or greater change in expression of target genes dependent on Lrp status (WT vs. KO) at the appropriate condition to identify the most biologically relevant interactions (Figure 3.9C); the transcription factors PhoP and FlhDC did not pass this filter and were eliminated from further analysis. EvgA, LeuO, ModE and TorR all likely represent convergent regulation due to the existence of no or limited overlap between transcription factor targets and direct Lrp targets. None of the global regulators appeared as significant co-regulators with Lrp based on our analysis, but there is the possibility of some cross-talk since Lrp directly regulates CRP and indirectly regulates ArcA and Fis. In addition, FNR is a NAP-target in eight conditions, suggesting that Lrp may regulate it under some stress conditions.

The transcription factor NtrC is a notable exception to the above trend suggesting convergent regulation as the reason for target overlap between Lrp and other transcription factors, as 33% of all its targets are also direct Lrp-activated targets (Figure 3.9D). This number is an underestimate since it only accounts for the genes classified in our scheme (namely those with annotated promoters); if we expand our classification to include the genes that comprise the transcription units of those classified genes, 74% of NtrC targets are also direct Lrp-activated targets. Two indirect Lrp-activated transcription units comprise the remainder of the NtrC regulon. NtrC is one of the transcription factors which can serve as an activator of  $\sigma^{54}$ , so the intersection between Lrp, NtrC and  $\sigma^{54}$  is interesting to consider. Activators of  $\sigma^{54}$  often bind at a distance from the promoter and so require significant DNA bending to come in physical contact with  $\sigma^{54}$  (L. Reitzer, 2003). IHF is a DNA bending protein known to facilitate DNA bending at some target genes, but our data indicates that Lrp may also have a role in DNA bending, and thus activation of NtrC/ $\sigma^{54}$  transcribed genes (see Discussion). Thus, while many instances of Lrp regulation appear to require coregulation with as yet unidentified regulatory factors, we are able to identify some likely possible mechanisms.

#### *Lrp binding sites have a condition- and time-specific motif preference*

While not as invariant as motifs for other *E. coli* transcription factors, a 15 base-pair motif comprising terminal inverted repeats and an AT-rich center was previously identified for Lrp (Cho et al., 2008; Yuhai Cui, 1995). We wanted to determine if a similar motif is apparent in our data, and how well Lrp binding is predicted by the presence of Lrp motifs. We used a logistic regression model to classify 500 bp windows of the genome as either containing a Lrp peak or not, using as predictors the presence of previously documented Lrp motifs and the AT

content (given the AT richness of the Lrp motif itself). Starting with a minimal model containing only an intercept term, we created more complex models by adding a single predictor at a time and scoring each new model with the Bayesian Information Criterion (BIC) as displayed in Figure 3.11A; *n.b.* a lower BIC indicates a more parsimonious model. A minimal model was chosen by adding to the new model the predictor with the largest decrease in BIC from the intercept-only model and iterating this process until the change in BIC switched sign (indicating that additional terms were no longer informative). A similar analysis was done in which we started with a full model containing all of the predictors and removed the predictor with the largest increase in BIC until the change in BIC switched sign (Figure 3.12). In both cases we arrived at the same set of minimal models for each condition. Intriguingly, among the minimal models for each condition, we see a shift between a non-specific preference for AT-rich regions at Log points and specific motif preference at later time points across all conditions (Figure 3.11A). In each condition, from early to late time points, there is a decrease in how much information is provided by the AT-content in terms of predicting Lrp binding. While their relative importance to the model shifts, the minimal variables needed to explain most of the data include a combination of AT-content and established Lrp motifs across all conditions. This suggests that Lrp binding is more non-specific in earlier phases of growth, and only gains specificity upon nutrient limitation and entrance into stationary phase, which also agrees with our observed increase in the number of peaks in later time points. Additionally, this pattern of specificity agrees with Lrp's proposed position of importance as a regulator of the transition to stationary phase. However, since we see the same lack of specificity in LIV\_Log and MIN\_Log (two conditions with dramatically different leucine concentrations), we can conclude that leucine level alone is not sufficient to shift the binding specificity of Lrp, but rather, that other signals

(such as, potentially, energy/carbon source availability) must also be integrated somehow into Lrp's binding.

The performance of the derived models is relatively good; the receiver operator curves, which show the recall for every potential false positive rate, trend toward the upper left corner where a perfect model would be (Figure 3.11B; quantified by area under the curve, ROC-AUC, in Table 3.8). In addition, the Matthews correlation coefficient (MCC), a combined measure of precision and recall which has potential values from 0 to 1, ranges from 0.25 to 0.62 (Table 3.8). These performance metrics were robust to withholding of shuffled subsets of the data, as indicated by minimum and maximum values found in five-fold cross-validation (values in parentheses in Table 3.8). Overall the specificity of these models is much better than their sensitivity, indicating that they perform well in rejecting locations where Lrp does not bind. However, there is still substantial room for improvement in calling Lrp bound sequences. Interestingly, the sensitivity drops in the conditions where specific sequence motifs are more informative. It is likely that we are missing additional features that would improve the sensitivity in these conditions; however, efforts to discover additional sequence determinants of Lrp binding were unsuccessful. This could simply indicate that sequence independent mechanisms, such as the well-established observation of Lrp cooperativity in binding (S. Chen et al., 2005), or recruitment of Lrp by binding of additional factors, could play a role in determining Lrp binding locations.

#### *Lrp binding peak length is relatively invariable*

Given that leucine modulates oligomerization, and that different oligomeric forms of Lrp could result in different regulatory effects, we next investigated whether the length of the called

Lrp ChIP peaks differed across conditions. A Lrp hexadecamer should protect approximately double the amount of DNA protected by a Lrp octamer, if the DNA is wrapped in a linear fashion with no loops, therefore we assumed a longer peak length might indicate hexadecamer rather than octamer binding. However, we detected no change in the called peak lengths across conditions or time points. If we could see a change in peak size, we would be most likely to see it when comparing MIN\_Log and LIV\_Log, conditions with the starkest difference in leucine concentration. Looking at those peaks specifically, it is evident that there is no substantial change in peak length (Figure 3.13). The lack of apparent differences may be due to lack of resolution in our ChIP data or the unknown effect of Lrp oligomerization on DNA binding. It is possible that hexadecamer-bound DNA is still easily accessible between Lrp octamers or that a region is looped out, making differences in peak length an unviable proxy for establishing Lrp oligomerization state.

## **Discussion**

### *Lrp regulates hundreds of genes in distinct categories by direct and indirect mechanisms*

By investigating Lrp activity under several media conditions and time points, and integrating binding data with changes in RNA expression, we are able to present an enhanced view of the Lrp regulon. Our use of a high-quality antibody against native Lrp removes any possibility of epitope tagging hindering native behavior in our experiments, and the use of modern sequencing-based methods provides us with a high resolution snapshot of both Lrp's binding and regulatory activity. We document hundreds of novel targets, and note the especially important effect of indirect regulation at MIN\_Trans and RDM\_Stat. The differences between direct and indirect targets are borne out by the GO-term analysis in which we see a shift between

GO-terms at direct targets (more transport and biosynthesis related genes) and those at indirect targets (flagellum associated genes among others). This could point to organization at a temporal level; the genes needing most urgent regulation (such as those involved directly in importing or generating needed nutrients) may be under direct Lrp control, while genes requiring less urgent modulation and instead governing foraging strategies may be indirectly regulated by Lrp.

In the most straight-forward transcriptional regulatory system, indirect targets should be traceable to a direct target. However, the complicated, interconnected nature of the regulatory system of *E. coli* may explain why we are unable to find connections explaining all Lrp indirect targets. In some cases, there may be another layer of regulation before indirect Lrp targets are affected, or intracellular signaling pathways may be triggered, leading to broader downstream effects, such as changes in metabolite levels. The cases of CysB, LrhA and GadW cleanly illustrate how some indirect regulation is accomplished. Other cases of missed identification may also arise simply due to our incomplete knowledge of the regulons of all *E. coli* transcription factors.

#### *Primed Lrp binding argues for interaction with coregulatory factors*

From our experiments, we identify many points at which Lrp binds the regulatory region of a gene without producing an effect on transcription, and even points at which an apparently identical Lrp binding pattern has no effect on transcription in one condition, but has a substantial effect under another. Given that Lrp binding is enriched in regulatory regions relative to other locations in the genome, this argues against a purely DNA-organizing role for these NAP-type sites. If that was the case, we would expect Lrp binding sites (the majority of which are NAP-

type sites in any condition) to be distributed more evenly across the genome. This poised regulation is also seen for some eukaryotic transcription factors such as the tumor suppressor p53 in binding to the *mdm2* gene (Xiao et al., 1998). Therefore, while Lrp itself is not conserved in eukaryotes, its ability to bind without regulating may have parallels to eukaryotic regulation, suggesting convergent evolution to a similar regulatory scheme. There are several possibilities for why Lrp may not have regulatory function in all cases where it binds, including 1) Lrp acts as a scaffold to interact directly with other proteins which are only present at certain conditions and modulate transcription, 2) Lrp wraps DNA in order to control DNA accessibility of other regulators, reminiscent of eukaryotic histone-like behavior, and/or 3) the presence of Lrp octamer or hexadecamer may control or influence the regulatory behavior of Lrp. We investigated the first possibility by analyzing if certain  $\sigma$  factors or transcription factors might be responsible for the condition-dependent regulation on a global scale. Although we do not see strong global evidence, gene-level studies have previously implicated Lrp in interacting with  $\sigma^{38}$  (Bouvier et al., 1998; Colland et al., 2000). While many potential connections appear to be cases of convergent regulation, we identified a few specific cases where Lrp appears to play a direct role in modulating the effects of other regulators.

There are several data points that indicate direct interaction between Lrp and  $\sigma^{54}$  at MIN\_Trans. First,  $\sigma^{54}$  is most active globally at MIN\_Trans, in agreement with when we see many of the overlapping regulated genes transition from NAP-type to direct targets. As noted above,  $\sigma^{54}$  is unique among the *E. coli*  $\sigma$  factors in that it requires an activator, such as NtrC or PspF (Zhang & Buck, 2015). We also document enrichment for NtrC targets at MIN\_Trans which argues for a role for Lrp in the nitrogen-limitation response. Known NtrC targets account for 33% of genes in the  $\sigma^{54}$  regulon, and almost all of those targets are in operons directly

controlled by Lrp. Activators of  $\sigma^{54}$ , such as NtrC, often bind to an upstream site and require precise looping of the DNA in order to bring the activator in contact with  $\sigma^{54}$ ; in previous studies, the bending has been documented as being intrinsic to the region or looping mediated by IHF (Shingler, 1996). In accordance with the possibility of intrinsic bending, the average AT content upstream of  $\sigma^{54}$  target genes is 70%, with the lowest being at 50% (Larry Reitzer & Schneider, 2001). As previously reported and seen in our data, Lrp is known to bind AT-rich regions preferentially (E B Newman & Lin, 1995). Lrp induces bending of 52° to 135° depending on the size of the binding sites (Q. Wang & Calvo, 1993). Thus, we hypothesize that Lrp may play a role in bending DNA to coordinate NtrC- $\sigma^{54}$  interaction at NtrC targets. This would agree with the connection between Lrp and nitrogen metabolism regulation seen previously in genome-wide studies (Ishihama et al., 2016). Analogous interactions with other transcription or regulatory factors may explain other NAP-type/direct target transitions. For example, Lrp interaction with H-NS is important for regulating rRNA promoters (U. Pul et al., 2007), and Lrp competition with DNA adenine methyltransferase is critical in regulating expression of the *pap* operon, which produces pili (Stacey N. Peterson & Reich, 2008). In addition, non-protein small molecules like ppGpp are known to affect some Lrp-regulated target genes (Traxler et al., 2011). Further studies are needed to investigate Lrp's interactions with other regulatory factors and the alternate mechanisms proposed above.

#### *Lrp binding activity is partially predicted by known sequence motifs*

While we identify a preference for Lrp binding at several related motifs and AT-rich regions, there are still a significant subset of peaks that are not predicted by these models. Attempts to improve Lrp binding prediction from additional sequence determinants were not

successful despite application of several state-of-the-art motif finders. As mentioned above, this could be due to Lrp binding initially at a sequence-specific location, and subsequent Lrp molecules binding due to cooperativity and the high local concentration of Lrp molecules provided by Lrp's oligomeric nature. Alternatively, Lrp itself may be recruited by other proteins. Due to Lrp's relatively high non-specific DNA binding affinity, especially under rich conditions (Shaolin Chen, Hao, et al., 2001), it is reasonable to find that not all of its binding locations can be predicted based on sequence alone. It is again important to note that the switch in DNA-binding specificity occurs regardless of the levels of leucine, suggesting that other small molecule regulators (Hart & Blumenthal, 2011) or potentially post-translational modifications (Baeza et al., 2014; Potel, Lin, Heck, & Lemeer, 2018) may play a role in Lrp regulatory activity. Additionally, despite extensive effort, we were unable to identify any sequence determinants capable of reliably explaining Lrp regulatory activity, either through predicting transitions from poised to active regulation, or distinguishing Lrp activation from Lrp repression. Possible mechanisms for this behavior include interactions with condition-specific factors that bind near the multifunctional Lrp sites (many potential partners have likely not yet been characterized), condition-dependent DNA looping triggered by the binding of Lrp to nearby sites or by octamer-hexadecamer transitions, or post-translational modifications to Lrp itself. Dissecting the detailed molecular mechanisms underlying the binding and regulatory landscape that we have revealed here will be a fruitful area for future research.

## **Materials and Methods**

### *Strains and media*

The WT strain used in this study was *E. coli* K-12 MG1655 (ATCC 47076). The Lrp deletion strain was constructed by homologous recombination resulting in the insertion of kanamycin resistance cassette (Datsenko & Wanner, 2000). Primers used for strain construction and validation are listed in Table 3.9. The *lrp::kanR* strain was validated by sizing of the P965/P1568/P1569 products and Sanger sequencing.

All routine cell growth during cloning was done in LB medium (10 g/liter tryptone, 5 g/liter yeast extract, 5 g/liter NaCl) or on LB plates (LB medium plus 15 g/liter Bacto agar) supplemented with 50 µg/mL kanamycin or 100 µg/mL ampicillin (both from US Biological; Salem, MA) as required. For the ChIP-seq and RNA samples, a single colony of wild type *E. coli* or the *lrp::kanR* strain was inoculated into MOPS media (Teknova; Hollister, CA) with 0.04% glucose (Neidhardt, Bloch, & Smith, 1974) and grown overnight. The cells were then back-diluted to OD<sub>600</sub>=0.003 in 100 mL of the appropriate target media. Experiments were performed in MOPS with 0.2% glycerol (the MIN media condition), MOPS with 0.04% glycerol and 0.2% (weight/volume) each leucine (Amresco; Solon, OH), isoleucine (Alfa-Aesar; Haverhill, MA) and valine (Amresco; Solon, OH; the LIV condition), or MOPS plus 0.4% glycerol, ACGU and EZ supplements (Teknova; Hollister, CA; the RDM condition). Media conditions are summarized in Table 3.10.

The cells were grown at 37°C with shaking (200 rpm) until the OD<sub>600</sub> was between 0.15 and 0.25 (for log phase samples), between 1.8 and 2.2 (for transition point in MIN or LIV media), between 2.3 and 2.7 (for transition point in RDM), or 12 hours past the log point (for stationary phase samples). The OD<sub>600</sub> range for transition point harvest was determined by monitoring the growth of cells grown in conditions identical to the experiment and selecting the

point in the OD600 range during which exponential growth becomes non-linear when visualized on a log scale.

### *ChIP-seq*

At the appropriate time, either WT or *lrp::kanR* cells were cross-linked by adding formaldehyde (37% Sigma-Aldrich; St. Louis, MO) to a final concentration of 1% (vol/vol) and incubated with shaking for 15 minutes at room temperature. Formaldehyde cross-linking was quenched by addition of Tris (pH 8) to a final concentration of 280 mM and incubation with shaking at room temperature for 10 minutes. The culture was then immediately centrifuged for 5 minutes at 5500xg at 4°C. The pellet was washed twice with 30 mL ice cold TBS (50 mM Tris, 150 mM NaCl, pH 7.5) before being resuspended in 1 mL TBS. Following a 3 minute centrifugation at 10,000xg at 4°C and removal of the supernatant, the pellet was flash-frozen in a dry ice/ethanol bath and then stored at -80°C. Two biological replicates, grown on different days, were prepared for each condition.

The cell pellet was resuspended in lysis buffer (PBS, 0.1% Tween 20, 1 mM EDTA, 1x Complete Mini EDTA-free Protease Inhibitors (Roche; Basel, Switzerland), 0.6 mg lysozyme (Amresco; Solon, OH)), vortexed for 3 seconds, and incubated at 37°C for 30 minutes. The sample was then sonicated in 3 bursts of 10 seconds each at 25% power (Branson Digital Sonifier). Cellular debris was removed by centrifugation at 16,000xg for 10 minutes at 4°C. As an input sample, 50 µL of the supernatant was removed and mixed with EDTA to 8.6 mM and 235 µL Elution Buffer (50 mM Tris (pH 8), 10 mM EDTA, 1% SDS (vol/vol)). The remainder of the lysate was added to 50 µL pre-washed SureBeads Protein G magnetic beads (Bio-Rad; Hercules, CA) and rocked for 1 hour at room temperature for pre-clearing. A separate aliquot of

100  $\mu$ L of pre-washed SureBeads Protein G magnetic beads was incubated with 10  $\mu$ g Lrp monoclonal antibody (Neoclone; Madison, WI) for 10 minutes at room temperature with rocking and then washed thrice with PBS/0.1% Tween-20 before the pre-cleared supernatant was added. The bead/lysate mixture was again incubated with rocking for 1 hour at room temperature. The beads were then washed thrice with PBS/0.1% Tween-20. To elute the cross-linked Lrp/DNA complexes, the beads were resuspended in 285  $\mu$ L of Elution Buffer and incubated at 65°C for 20 min, vortexing every 5 minutes. The resulting eluate was incubated overnight at 65°C to reverse the cross-links.

The sample was treated with 0.05 mg RNase A (Thermo Fisher; Waltham, MA) for 2 hours at 37°C, then 0.2 mg Proteinase K (Thermo Fisher; Waltham, MA) for 2 hours at 50°C before the DNA was isolated by phenol-chloroform extraction and ethanol precipitation. The samples were quantified (QuantiFluor dsDNA Kit, Promega; Madison, WI) and prepared for sequencing using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB; Ipswich, MA). The library was checked for quality by 2% agarose gel electrophoresis using GelRed stain (Biotium; Fremont, CA). Samples were pooled and the sequencing performed on an Illumina NextSeq -500, with 38x37 bp paired end reads. We obtained at least three million reads that passed all filters and aligned properly to the genome per biological replicate with an average of nine million reads per replicate (Table S10). Input samples were treated identically to the ChIP extracted samples beginning at the RNase A treatment.

### *RNA-seq*

For RNA-seq samples in both WT and *lrp::kanR* cells, 2.5 ml of culture was removed when cells had reached the appropriate OD and mixed with 5 mL Qiagen RNAProtect Bacteria

Reagent (Qiagen; Hilden, Germany), vortexed, incubated 5 minutes at room temperature, and then centrifuged for 10 minutes at 5,000xg in a fixed angle rotor at 4°C. The supernatant was removed and the pellet was flash-frozen in a dry ice/ethanol bath before being stored at -80°C. The pellet was resuspended in TE and treated with 177 kilounits Ready-Lyse Lysozyme Solution (Epicentre; Madison, WI) and 0.2 mg Proteinase K (Thermo Fisher; Waltham, MA) for ten minutes at room temperature, vortexing every two minutes. The RNA was purified using the Zymo RNA Clean and Concentrator kit (Zymo; Irvine, CA), treated with 5 units Baseline Zero DNase (Epicentre; Madison, WI), in the presence of RNase Inhibitor (NEB; Ipswich, MA), for 30 minutes at 37°C, and then again purified with the Zymo RNA Clean and Concentrator kit. RNA quality was assessed by electrophoresis in a denaturing agarose-guanidinium gel (Goda & Minton, 1995). rRNA depletion was performed using the Ribo-Zero rRNA Removal Kit for Bacteria (Illumina; San Diego, CA), halving all reagent and input quantities but otherwise following the manufacturer's instructions. cDNA synthesis and sequencing library preparation were performed following the NEBNext Ultra Directional RNA Library Prep Kit (NEB; Ipswich, MA). The library was checked for quality by 2% agarose gel electrophoresis using GelRed stain (Biotium; Fremont, CA). Samples were pooled and the sequencing performed on a NextSeq -500 at the University of Michigan's DNA Sequencing Core Facility.

#### *Preprocessing and alignment of ChIP-seq data*

Sequencing analysis was performed by Michael Wolfe. Full methods description can be found in (Kroner, Wolfe, & Freddolino, 2018). Removal of sequencing adaptors was accomplished using CutAdapt version 1.8.1, and low quality reads were discarded using Trimmomatic version 0.32. We evaluated the quality of raw and preprocessed sequencing fastq

files using FastQC version 0.10.1 and MultiQC version 1.2. Samples were aligned to the MG1655 U00096.2 genome with modifications for the ATCC 47076 variant (Freddolino, Amini, & Tavazoie, 2012) using bowtie version 2.1.0.

### *Calculation of ChIP-seq summary signal*

The raw coverage (calculated from alignments of ChIP-extracted and input samples individually) was scaled using the median coverage of the entire genome, and then the raw enrichment (RE) was obtained from the log<sub>2</sub> ratio of the scaled extracted to the scaled input data. Since the WT and *lrp::kanR* samples were not paired, we calculated a raw subtracted Lrp enrichment signal (RSE) for each potential combination (four total) of WT and *lrp::kanR* samples by subtracting the *lrp::kanR* RE (if positive) from the WT RE. Each of these raw subtracted Lrp enrichment signals was converted to robust Z-score estimates to allow comparisons between conditions using the following formula:

$$RZ_{(n)} = \frac{RSE_{(n)} - \text{median}(RSE)}{\text{median}(|RSE_{(n)} - \text{median}(RSE)|)} \cdot 1.4826$$

The four robust Z-score replicates were averaged to produce the final Lrp occupancy signal used in later analysis. Reproducibility of both the RE and RSE for each replicate can be seen in Figure 3.14A.

### *Determination of high-confidence Lrp binding sites*

To identify areas of high-confidence Lrp binding, we established three required criteria for Lrp enrichment: 1) the enrichment must be technically reproducible, 2) the enrichment must be above the input background, 3) the enrichment must be biologically reproducible.

Technical reproducibility was established by calculating the RSE (as above) for 1000 bootstrap replicates obtained by sampling with replacement from the aligned reads for each set of extracted and input samples separately. We then calculated a Z-score, with the null hypothesis that the RSE is normally distributed around 0, and converted the Z-score to a p-value using a one-sided Z-test from the *scipy.stats* python package. The p-values were FDR corrected using the procedure described by Benjamini and Hochberg (Hochberg & Benjamini, 1990) and a region was considered technically reproducible if the q-value was less than 0.001.

Given some non-specific antibody activity apparent in the *lrp::kanR* strain pulldowns, it is important to establish Lrp-specific enrichment above the input. Using the robust Z-scores calculated previously for each potential combination of WT and *lrp::kanR*, we tested for enrichment of the robust Z-score above the median robust Z-score for that pair using a one-sided Z-test and FDR corrected the p-value to a q-value as above. Regions with a q-value less than 0.001 were considered to have enrichment above background.

Biological reproducibility was tested by calculating the irreducible discover rate (IDR) (Li Q, 2011) for each data point between the robust Z-score signals for all four WT/*lrp::kanR* subtractions. A region was considered biologically reproducible if the FDR-corrected IDR q-values was less than 0.01 for both possible combinations of RSE replicates.

A region was considered a Lrp binding site if it passed the biological reproducibility filter and if at least one of the four WT/*lrp::kanR* RSE combinations passed both the technical and enrichment filters. If regions were within 30 base pairs, they were combined into one called Lrp-binding site. We verified the applied cutoffs by manually inspecting called and candidates peaks under a number of different threshold combinations. An example peak in comparison to a non-Lrp-specific peak can be seen in Figure 3.14B, C.

### *Preprocessing of RNA-Seq data*

Similar to the ChIP-Seq reads, sequencing adapters were removed from all sequences using CutAdapt version 1.8.1, and low quality reads were trimmed with Trimmomatic version 0.32. We again assessed the quality of the raw and preprocessed fastq files using FastQC version 0.10.1 and MultiQC version 1.2. In some of our samples up to 70% of our RNA-seq reads were ribosomal reads or the highly abundant RNA products from *ssrA* and *ssrS*. To avoid having variations in ribosome depletion efficiency affect proper normalization, we filtered these highly abundant RNA reads by aligning all RNA-seq reads to the same ATCC 47076-modified version of the U00096.2 genome used for the ChIP-Seq data using bowtie version 2.1.0. New fastq files were written that only included RNA-seq reads that did not overlap with ribosomal regions in a strand specific manner, and these files were used for downstream gene expression analysis. At least two million reads remained for each replicate after this filtering step.

### *Determination of Lrp-dependent changes in Transcription*

To determine Lrp-dependent changes in transcription, we first established gene-centric quantification of RNA expression for all samples using kallisto version 0.43.0. We then employed kallisto's companion post-processing data analysis software sleuth to model the transcript abundance for each condition and time point. Differential expression between the WT and the *lrp::kanR* strains was established using a Wald test on the genotype type of a model where transcript abundance is dependent on genotype; the *lrp::kanR* is the baseline condition. Transcripts that passed both an FDR corrected p-value of less than 0.05 and a genotype term

magnitude of greater than 0.5 were considered as having a significant Lrp-dependent RNA expression change under that condition.

To visualize Lrp-dependent changes in expression more intuitively, we reported the log<sub>2</sub> ratio of the average WT transcripts per million (TPM) over *lrp::kanR* TPM. To generate the error bars on all RNA expression bar plots, the log<sub>2</sub>(WT/KO) TPM was calculated for all 100 bootstrap replicates from kallisto, and a percentile based 95% confidence interval from these bootstrap replicates was taken to be the lower and upper bounds of the ratio.

### *Antibody development and testing*

The monoclonal antibody used in these experiments was developed via a contract with NeoClone (Madison, WI). Using purified His-tagged Lrp, several rounds of potential antibodies were developed. The potential antibodies were tested for cross-reactivity with the known Lrp homologues AsnC and YbaO by ELISA at NeoClone. We used an *in vitro* DNA pull-down assay to ensure that the potential antibodies did not inhibit Lrp-DNA binding (Figure 3.15A). In addition, we tested the antibody for use in Western blotting (Figure 3.15B), and confirmed that the antibody did not bind the oligomerization interface by observing bands corresponding to Lrp octamers and hexadecamers in native Western blots.

### *Filtering of genes into Lrp-dependent categories*

For gene target filtering, we established four categories through a two-level filtering scheme (Figure 3.3A). We first tested whether the gene had a Lrp-dependent change in RNA expression by comparing the target gene's expression in WT and *lrp::kanR* strains using a Wald test as described above. We next asked if the gene had a high confidence Lrp binding site, as

defined above, within the regulatory region, defined as 500 bp upstream and downstream from the annotated transcription start site (TSS; annotations from RegulonDB (Gama-Castro et al., 2016)). If multiple TSSs were annotated for a gene, the regulatory region included 500 bp upstream of the most distal TSS and 500 bp downstream of the most proximal TSS.

Using our high confidence Lrp binding regions, we then determined which regulatory regions fell within a high-confidence Lrp binding site; any regulatory region that overlapped with a high-confidence Lrp binding site was classified as bound by Lrp. Genes were thus categorized as either a direct target (RNA expression change and Lrp binding), an indirect target (RNA expression change but no Lrp binding), a NAP target (no RNA expression change but Lrp binding), or unconnected to Lrp (neither RNA expression change or Lrp binding).

For comparing enrichment of Lrp targets with  $\sigma$  factor targets, we used permutation tests as noted in the text, implemented using custom python scripts and 1000-10000 permutations. When testing for enrichment across several different  $\sigma$  factors, we corrected for multiple hypothesis testing using the statsmodels.sandbox.stats.multicomp.multipletests module using the Benjamini-Hochberg method (Hochberg & Benjamini, 1990; Seabold & Perktold, 2010).

All plots except where noted were created using ggplot2 (Wickham, 2016) or Matplotlib (Hunter, 2007).

### *Data Availability*

Raw sequencing data has been deposited in the Gene Expression Omnibus with accession number GSE111874. Source code for standalone analysis of sequencing data are publicly available from [https://github.com/freddolino-lab/2018\\_Lrp\\_ChIP](https://github.com/freddolino-lab/2018_Lrp_ChIP).

## Tables

Condition	Total Genes Significantly Upregulated by Lrp	Total Genes Significantly Downregulated by Lrp
MIN_Log	164 (3.52%)	162 (3.48%)
MIN_Trans	423 (9.08%)	590 (12.67%)
MIN_Stat	63 (1.35%)	63 (1.35%)
LIV_Log	109 (2.34%)	217 (4.66%)
LIV_Trans	87 (1.87%)	123 (2.64%)
LIV_Stat	80 (1.72%)	68 (1.46%)
RDM_Log	36 (0.77%)	84 (1.80%)
RDM_Trans	58 (1.25%)	21 (0.45%)
RDM_Stat	728 (15.63%)	622 (13.35%)

**Table 3.1: Genes with significant Lrp-dependent changes in expression.** Percentage is out of the total number of genes in *E. coli* (4658).

<i>aceA</i>	<i>ftsH</i>	<i>leuC</i>	<i>oppD</i>
<i>aroA</i>	<i>fumC</i>	<i>leuD</i>	<i>oppF</i>
<i>astE</i>	<i>gadC</i>	<i>livH</i>	<i>pnp</i>
<i>atpG</i>	<i>glgA</i>	<i>mazE</i>	<i>pntB</i>
<i>deoD</i>	<i>glgP</i>	<i>metI</i>	<i>rplO</i>
<i>dppB</i>	<i>gltD</i>	<i>nuoC</i>	<i>rpmD</i>
<i>dppC</i>	<i>ilvA</i>	<i>nuoE</i>	<i>rpmJ</i>
<i>dppD</i>	<i>ilvD</i>	<i>nuoF</i>	<i>rpsE</i>
<i>dppF</i>	<i>ilvE</i>	<i>opgH</i>	<i>rpsO</i>
<i>fimD</i>	<i>ilvM</i>	<i>oppB</i>	<i>secY</i>
<i>fimG</i>	<i>leuB</i>	<i>oppC</i>	<i>yeeD</i>

**Table 3.2: Ambiguous indirect targets.** Genes classified as indirect Lrp targets with internally annotated transcription start sites and in an operon whose first gene is a direct Lrp target.

Condition	p-value
MIN_Log	$< 1.0 \times 10^{-3}$
MIN_Trans	$< 1.0 \times 10^{-3}$
MIN_Stat	$< 1.0 \times 10^{-3}$
LIV_Log	$< 1.0 \times 10^{-3}$
LIV_Trans	$2.0 \times 10^{-3}$
LIV_Stat	$< 1.0 \times 10^{-3}$
RDM_Log	$< 1.0 \times 10^{-3}$
RDM_Trans	$< 1.0 \times 10^{-3}$
RDM_Stat	$< 1.0 \times 10^{-3}$

**Table 3.3: *Lrp* preferentially binds regulatory regions.** Results of permutation test for enrichment of *Lrp* binding in regulatory regions.

Condition	Percentage in gene region	Percentage in transcription-unit	Percentage in intergenic region
Genome-wide	45.8	23.8	30.4
MIN_Log	41.3	17.3	41.3
MIN_Trans	44.9	26.0	29.1
MIN_Stat	47.6	23.3	29.1
LIV_Log	28.6	14.3	57.1
LIV_Trans	38.3	22.7	39.0
LIV_Stat	39.8	23.5	36.7
RDM_Log	48.4	29.0	22.6
RDM_Trans	50.0	19.8	30.2
RDM_Stat	48.6	22.9	28.6

**Table 3.4: Locations of non-regulatory region peaks.** Percentages of non-regulatory region peaks that annotate to other regions of the genome. All categories are mutually exclusive: if a position is in regulatory region, it cannot be classified as being in a gene; if a position is in a gene, it cannot be in a gene; if a position is in a gene, it cannot be in a transcription unit. Anything not assigned to either of those categories is labeled intergenic. Percentages genome-wide were determined at a 1 bp resolution.

	Total indirect targets	Number (Percentage) of indirect targets explained by direct targets	Fold-enrichment	p-value	q-value
MIN_Log	152	6 (3.95%)	4.36	0.023	0.041
MIN_Trans	509	40 (7.86%)	1.4	0.011	0.033
MIN_Stat	55	1 (1.82%)	1.07	0.797	1
LIV_Log	190	14 (7.37%)	7.06	0.001	0.009
LIV_Trans	98	1 (1.02%)	4.89	0.337	0.505
LIV_Stat	84	0 (0%)	0	1	1
RDM_Log	49	2 (4.08%)	14.66	0.021	0.041
RDM_Trans	34	0 (0%)	0	1	1
RDM_Stat	919	13 (1.41%)	2.14	0.005	0.022

**Table 3.5: Indirect target annotation.** Numbers of indirect targets in different conditions and results of a permutation test for enrichment of indirect Lrp targets among known targets of Lrp direct targets. Fold enrichment is calculated by dividing the fraction of indirect targets regulated by direct targets, by the fraction of all classified genes that are regulated by direct targets.

		$\sigma^{24}$	$\sigma^{28}$	$\sigma^{32}$	$\sigma^{38}$	$\sigma^{54}$	$\sigma^{70}$
MIN_Log	q-value	1	0.699	1	0.542	0.485	1
	Fold change	0	1.39	0.56	1.35	1.77	0.99
MIN_Trans	q-value	0.923	0.705	1	0.947	0.076	1
	Fold change	1.01	1.14	0.86	1.00	1.79	0.92
MIN_Stat	q-value	1	0.705	1	0.606	0.485	1
	Fold change	0	1.33	0.66	1.46	2.24	0.94
LIV_Log	q-value	1	1	1	0.705	1	0.485
	Fold change	0	0	0	1.09	0	1.20
LIV_Trans	q-value	1	0.699	1	0.705	1	0.485
	Fold change	0	1.51	0	1.26	0	1.17
LIV_Stat	q-value	0.947	0.705	1	1	0.520	1
	Fold change	1.01	1.51	0	0.90	2.42	0.99
RDM_Log	q-value	1	0.684	1	1	0.485	0.520
	Fold change	0	1.94	0.65	0	2.59	1.19
RDM_Trans	q-value	1	1	1	1	0.699	0.076
	Fold change	0	0	0	0.39	1.67	1.39
RDM_Stat	q-value	1	1	0.542	0.485	0.304	1
	Fold change	0.50	0	1.34	1.44	1.89	0.83

**Table 3.6: Results of permutation test for enrichment of direct Lrp targets relative to NAP-type targets within the known  $\sigma$  factor regulons at each condition.** Fold change is calculated by dividing the fraction of bound  $\sigma$  factor targets (either direct or NAP) which are classified as direct targets, by the overall fraction of Lrp-bound targets which are direct targets.

		$\sigma^{24}$	$\sigma^{28}$	$\sigma^{32}$	$\sigma^{38}$	$\sigma^{54}$	$\sigma^{70}$
MIN_Log	q-value	1	0.802	1	0.802	0.802	1
	Fold change	0	1.42	0.38	1.38	1.80	1.01
MIN_Trans	q-value	0.802	0.701	0.802	1	0.016	1
	Fold change	1.25	1.67	1.25	0.93	3.05	0.72
MIN_Stat	q-value	1	1	0.942	0.802	0.701	1
	Fold change	0	0	1.06	1.56	2.39	0.91
LIV_Log	q-value	1	1	1	0.920	1	0.606
	Fold change	0	0	0	1.01	0	1.23
LIV_Trans	q-value	1	0.802	1	0.891	1	0.802
	Fold change	0	1.82	0	1.22	0	1.15
LIV_Stat	q-value	0.802	1	1	0.802	0.606	1
	Fold change	1.56	0	0	1.39	3.75	0.83
RDM_Log	q-value	1	1	1	1	0.802	0.606
	Fold change	0	0	1	0	2	1.27
RDM_Trans	q-value	1	1	1	1	1	0.310
	Fold change	0	0	0	0.70	0	1.41
RDM_Stat	q-value	1	1	1	0.059	0.606	1
	Fold change	0.80	0	0.27	2.29	2	0.79

**Table 3.7: Results of permutation tests for enrichment of direct Lrp-activated targets relative to direct Lrp-repressed and NAP-type targets within the known  $\sigma$  factor regulons at each condition.** Fold change is calculated as for Table 3.6 except that the number of direct targets is replaced with the number of direct Lrp-activated targets.

Condition	MCC	Specificity	Sensitivity	ROC-AUC
LIV_Log	0.62 (0.52-0.67)	0.85 (0.69-0.92)	0.81 (0.62-0.93)	0.85 (0.70-0.91)
LIV_Trans	0.37 (0.31-0.46)	0.77 (0.73-0.82)	0.63 (0.55-0.77)	0.79 (0.74-0.84)
LIV_Stat	0.39 (0.27-0.55)	0.77 (0.70-0.85)	0.65 (0.59-0.71)	0.79 (0.73-0.87)
MIN_Log	0.35 (0.28-0.43)	0.74 (0.69-0.80)	0.64 (0.55-0.71)	0.77 (0.74-0.78)
MIN_Trans	0.26 (0.15-0.29)	0.72 (0.67-0.74)	0.56 (0.49-0.62)	0.69 (0.61-0.74)
MIN_Stat	0.25 (0.20-0.28)	0.72 (0.68-0.76)	0.56 (0.46-0.63)	0.69 (0.66-0.73)
RDM_Log	0.44 (0.30-0.47)	0.78 (0.74-0.83)	0.70 (0.59-0.80)	0.82 (0.75-0.87)
RDM_Trans	0.30 (0.19-0.34)	0.73 (0.72-0.74)	0.60 (0.48-0.64)	0.73 (0.68-0.76)
RDM_Stat	0.33 (0.09-0.48)	0.74 (0.64-0.84)	0.63 (0.36-0.74)	0.77 (0.68-0.81)

**Table 3.8: Performance of Lrp binding site prediction models.** The minimum and maximum values from 5-fold cross-validation for each metric are indicated in parentheses.

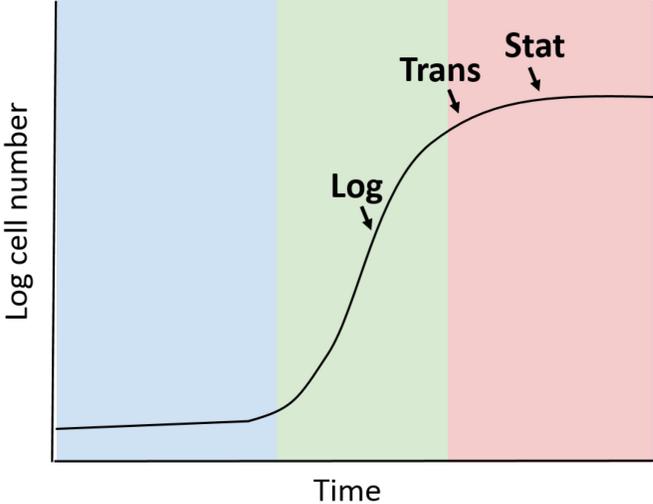
Identifier	Sequence	Notes
P1582	TCAGACAGGAGTAGGGAAGGAATACAGAGAGACAATAAT ATGTGTAGGCTGGAGCTGCTTC	Generate Kan cassette to delete <i>lrp</i>
P1583	GAGTGTAATCAAAATACGCCGATTTTGCACCTGTTCCGTG CATATGAATATCCTCCTTA	
P965	GAACTTCGAAGCAGCTCCAG	Test <i>lrp::kanR</i> deletion
P1568	CAAGGCAACGGTCTTCTCAC	
P1569	CCTGGCTCAAGAAAGGCTCT	

**Table 3.9: Primers used for *lrp::kanR* construction.**

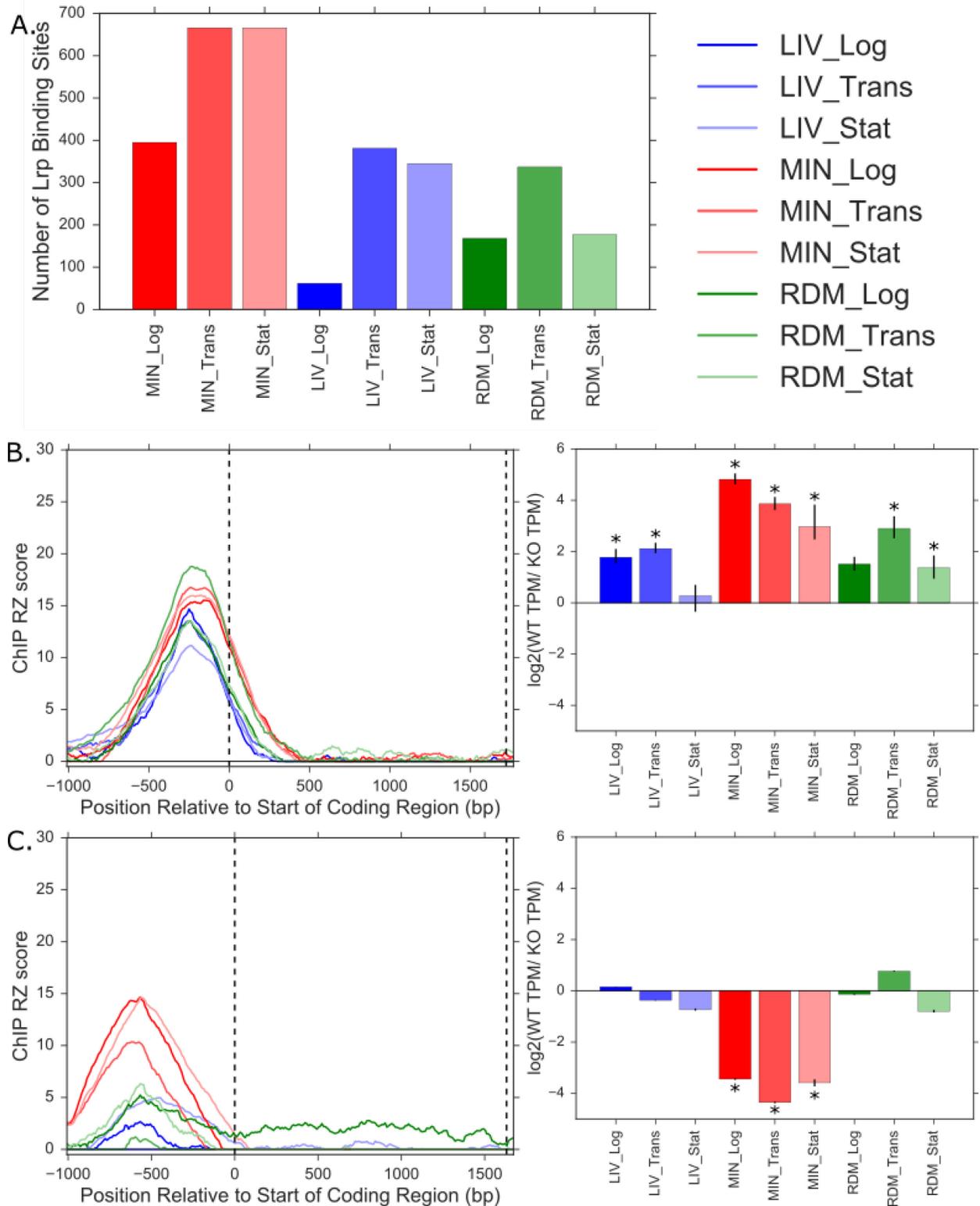
	Pre-growth media	Minimal	Min+LIV	RDM
Media Base	MOPS	MOPS	MOPS	MOPS RDM
Carbon Source (weight/volume)	0.04% glucose	0.2% glycerol	0.2% glycerol	0.4% glycerol
Leucine, Isoleucine, Valine Supplement			0.2% (weight/volume)	

**Table 3.10: Media conditions for cell growth.** All MOPS media formulations are based on (Neidhardt, Bloch et al. 1974).

**Figures**

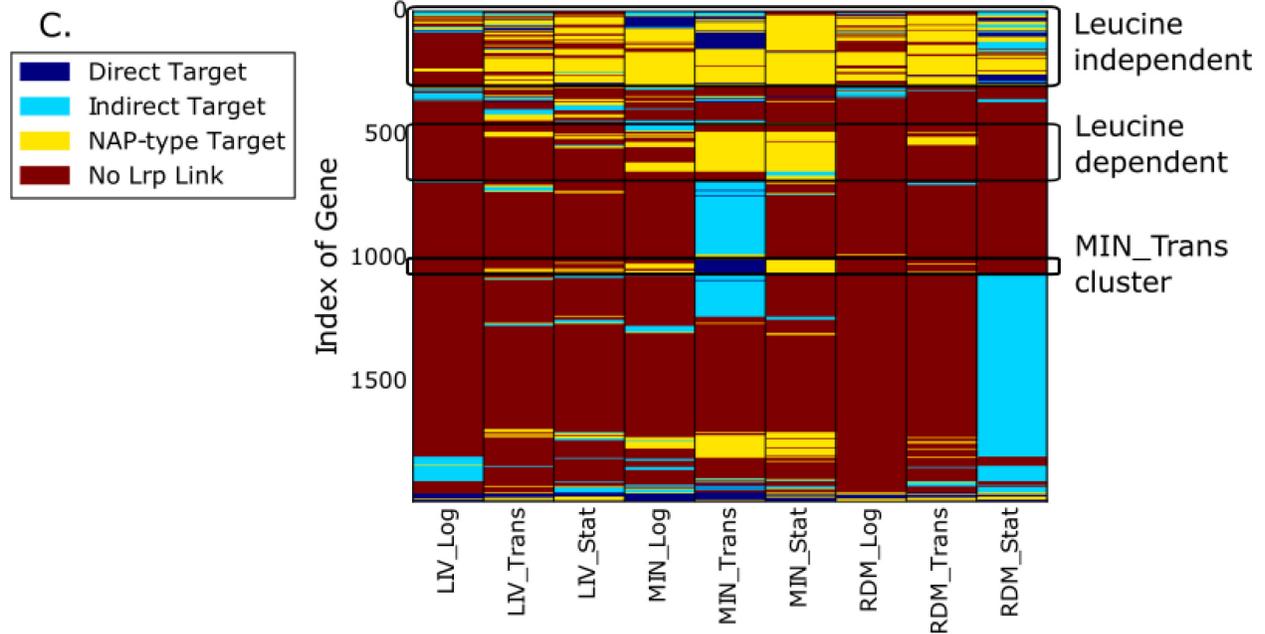
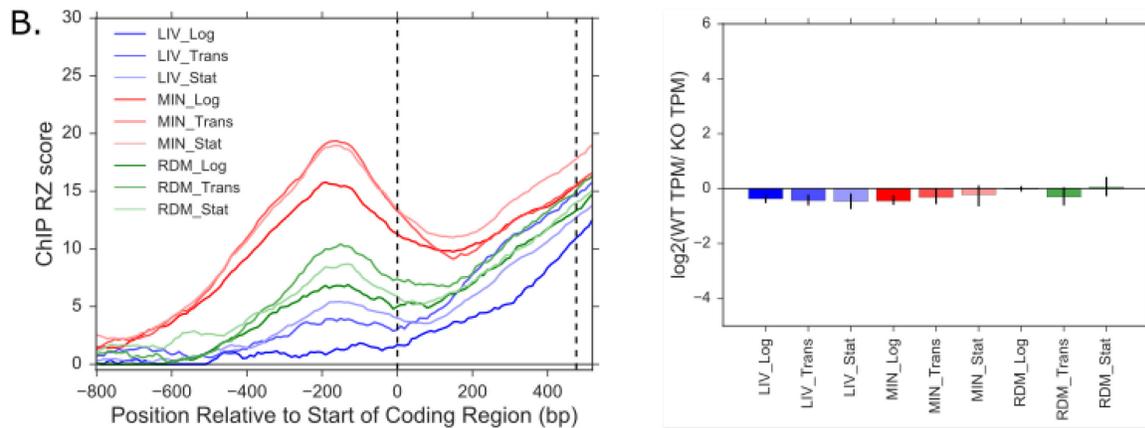
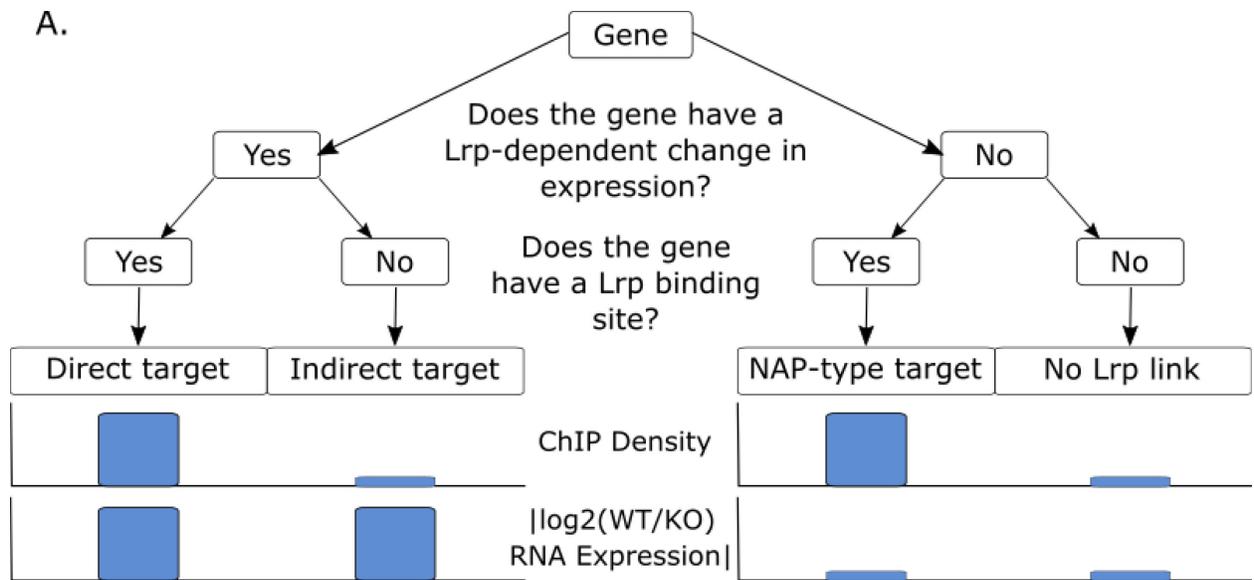


*Figure 3.1: Depiction of experimental time points.*

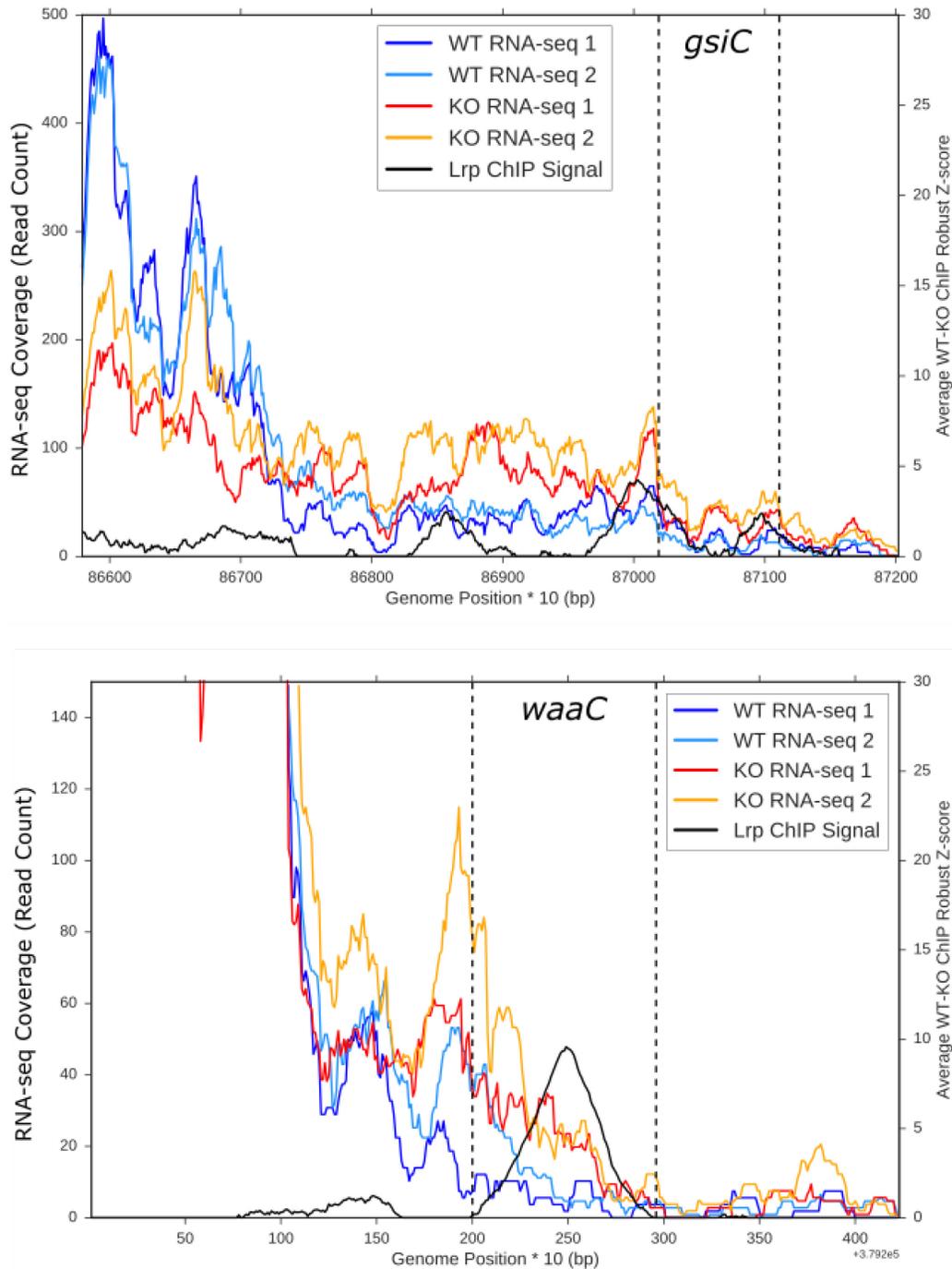


**Figure 3.2: ChIP-seq data shows agreement with previous data and reveals novel *Lrp* binding sites.** *A.* Total number of non-overlapping high-confidence *Lrp* binding sites identified in each condition. *B.* ChIP robust Z-score (left) and RNA-seq expression change ( $\log_2(\text{WT}/\text{KO})$ ); right)

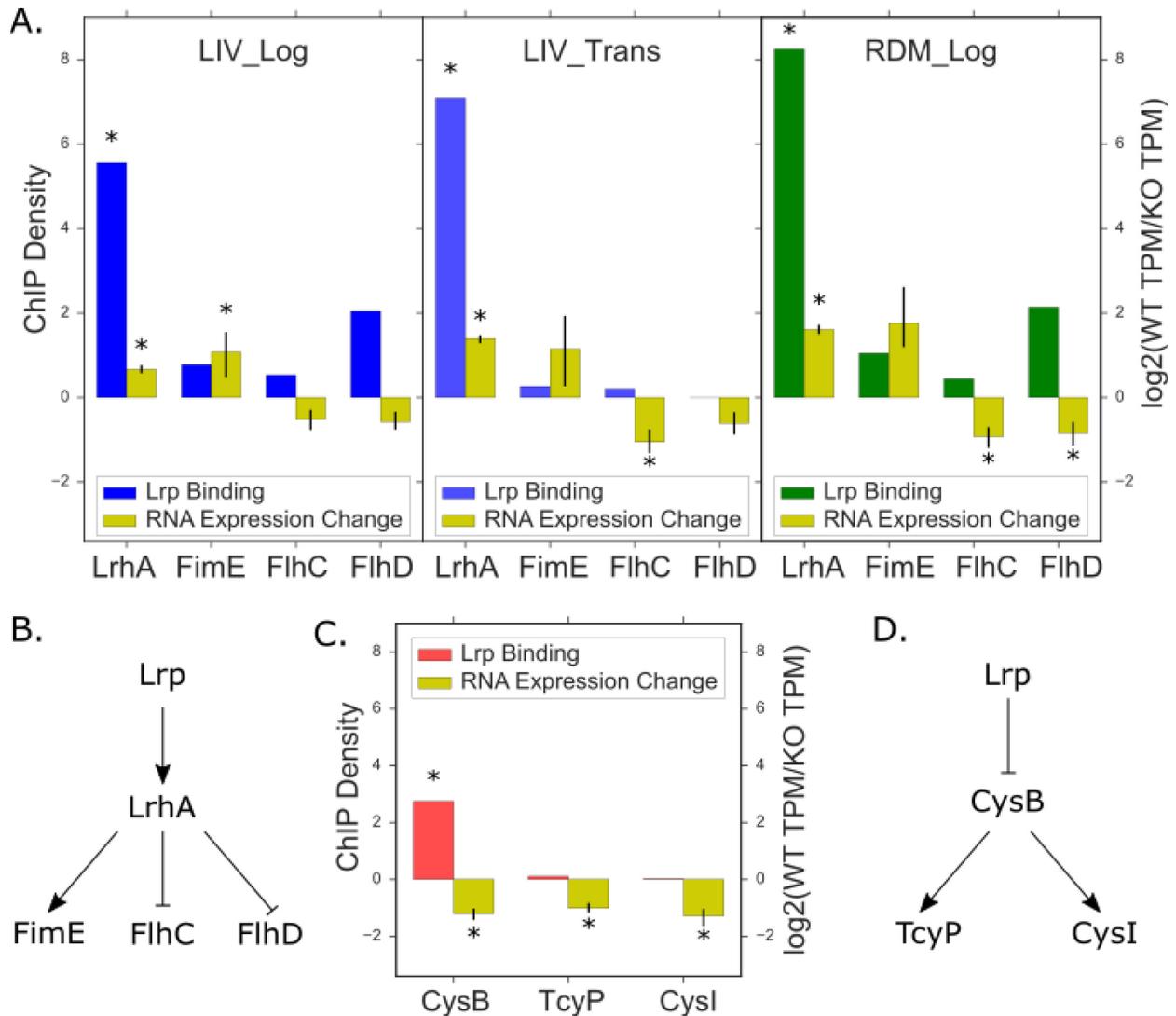
for known *Lrp* activated target *ilvI*. Dashed vertical lines on the ChIP robust Z-score graph mark the start and end of the gene coding region. Error bars for the RNA-seq data indicate a percentile based 95% confidence interval from 100 bootstrap replicates of TPM estimates. Stars indicate a significant difference in RNA abundance between WT and *lrp::kanR* strains (Wald Test  $q$ -value of  $< 0.05$  and a genotype log fold change coefficient magnitude of  $> 0.5$ ; see Methods for details). **C.** ChIP robust Z-score (left) and RNA-seq expression change ( $\log_2(WT/KO)$ ; right) for known *Lrp* repressed target *oppA*, panels as in **B**.



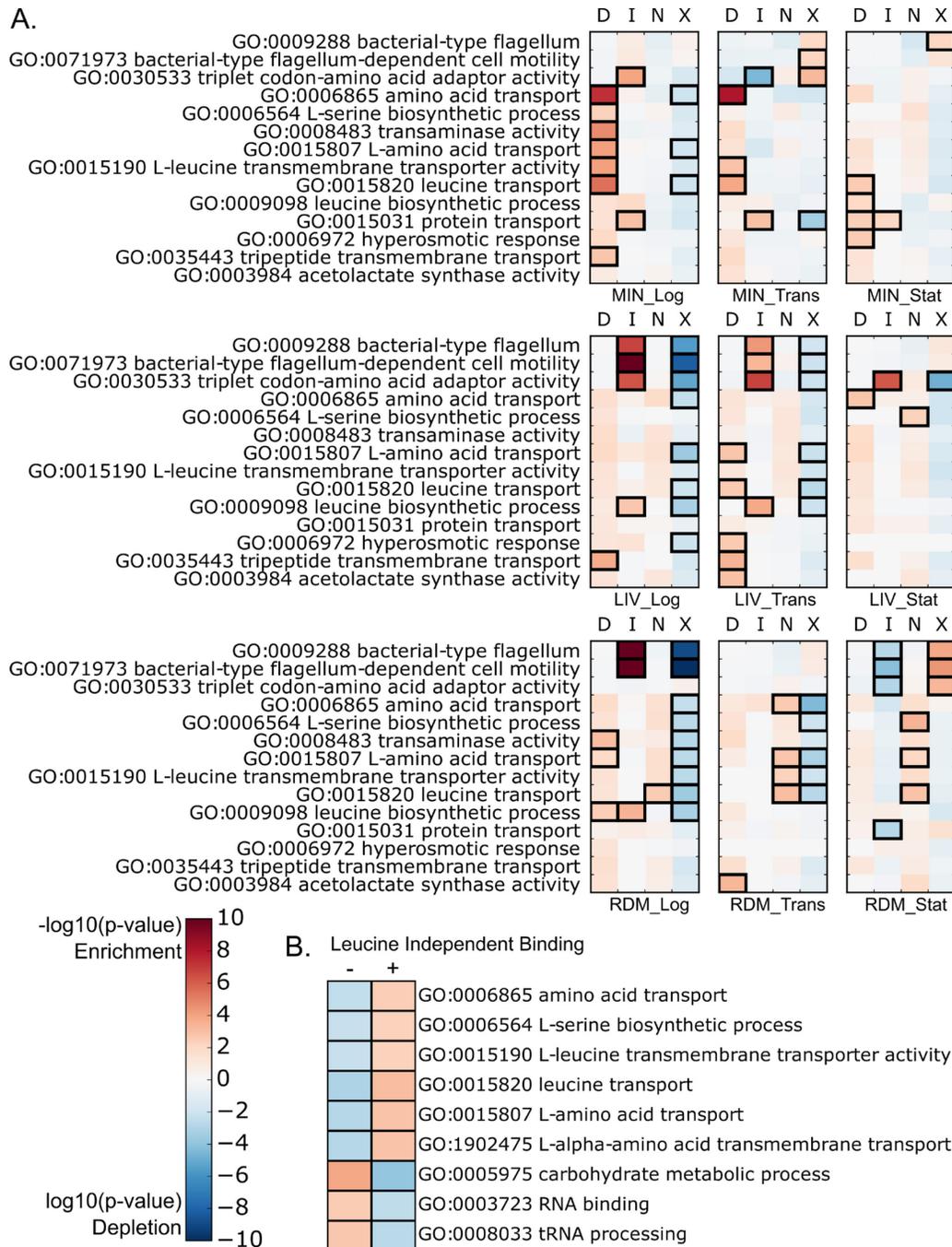
**Figure 3.3: Lrp regulates genes both directly and indirectly.** *A. Schematic showing how genes were categorized: direct targets of Lrp (Lrp-bound regulatory region and with a significant RNA expression change between WT and *lrp::kanR* cells), indirect targets (not bound but with a significant RNA expression change), NAP targets (bound but with no significant RNA expression change), or not linked (not bound and no significant RNA expression change). Filtering was done independently for each condition. B. ChIP robust Z-score (left) and RNA-seq expression change ( $\log_2(WT/KO)$ ; right) for Lrp NAP-type target *ybjN* (as in Fig 1B). C. Heat map indicating how each gene was classified in the nine experimental conditions. Genes with no Lrp link in any condition were removed from visualization. Genes were hierarchically clustered using a Manhattan distance metric and average linkage clustering. Black boxes mark out notable clusters of genes: those with leucine-dependent or -independent binding and those that are direct targets only under MIN\_Trans.*



**Figure 3.4: Intragenic Lrp peaks do not systematically affect transcription.** ChIP RZ signal and RNA-seq coverage for WT and KO cells over two operons in the MIN\_Trans condition. The coding region of the gene which shows a different behavior in response to Lrp compared to the first gene in the operon is indicated by dashed black lines. Top panel shows the *iaaA/gsiABCD* operon, where *gsiC* is Lrp-repressed and *iaaC* is an indirect Lrp-activated target. Bottom panel shows the *rfaD/waaFCL* operon, where *rfaD* is not regulated by Lrp and *waaC* is Lrp-repressed. Note axis definitions for RNA-seq (left) and ChIP-seq (right) data.



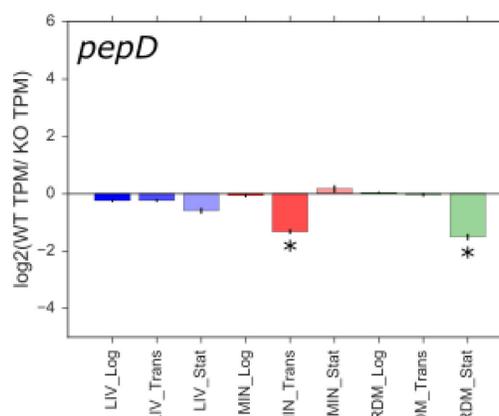
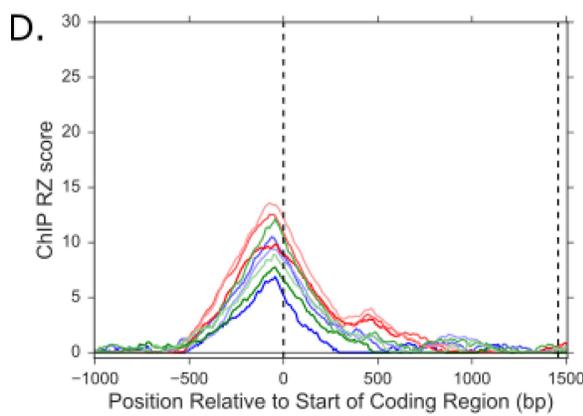
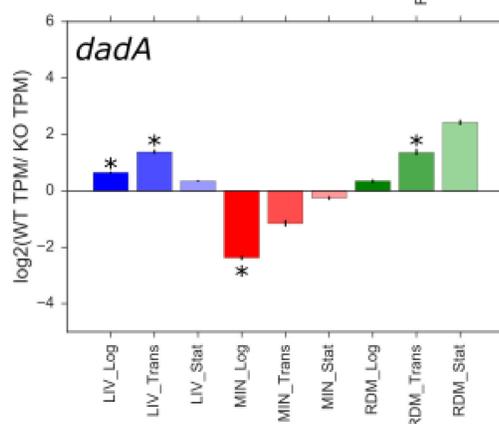
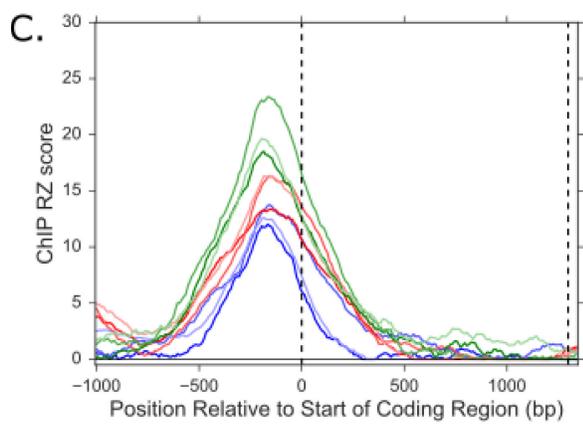
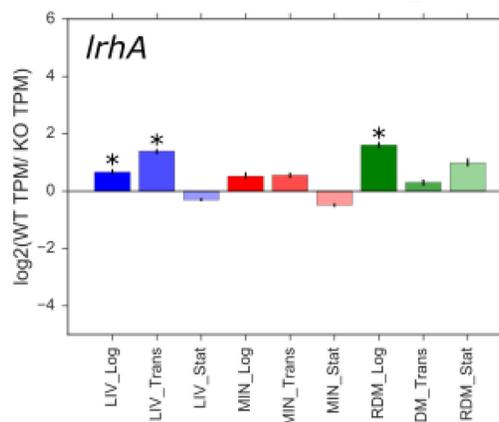
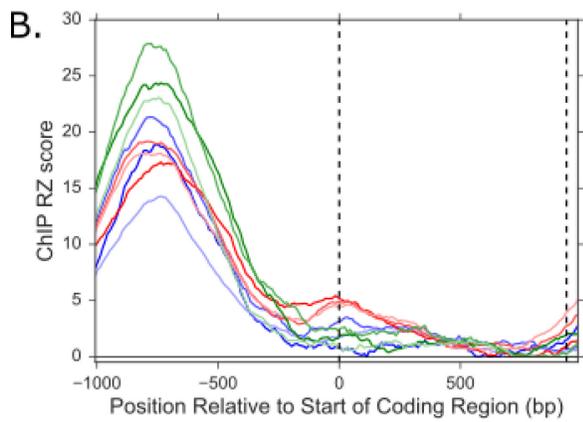
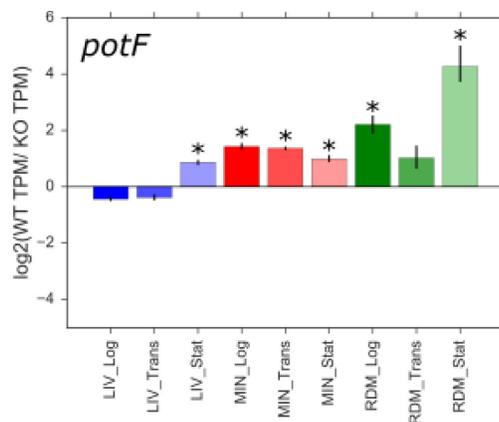
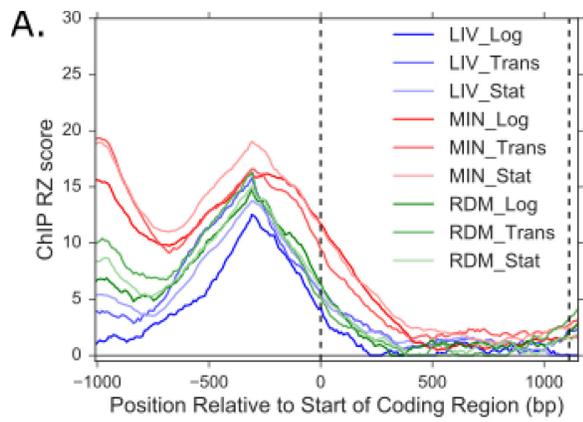
**Figure 3.5: Known targets of direct Lrp targets explain the mechanism of indirect Lrp regulation at some genes.** *A.* ChIP density and RNA-seq expression change ( $\log_2(\text{WT/KO})$ ) for direct Lrp target LrhA and its known target genes, FimE, FlhC and FlhD (Gama-Castro et al., 2016). Error bars for the RNA-seq data indicate a percentile based 95% confidence interval from 100 bootstrap replicates of TPM estimates. Stars indicate a significant difference in RNA abundance between WT and *lrp::kanR* strains (Wald Test *q*-value of  $< 0.05$  and a genotype log fold change coefficient magnitude of  $> 0.5$ ; see Methods for details). *B.* Proposed model of Lrp/LrhA mediated regulation of LrhA targets. *C.* ChIP density and RNA-seq expression change ( $\log_2(\text{WT/KO})$ ) for direct Lrp target CysB and some of its known target genes, TcyP and CysI (Gama-Castro et al., 2016), as in *A.* *D.* Proposed model of Lrp/CysB mediated regulation of CysB targets.



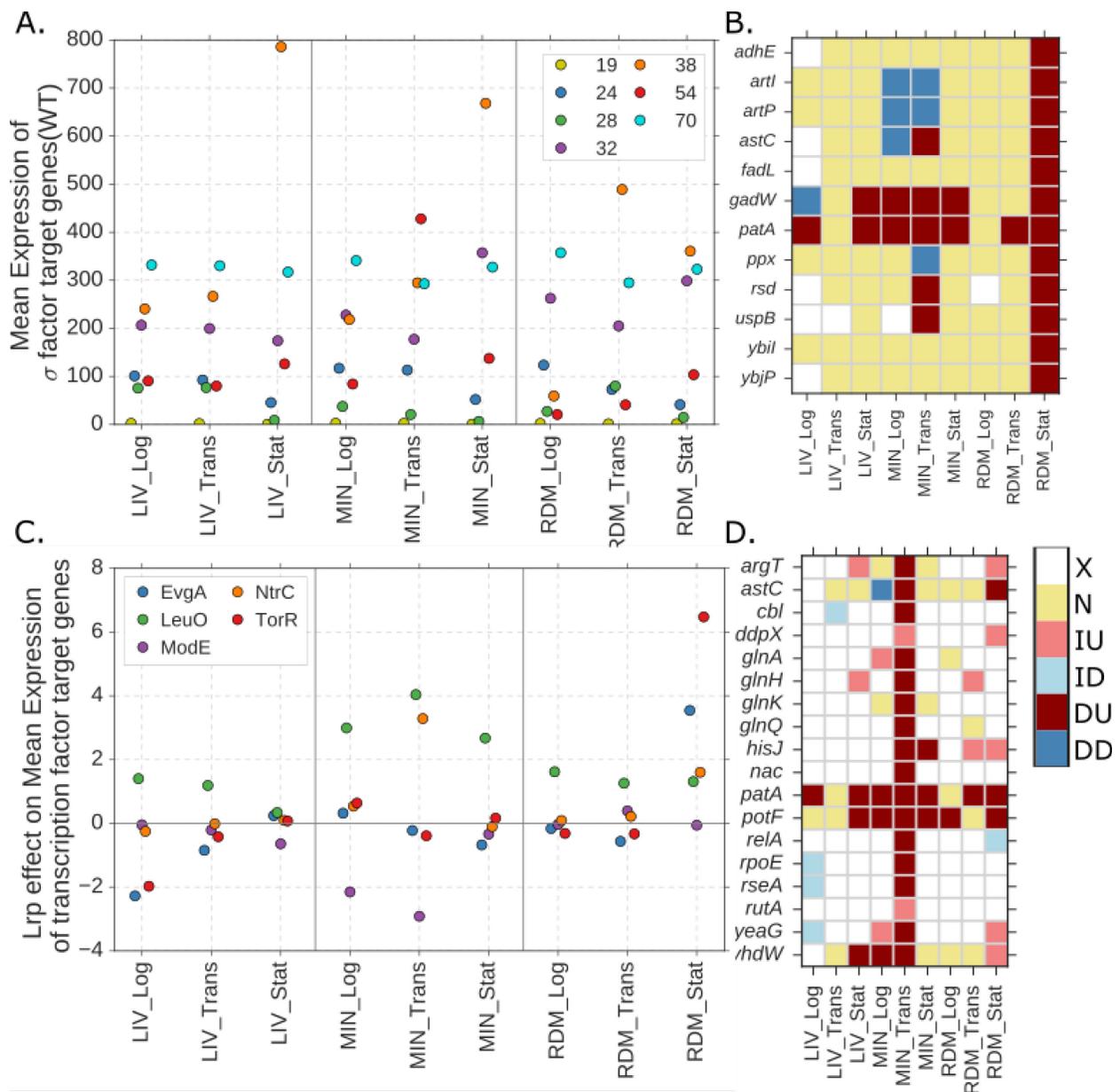
**Figure 3.6: Enriched GO-terms differ for direct and indirect Lrp targets.** *A.* A subset of GO-terms enriched or depleted within various conditions and groups of targets are listed to the left. Abbreviations are as follows: D - direct targets, I - indirect targets, N - NAP-type targets, X - no Lrp link genes. *B.* GO-terms enriched or depleted in genes with Lrp binding in at least 8 of the 9 conditions in this study. + indicates genes that meet this criteria and - indicates genes that do not meet this criteria. Boxes around a specific GO-term/condition/target group indicates a significant enrichment or depletion as indicated by a hypergeometric test ( $p\text{-value} < 0.01$ ). Color inside the box specifies the magnitude of enrichment (red) or depletion (blue) as indicated by the color bar.



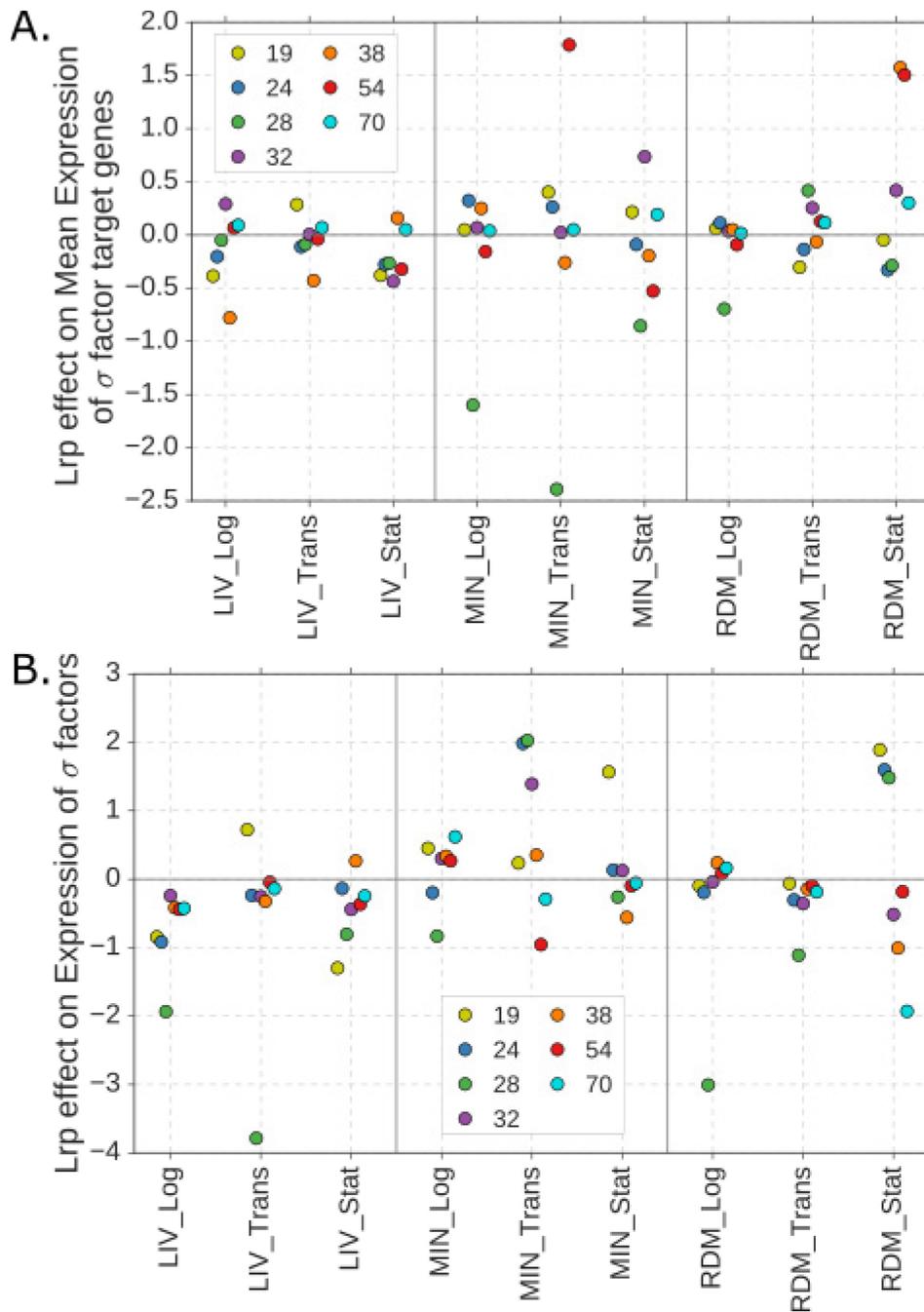
**Figure 3.7: Full GO-term enrichment results for general target classification and sub-classification by direction of *Lrp* regulatory change.** **A.** All GO-terms identified by iPAGE as having significant mutual information with our target classification within various conditions are listed to the left. Abbreviations are as follows: D - direct targets, I - indirect targets, N - NAP-type targets, X - no *Lrp* link genes. Boxes around specific GO-term/condition/target groups indicate significant enrichment or depletion (indicated by a hypergeometric test  $p$ -value  $< 0.01$ ). Color inside the box specifies the magnitude of enrichment (red) or depletion (blue) as indicated by the color bar. **B.** Similar to panel A, but downregulated and upregulated targets are treated as separate groups for target classification. Abbreviations are as follows: DD - direct downregulated targets, DU - direct upregulated targets, ID - indirect downregulated targets, IU - indirect upregulated targets, N - NAP-type targets, X - no *Lrp* link genes.



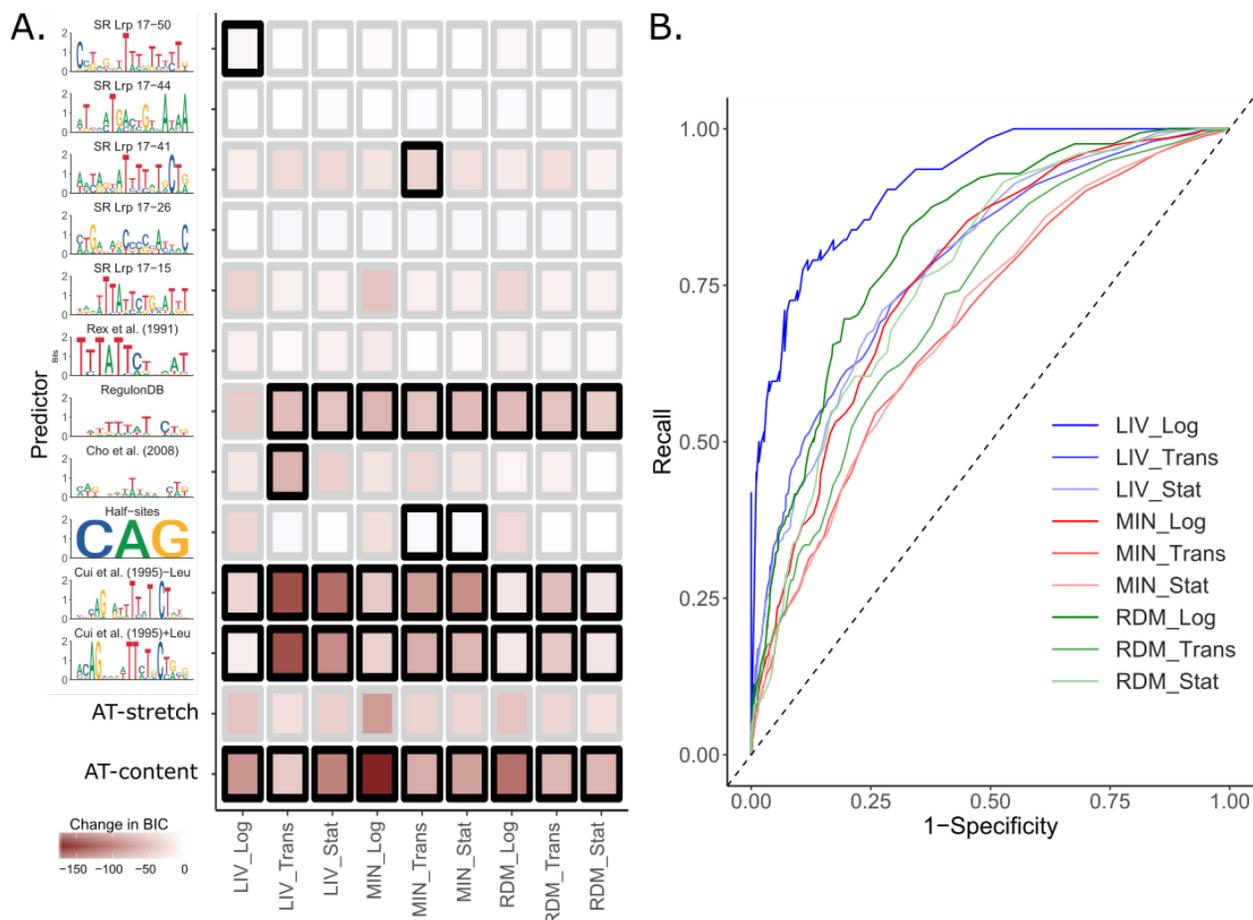
**Figure 3.8: Lrp sits at genes in poised position in preparation for regulatory activity.** ChIP robust Z-score (left) and RNA-seq expression change ( $\log_2(WT/KO)$ ; right) for four Lrp targets. *potF* (A) and *dadA* (C) are previously known targets, and *lrhA* (B) and *pepD* (D) are novel targets. Dashed vertical lines on the ChIP robust Z-score graph mark the start and end of the gene coding region. Error bars for the RNA-seq data indicate a percentile based 95% confidence interval from 100 bootstrap replicates of TPM estimates. Stars indicate a significant difference in RNA abundance between WT and *lrp::kanR* strains (Wald Test q-value of  $< 0.05$  and a genotype log fold change coefficient magnitude of  $> 0.5$ ; see Methods for details).



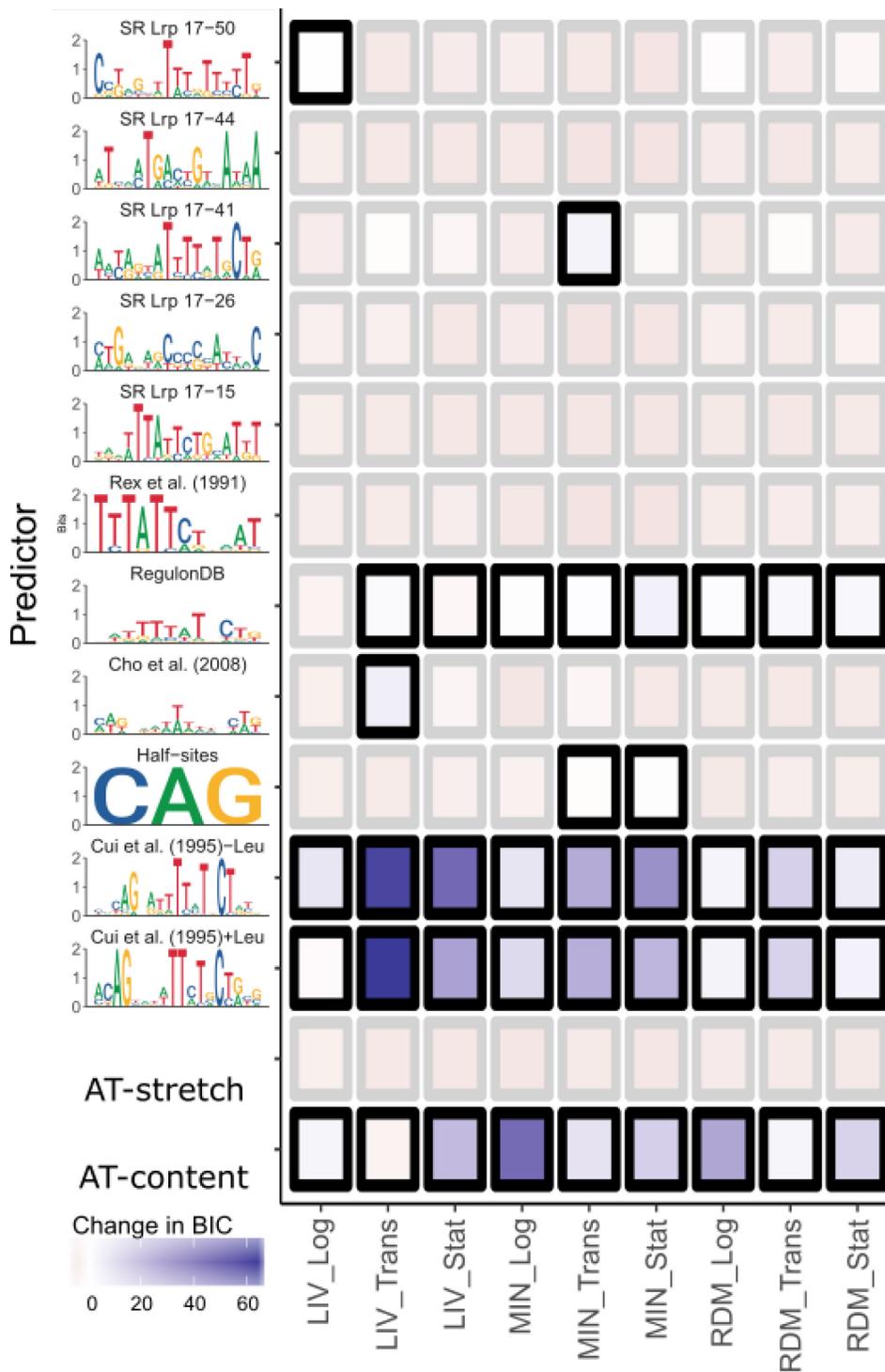
**Figure 3.9: Lrp interacts with other regulatory factors to control some targets' expression.** *A.* Average expression (TPM) of known targets of each  $\sigma$  factor in WT cells at each condition. *B.* Heatmap showing classification of a subset of  $\sigma^{38}$  targets which are direct Lrp-activated targets at RDM\_Stat. Abbreviations on the color bar are as follows: DD - direct downregulated targets, DU - direct upregulated targets, ID - indirect downregulated targets, IU - indirect upregulated targets, N - NAP-type targets, X - no Lrp link. *C.* Average  $\log_2(\text{WT}/\text{KO})$  expression ratio of known transcription factor targets for selected transcription factors at each condition. *D.* Heatmap showing classification of those NtrC targets which have an annotated transcription start site and thus are classified in our analysis. Abbreviations as for *B.*



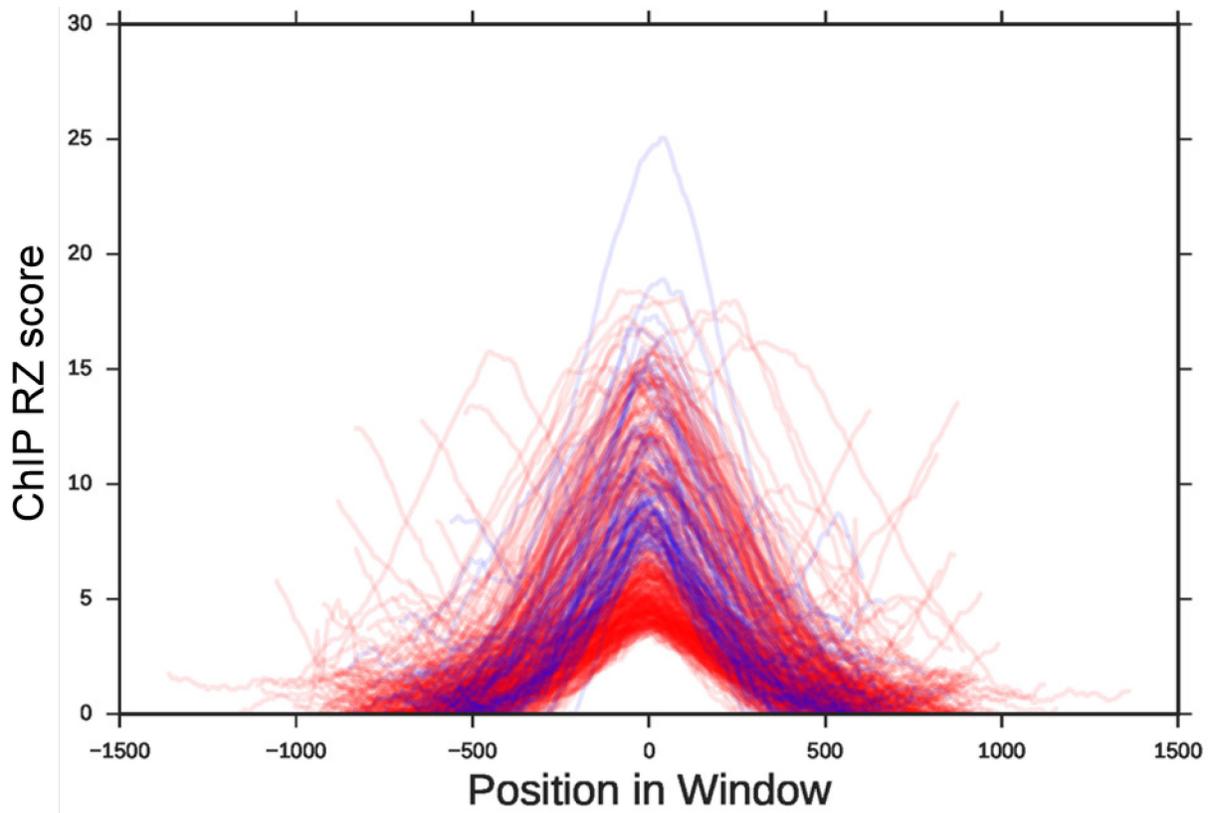
**Figure 3.10: Characteristics and regulatory activities of potential Lrp partners.** *A.* Average  $\log_2(WT/KO)$  expression ratio of known  $\sigma$  factor targets at each condition. *B.*  $\log_2(WT/KO)$  expression ratio of  $\sigma$  factors at each condition.



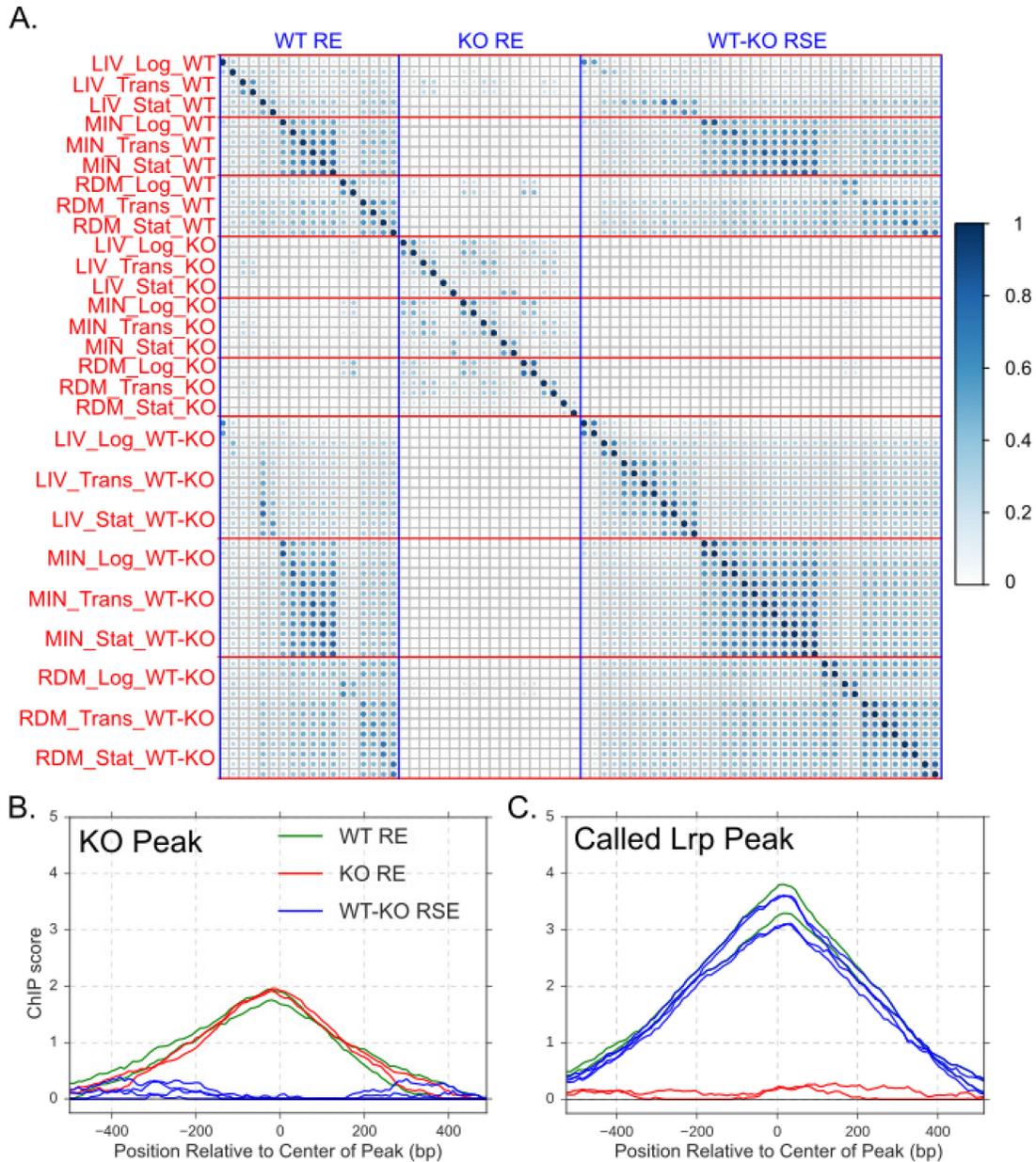
**Figure 3.11: Lrp exhibits condition-dependent sequence-preference.** **A.** Change in BIC for add-one-in logistic regression models. The y axis displays the Position Weight Matrix (PWM) used to create a particular feature. PWMs were obtained from the publication indicated above the PWM (Cho et al., 2008; Rex, Aronson, & Somerville, 1991; Yuhai Cui, 1995), RegulonDB (Gama-Castro et al., 2016) or, in the case of SR motifs, the SwissRegulon (Pachkov, Erb, Molina, & van Nimwegen, 2007). Features were created from a given PWM by dividing the count of matches within a sequence (as obtained by FIMO (Grant, Bailey, & Noble, 2011) with  $p$ -value  $< 0.0001$ ) by the length of the sequence. AT-stretch indicates the longest stretch of continuous As and Ts normalized by the length of the sequence. AT-content indicates the number of As and Ts normalized by the length of the sequence. Colors then indicate the change in BIC when a given term is added to a minimal model containing only an intercept term. Heavy boxes indicate a feature was included in the final model for that condition. For both this panel and panel **B**, the positive class of sequences was obtained by taking 500 bp around the center of each peak for each condition. The negative class of sequences was obtained by taking three times the number of equal-sized random sequences from the subset of the genome that was not in a peak for that condition. **B.** Receiver Operator Characteristic curves for each final model by condition. Curves were calculated at 0.01 increments from 0 to 1 for a predicted probability cutoff from the logistic regression. Full statistics including five-fold cross-validation are included in Table 3.8.



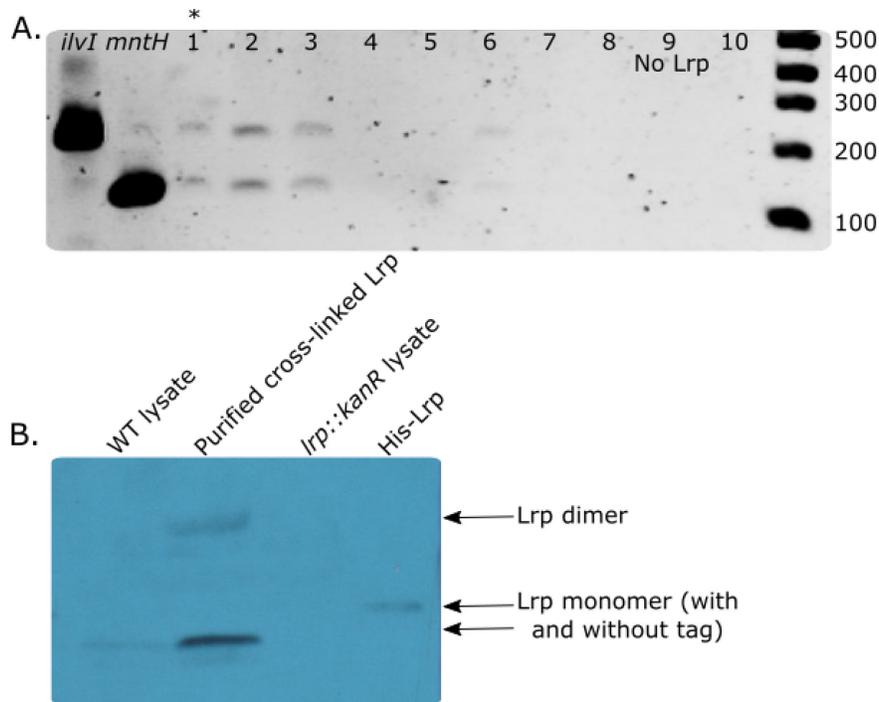
**Figure 3.12: Changes in BIC for leave-one-out logistic regression models.** Same as for Figure 3.10 except that coloring indicates the change in BIC when a particular term is dropped from the original model (containing all possible terms) under that condition. Heavy boxes indicate that a feature was included in the final model for that condition.



**Figure 3.13: Lrp peaks are a consistent length.** Overlay of peaks at MIN\_Log and LIV\_Log showing peak length consistency. All peaks are centered. Called peaks at MIN\_Log are in red, LIV\_Log in blue.



**Figure 3.14: *Lrp* ChIP-Seq data is highly reproducible.** **A.** Heatmap displaying the similarity between replicates based on shared locations in the highest 2% of signal in each replicate as quantified by the Jaccard statistic ( $\frac{A \cap B}{A \cup B}$ ). Replicates for each WT Raw Enrichment (WT RE), *lrp::kanR* Raw Enrichment (KO RE), and WT-*lrp::kanR* Raw Subtracted Enrichment (WT-KO RSE) are shown. (Details for each signal calculation in the methods). Red lines separate replicates in the same nutrient conditions, Blue lines separate replicates in the same genotype. Plot generated with the corrplot R package (Simko, 2017). **B.** Representative non-specific peak from the MIN\_Trans condition. Since the peak is seen in both the *lrp::kanR* and WT strains, it does not qualify as a *Lrp* peak in our data analysis pipeline. Green traces represent the WT Raw Enrichment (RE) from each of two replicates. Likewise, red traces indicate *lrp::kanR* RE and blue traces represent the *lrp::kanR* subtracted replicates. **C.** Representative true *Lrp* peak from the MIN\_Trans condition. Colors are the same as in **B**.



**Figure 3.15: Lrp antibody does not interfere with DNA binding and is specific for Lrp.** **A.** Pull down assay to test ability of antibody to bind DNA-bound Lrp. The first two labeled lines show the expected size of the band for pull down of the specific (*ilvI*) and non-specific (*mntH*) DNA-fragments. Lanes 1-8 are candidate anti-Lrp antibodies. Lane 9 is a negative control with no Lrp. Lane 10 is a positive control with a previously successful antibody clone (which had suffered degradation at the time of this assay). The star above lane 1 indicates that this is the antibody subclone we selected to produce. **B.** Western blot using the selected antibody subclone. Monomer Lrp bands (with some size discrepancy due to the presence of a tag) are apparent in the WT lysate and the two lanes with purified Lrp. No bands are visible in the *lrp::kanR* lysate lane.

## Chapter 4 Conclusions

### Summary of Thesis

#### *E. coli chromosome architecture is dominated by interaction propensities*

In our 3C-sequencing experiments, we identify a novel feature, each region's interaction propensity, that strongly contributes to the overall pattern of interactions we observe. The strong consistency of the interaction propensity when compared between replicates, time points and strains is striking and suggests that it plays a role in chromosome architecture. This conclusion is supported by the observed correlation between the CID centers and regions of low interaction propensity. In addition, several genome-wide features exhibit strong positive or negative correlations with the interaction propensity. These experiments are not able to distinguish causality in these correlations, but they provide directions for future studies. Taken together, the data suggest that the genome is divided into two regions: an interactive, highly expressed region, and a transcriptionally more silent, protein bound region isolated from making DNA-DNA interactions. The similarity between this organization and the euchromatin/ heterochromatin based structure in eukaryotes is intriguing and implies that fundamental principles of genome organization may be conserved across kingdoms of life.

#### *rRNA operon clustering represents an example of a secondary organizing feature*

Another feature that echoes what is known about eukaryotic genome organization is the clustering of the rRNA operons, as seen in our data and previously published work (Gaal et al., 2016). The eukaryotic nucleolus is the site of ribosome synthesis, and it appears a similar

organizational feature occurs in bacteria. The rRNA clustering is not just due to high levels of transcription, though the clustering is attenuated during later points of growth, since we do not observe clustering for groups of highly-expressed genes related either to translation, secreted proteins, or metabolic functions. Additionally, the rRNA operons are in regions of relatively low interaction propensity, indicating that the clustering happens as a result of factors other than the interaction propensity. Careful investigation of 3C-seq data is important to ensure that patterns such as this are not overlooked.

#### *Macrodomains are not universally observed in our data*

Macrodomains represent another feature that might be obscured by the population level analysis in 3C-seq. While our data strongly supports the presence of the left MD, the other MDs are not apparent, either visually or computationally. However, MD-like structures appear upon polymer modeling based on parameters in our data, suggesting that the technique used to evaluate genome organization may influence what features can be documented. This is an important factor to consider for future comparisons and studies.

#### *Regulatory patterns provide strong evidence for Lrp as a global regulator*

By interrogating nine physiological conditions with both ChIP-seq and RNA-seq, we are able to identify a larger regulon for Lrp than previously established. We observe that 35% of genes with annotated transcription start sites are directly or indirectly regulated by Lrp in at least one condition, and that up to 50% of all *E. coli* genes exhibit Lrp-dependent RNA expression in at least one condition. This scope illustrates the important role that Lrp plays in *E. coli* transcriptional regulation.

The combined binding and expression data also allows us to identify genes at which Lrp acts directly and those that are indirectly regulated by Lrp. Indirect targets represent 56.1 to 91.9% of total targets (direct and indirect) at each condition, marking the vital role of indirect regulation in the Lrp regulon. This pattern reinforces our understanding of Lrp as a global regulator, since we expect global regulators to have far-ranging indirect regulatory effects due to the hierarchical nature of bacterial regulatory systems.

The differing GO-terms enriched among direct targets as opposed to indirect targets also exemplifies the hierarchical organization. Genes encoding proteins required for nutrient transport and utilization are often direct targets, while those impacting a wider variety of cellular behavior (such as motility) appear as indirect targets. In many of the examples we investigated, indirect targets are also known to be regulated by other transcription factors besides the direct Lrp-target that we identify, leading to the possibility of combinatorial regulation depending on the exact conditions with which the cell is confronted.

#### *Lrp binding exhibits changing specificity during later stages of growth*

Many of the binding sites in our data contain sequences similar to previously identified Lrp binding motifs. Other more degenerate binding motifs often display high AT-content. We performed modeling to predict Lrp binding sites based on various sequence features, either previously-identified specific Lrp binding sites or general AT-content variables. We found that the relative importance of specific motifs increases at the Trans and Stat time points, while the AT-content is most important at the Log time points across all three media conditions. While these models do not provide complete sensitivity or specificity, they indicate an important pattern about Lrp behavior. Genes with AT-rich promoters and perhaps a weak Lrp consensus

site might be more likely to be bound and regulated (in some manner) during logarithmic growth rather than genes with GC-rich promoters and stronger Lrp consensus sites. In addition, since we see this pattern across media conditions, we must conclude that Lrp binding specificity is not solely determined by leucine since AT-content is similarly important in MIN\_Log and LIV\_Log. If leucine levels alone determined specificity, we would not expect to see AT-content being important at MIN\_Log, when leucine levels would already be low.

#### *Lrp likely interacts with a variety of coregulators*

We document a high level of Lrp binding that does not result in a regulatory change for the bound gene, dubbed NAP-type activity. Over 90% of direct targets exhibit NAP-type binding in at least one condition, suggesting that Lrp remains bound and poised to regulate at many of its target genes. The question then remains of how non-functional binding turns into binding that causes regulatory changes. For most NAP-direct gene transitions, we do not see a change in the Lrp binding profile, arguing for the interaction of Lrp with other coregulators to affect expression. From our data, we propose that Lrp likely interacts with the nitrogen-response  $\sigma$  factor,  $\sigma^{54}$  and one of its known activators NtrC, likely through modulation of the DNA structure near the promoter. It is probable that further research will identify other coregulators that interact with Lrp to affect gene expression.

#### **Future Directions**

There is still much to be learned about regulation in *E. coli* and other bacteria. Studies analyzing chromosome architecture are inherently limited by the fact that some interactions we measure may be products of intercellular interactions produced during sample preparation.

While the current 3C technique has been optimized to limit that occurrence, adjustment of the method to allow many of the preparatory steps to be performed in permeabilized cells (as in *in situ* HiC on eukaryotic systems (Ramani et al., 2016)) might decrease that problem. Comparing such data to our current experiments would allow identification of what interactions might be off-target. In addition, subtler patterns of genome organization might be apparent. For these experiments, rather than lysing the cells after cross-linking, the cells would be permeabilized using toluene so that the digestion and ligation of the DNA would occur within each cell. The resulting DNA would then be prepared for sequencing and analyzed in a similar manner. Experiments like these should provide data that is less noisy, and thus enable more accurate conclusions to be drawn about the chromosome organization in *E. coli*.

In order to investigate further the proteins or interactions that cause a certain type of genome organization, it would be interesting to perform 3C-sequencing studies in strains lacking one or more of the canonical nucleoid associated proteins. While other studies in the lab have suggested that NAP-binding is fairly redundant (Goss & Freddolino, in preparation), it would be interesting to investigate if NAP-mediated effects on global chromosome architecture are as redundant. I suspect that the type of effect the protein has on DNA (either bending or bridging) would be important. Knocking out two DNA bending proteins, for example, would likely have a more significant effect than knocking out a combination of a DNA bending and a DNA bridging protein.

There are many areas worthy of further investigation concerning Lrp as well. Lrp, or one of its homologues, has been identified in several global studies of post-translational modifications (PTM) as undergoing a variety of acetylations and phosphorylations (Baeza et al., 2014; Potel et al., 2018; Yokoyama et al., 2007). Several of the acetylation and phosphorylation

sites are in the DNA-binding domain, so this immediately suggests that modifications likely alter the electrostatics of DNA-Lrp binding and so may eliminate binding at some or all targets. This could be an interesting method of regulating Lrp activity in addition to standard alteration of gene expression levels, and is in line with the discovery of more PTM control in bacteria (Cain, Solis, & Cordwell, 2014; Hentchel & Escalante-Semerena, 2015). Changes in binding affinity could also selectively affect certain genes targets, which might aid in explaining the variant modes of Lrp regulation in response to leucine.

Therefore, investigating the effect of Lrp mutants (with either phosphomimetics or alanine mutations that cannot be acetylated or phosphorylated) on cell growth and the pattern of Lrp binding and regulatory control would be a useful area of research. First, it would be important to assess the DNA binding affinity of Lrp containing one or more serine/threonine to glutamate (S/T→E) mutations at proposed phosphorylation sites or one or more lysine to alanine (K→A) mutations. The disruption of attractive electrostatic interactions would be expected to decrease the binding affinity in both cases. Second, genomically encoded mutants could be used to assess how continued phosphorylation (the S/T→E mutations) or elimination of phosphorylation regulation (S/T→alanine or K→A mutations) affect cell growth, by monitoring the growth rate, and Lrp activity, by performing ChIP-seq and RNA-seq. If PTMs on Lrp appears to play a role in its activity, it would also be interesting to perform targeted mass-spectrometry on Lrp under different conditions and time points to determine when various PTMs are present. Acetylation is known to affect the DNA-binding activity of transcription factors in bacteria (Thao, Chen, Zhu, & Escalante-Semerena, 2010), and so I expect that PTMs would strongly affect Lrp activity, and may be a critical method of controlling when Lrp is functional.

In addition, given that amino acids besides leucine have been identified as potential coregulators (Hart & Blumenthal, 2011), identifying the array of Lrp binding sites and regulated targets upon supplementation of other amino acids, such as alanine or methionine, would be interesting. Just as we used minimal media supplemented with leucine, isoleucine and valine to assess the effect of leucine on the Lrp regulon, cells grown in minimal media supplemented with alanine or methionine could be used for ChIP-seq and RNA-seq in order to identify Lrp binding locations and Lrp-regulated genes. By again comparing the WT and Lrp KO cells, the members of the Lrp regulon that are affected by varying concentrations of alanine and methionine could be identified. A number of possibilities exist, including that amino acid coregulators may be 1) combinatorial or 2) redundant in their effect on Lrp or 3) that they may somehow induce unique structural conformations that have varying affinities for DNA sites or protein-interaction partners. Thus, genes that were bound only in low leucine conditions may also be bound only under low alanine or methionine conditions (redundant behavior), or it might be that distinct subsets of genes are bound under those conditions (unique behavior). These studies would provide valuable insight into what serves as the input to influence Lrp regulatory activity.

This study is not able to provide much information about the influence of Lrp oligomerization on gene regulatory activity. We do not observe a condition-dependent change in peak length, which might imply occupancy by an octamer or hexadecamer, but that may be due to not employing a separate digestion step as done in recent high-resolution ChIP methods (Skene & Henikoff, 2015). The ability of octamers and hexadecamers to wrap and conceal DNA from, or move and expose DNA to, other transcription or regulatory factors is a very intuitive hypothesis for why Lrp might have different effects at different sites. High-resolution ChIP studies, employing an additional DNA digestion step on cross-linked protein/DNA complexes,

would allow investigation of which targets are bound by octamers or hexadecamers and would demonstrate how skewed the binding preference is. For example, perhaps gene A is only ever bound by Lrp octamers, gene B is only bound by hexadecamers and gene C can be bound by either octamers or hexadecamers depending on the most prevalent state of Lrp.

Finally, while we identified some mechanisms of Lrp-mediated regulation, many of the reasons for positive or negative regulation remain elusive. A technique is currently being developed in our lab to pull-down a specific sequence of DNA *in vivo* and identify its protein binding partners by mass spectrometry. It would be informative to perform this experiment on several positively and negatively regulated Lrp targets in order to identify candidate coregulators. The candidate coregulators could then be tested *in vitro* for physical interaction with Lrp by isothermal titration calorimetry. If the coregulator is not essential, quantitative reverse-transcriptase PCR (qPCR) on RNA isolated from WT and candidate coregulator-knockout strains could be used to determine whether Lrp-dependent regulation is eliminated or greatly attenuated upon elimination of the candidate coregulator. If the coregulator is essential, it may be possible to mutate or eliminate its DNA binding sites at certain target genes and again assess whether Lrp-mediated regulation of the target gene is eliminated via qPCR on isolated RNA. While much of the *E. coli* regulatory network is annotated, there are still many gaps, and so biochemical identification of interaction partners would be a critical supplement to any future computational analysis of the Lrp regulon.

## Bibliography

- Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., & Ishihama, A. (1999). Growth Phase-Dependent Variation in Protein Composition of the Escherichia coli Nucleoid. *J Bacteriol*, 181(20), 6361-6370.
- Azam, T. A., & Ishihama, A. (1999). Twelve Species of the Nucleoid-associated Protein from Escherichia coli : SEQUENCE RECOGNITION SPECIFICITY AND DNA BINDING AFFINITY. *Journal of Biological Chemistry*, 274(46), 33105-33113. doi: 10.1074/jbc.274.46.33105
- Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., . . . Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2, 2006.0008-2006.0008. doi: 10.1038/msb4100050
- Baek, C.-H., Wang, S., Roland, K. L., & Curtiss, R. (2009). Leucine-Responsive Regulatory Protein (Lrp) Acts as a Virulence Repressor in Salmonella enterica Serovar Typhimurium. *J Bacteriol*, 191(4), 1278-1292. doi: 10.1128/JB.01142-08
- Baeza, J., Dowell, J. A., Smallegan, M. J., Fan, J., Amador-Noguez, D., Khan, Z., & Denu, J. M. (2014). Stoichiometry of site-specific lysine acetylation in an entire proteome. *J Biol Chem*, 289(31), 21326-21338. doi: 10.1074/jbc.M114.581843
- Battesti, A., Majdalani, N., & Gottesman, S. (2011). The RpoS-mediated general stress response in Escherichia coli. *Annu Rev Microbiol*, 65, 189-213. doi: 10.1146/annurev-micro-090110-102946
- Berthiaume, F., Crost, C., Labrie, V., Martin, C., Newman, E. B., & Harel, J. (2004). Influence of l-Leucine and l-Alanine on Lrp Regulation of foo, Coding for F165(1), a Pap Homologue. *J Bacteriol*, 186(24), 8537-8541. doi: 10.1128/JB.186.24.8537-8541.2004
- Blumer, C., Kleefeld, A., Lehnen, D., Heintz, M., Dobrindt, U., Nagy, G., . . . Uden, G. (2005). Regulation of type 1 fimbriae synthesis and biofilm formation by the transcriptional regulator LrhA of Escherichia coli. *Microbiology*, 151(10), 3287-3298. doi: doi:10.1099/mic.0.28098-0
- Bouvier, J., Gordia, S., Kampmann, G., Lange, R., Hengge-Aronis, R., & Gutierrez, C. (1998). Interplay between global regulators of Escherichia coli : effect of RpoS, Lrp and H-NS on transcription of the gene osmC. *Molecular Microbiology*, 28(5), 971-980. doi: 10.1046/j.1365-2958.1998.00855.x

- Brinkman, A. B., Ettema, T. J. G., De Vos, W. M., & Van Der Oost, J. (2003). The Lrp family of transcriptional regulators. *Molecular Microbiology*, *48*(2), 287-294. doi: 10.1046/j.1365-2958.2003.03442.x
- Browning, D. F., Grainger, D. C., & Busby, S. J. W. (2010). Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Current Opinion in Microbiology*, *13*(6), 773-780. doi: <https://doi.org/10.1016/j.mib.2010.09.013>
- Bryant, J. A., Sellars, L. E., Busby, S. J., & Lee, D. J. (2014). Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Res*, *42*(18), 11383-11392. doi: 10.1093/nar/gku828
- Cabrera Julio, E., & Jin Ding, J. (2003). The distribution of RNA polymerase in Escherichia coli is dynamic and sensitive to environmental cues. *Molecular Microbiology*, *50*(5), 1493-1505. doi: 10.1046/j.1365-2958.2003.03805.x
- Cagliero, C., Grand, R. S., Jones, M. B., Jin, D. J., & O'Sullivan, J. M. (2013). Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. *Nucleic Acids Research*, *41*(12), 6058-6071. doi: 10.1093/nar/gkt325
- Cain, J. A., Solis, N., & Cordwell, S. J. (2014). Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J Proteomics*, *97*, 265-286. doi: 10.1016/j.jprot.2013.08.012
- Caldara, M., Charlier, D., & Cunin, R. (2006). The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. *Microbiology*, *152*(Pt 11), 3343-3354. doi: 10.1099/mic.0.29088-0
- Calvo, J. M., & Matthews, R. G. (1994). The leucine-responsive regulatory protein, a global regulator of metabolism in Escherichia coli. *Microbiological Reviews*, *58*(3), 466-490.
- Casadesús, J., & Low, D. A. (2013). Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria. *J Biol Chem*, *288*(20), 13929-13935. doi: 10.1074/jbc.R113.472274
- Chen, S., & Calvo, J. M. (2002). Leucine-induced Dissociation of Escherichia coli Lrp Hexadecamers to Octamers. *Journal of Molecular Biology*, *318*(4), 1031-1042. doi: 10.1016/s0022-2836(02)00187-0
- Chen, S., Hao, Z., Bieniek, E., & Calvo, J. M. (2001). Modulation of Lrp action in Escherichia coli by leucine: effects on non-specific binding of Lrp to DNA1. *Journal of Molecular Biology*, *314*(5), 1067-1075. doi: <http://dx.doi.org/10.1006/jmbi.2000.5209>

- Chen, S., Iannolo, M., & Calvo, J. M. (2005). Cooperative binding of the leucine-responsive regulatory protein (Lrp) to DNA. *J Mol Biol*, 345(2), 251-264. doi: 10.1016/j.jmb.2004.10.047
- Chen, S., Rosner, M. H., & Calvo, J. M. (2001). Leucine-regulated self-association of leucine-responsive regulatory protein (Lrp) from *Escherichia coli*. *Journal of Molecular Biology*, 312(4), 625-635. doi: <http://dx.doi.org/10.1006/jmbi.2001.4955>
- Cho, B. K., Barrett, C. L., Knight, E. M., Park, Y. S., & Palsson, B. O. (2008). Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 105(49), 19462-19467. doi: 10.1073/pnas.0807227105
- Cohen, N. R., Ross, C. A., Jain, S., Shapiro, R. S., Gutierrez, A., Belenky, P., . . . Collins, J. J. (2016). A role for the bacterial GATC methylome in antibiotic stress survival. *Nat Genet*, 48(5), 581-586. doi: 10.1038/ng.3530
- Colland, F., Barth, M., Hengge-Aronis, R., & Kolb, A. (2000).  $\sigma$  factor selectivity of *Escherichia coli* RNA polymerase: role for CRP, IHF and Lrp transcription factors. *The EMBO Journal*, 19(12), 3028-3037. doi: 10.1093/emboj/19.12.3028
- Cui, Y., Midkiff, M. A., Wang, Q., & Calvo, J. M. (1996). The Leucine-responsive Regulatory Protein (Lrp) from *Escherichia coli*: STOICHIOMETRY AND MINIMAL REQUIREMENTS FOR BINDING TO DNA. *Journal of Biological Chemistry*, 271(12), 6611-6617. doi: 10.1074/jbc.271.12.6611
- Dame, R. T. (2005). The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular Microbiology*, 56(4), 858-870. doi: 10.1111/j.1365-2958.2005.04598.x
- Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A*, 97(12), 6640-6645.
- Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 14(6), 390-403. doi: 10.1038/nrg3454
- Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558), 1306-1311. doi: 10.1126/science.1067799
- Doyle, B., Fudenberg, G., Imakaev, M., & Mirny, L. A. (2014). Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. *PLoS Computational Biology*, 10(10), e1003867. doi: 10.1371/journal.pcbi.1003867
- E B Newman, a., & Lin, R. (1995). Leucine-Responsive Regulatory Protein: A Global Regulator of Gene Expression in *E. Coli*. *Annual Review of Microbiology*, 49(1), 747-775. doi: doi:10.1146/annurev.mi.49.100195.003531

- Eisen, J. A., Heidelberg, J. F., White, O., & Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology*, *1*(6), research0011.0011-0011.0019.
- Endesfelder, U., Finan, K., Holden, Seamus J., Cook, Peter R., Kapanidis, Achillefs N., & Heilemann, M. Multiscale Spatial Organization of RNA Polymerase in *Escherichia coli*. *Biophysical Journal*, *105*(1), 172-181. doi: 10.1016/j.bpj.2013.05.048
- Engstrom, M. D., & Mobley, H. L. T. (2016). Regulation of Expression of Uropathogenic *Escherichia coli* Nonfimbrial Adhesin TosA by PapB Homolog TosR in Conjunction with H-NS and Lrp. *Infection and Immunity*, *84*(3), 811-821. doi: 10.1128/iai.01302-15
- Espeli, O., Mercier, R., & Boccard, F. (2008). DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Molecular Microbiology*, *68*(6), 1418-1427. doi: 10.1111/j.1365-2958.2008.06239.x
- Ettema, T. J. G., Brinkman, A. B., Tani, T. H., Rafferty, J. B., & van der Oost, J. (2002). A Novel Ligand-binding Domain Involved in Regulation of Amino Acid Metabolism in Prokaryotes. *Journal of Biological Chemistry*, *277*(40), 37464-37468.
- Franklin, F. C., & Venables, W. A. (1976). Biochemical, genetic, and regulatory studies of alanine catabolism in *Escherichia coli* K12. *Mol Gen Genet*, *149*(2), 229-237.
- Freddolino, P. L., Amini, S., & Tavazoie, S. (2012). Newly Identified Genetic Variations in Common *Escherichia coli* MG1655 Stock Cultures. *J Bacteriol*, *194*(2), 303-306. doi: 10.1128/JB.06087-11
- Friedberg, D., Midkiff, M., & Calvo, J. M. (2001). Global versus local regulatory roles for Lrp-related proteins: *Haemophilus influenzae* as a case study. *J Bacteriol*, *183*(13), 4004-4011. doi: 10.1128/JB.183.13.4004-4011.2001
- Gaal, T., Bratton, B. P., Sanchez-Vazquez, P., Sliwicki, A., Sliwicki, K., Vogel, A., . . . Gourse, R. L. (2016). Colocalization of distant chromosomal loci in space in *E. coli*: a bacterial nucleolus. *Genes & Development*, *30*(20), 2272-2285. doi: 10.1101/gad.290312.116
- Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res*, *44*(D1), D133-143. doi: 10.1093/nar/gkv1156
- Gaston, K., & Jayaraman, P. S. (2003). Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cellular and Molecular Life Sciences CMLS*, *60*(4), 721-741. doi: 10.1007/s00018-003-2260-3

- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., . . . Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, *489*, 91. doi: 10.1038/nature11245
- Goda, S. K., & Minton, N. P. (1995). A simple procedure for gel electrophoresis and northern blotting of RNA. *Nucleic Acids Res*, *23*(16), 3357-3358.
- Goodarzi, H., Elemento, O., & Tavazoie, S. (2009). Revealing global regulatory perturbations across human cancers. *Molecular Cell*, *36*(5), 900-911. doi: 10.1016/j.molcel.2009.11.016
- Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017-1018. doi: 10.1093/bioinformatics/btr064
- Graunke, D. M., Fornace, A. J., & Pieper, R. O. (1999). Presetting of chromatin structure and transcription factor binding poise the human GADD45 gene for rapid transcriptional up-regulation. *Nucleic Acids Research*, *27*(19), 3881-3890.
- Green, M. R. (2005). Eukaryotic Transcription Activation: Right on Target. *Molecular Cell*, *18*(4), 399-402. doi: <https://doi.org/10.1016/j.molcel.2005.04.017>
- Guimaraes, J. C., Rocha, M., & Arkin, A. P. (2014). Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Research*, *42*(8), 4791-4799. doi: 10.1093/nar/gku126
- Hart, B. R., & Blumenthal, R. M. (2011). Unexpected coregulator range for the global regulator Lrp of Escherichia coli and Proteus mirabilis. *J Bacteriol*, *193*(5), 1054-1064. doi: 10.1128/JB.01183-10
- Hart, B. R., Mishra, P. K., Lintner, R. E., Hinerman, J. M., Herr, A. B., & Blumenthal, R. M. (2011). Recognition of DNA by the helix-turn-helix global regulatory protein Lrp is modulated by the amino terminus. *J Bacteriol*, *193*(15), 3794-3803. doi: 10.1128/JB.00191-11
- Hentchel, K. L., & Escalante-Semerena, J. C. (2015). Acylation of Biomolecules in Prokaryotes: a Widespread Strategy for the Control of Biological Function and Metabolic Stress. *Microbiology and Molecular Biology Reviews*, *79*(3), 321-346. doi: 10.1128/membr.00020-15
- Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med*, *9*(7), 811-818.
- Holmqvist, E., Unoson, C., Reimegård, J., & Wagner, E. G. H. (2012). A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp. *Molecular Microbiology*, *84*(3), 414-427. doi: 10.1111/j.1365-2958.2012.07994.x

- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), 90-95.
- Ishihama, A., Shimada, T., & Yamazaki, Y. (2016). Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res*, 44(5), 2058-2074. doi: 10.1093/nar/gkw051
- Kahramanoglou, C., Seshasayee, A. S., Prieto, A. I., Ibberson, D., Schmidt, S., Zimmermann, J., . . . Luscombe, N. M. (2011). Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res*, 39(6), 2073-2091. doi: 10.1093/nar/gkq934
- Kawashima, T., Aramaki, H., Oyamada, T., Makino, K., Yamada, M., Okamura, H., . . . Suzuki, M. (2008). Transcription Regulation by Feast/Famine Regulatory Proteins, FFRPs, in Archaea and Eubacteria. *Biological and Pharmaceutical Bulletin*, 31(2), 173-186. doi: 10.1248/bpb.31.173
- Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., . . . Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, 45(D1), D543-D550. doi: 10.1093/nar/gkw1003
- Kiupakis, A. K., & Reitzer, L. (2002). ArgR-independent induction and ArgR-dependent superinduction of the astCADBE operon in Escherichia coli. *J Bacteriol*, 184(11), 2940-2950.
- Klepsch, M. M., Kovermann, M., Löw, C., Balbach, J., Permentier, H. P., Fusetti, F., . . . Berntsson, R. P. A. (2011). Escherichia coli Peptide Binding Protein OppA Has a Preference for Positively Charged Peptides. *Journal of Molecular Biology*, 414(1), 75-85. doi: <https://doi.org/10.1016/j.jmb.2011.09.043>
- Kroner, G. M., Wolfe, M. B., & Freddolino, P. (2018). Escherichia coli Lrp regulates one-third of the genome via direct, cooperative, and indirect routes. *bioRxiv*.
- Lal, A., Dhar, A., Trostel, A., Kouzine, F., Seshasayee, A. S., & Adhya, S. (2016). Genome scale patterns of supercoiling in a bacterial chromosome. *Nat Commun*, 7, 11055. doi: 10.1038/ncomms11055
- Landgraf, J. R., Wu, J., & Calvo, J. M. (1996). Effects of nutrition and growth rate on Lrp levels in Escherichia coli. *J Bacteriol*, 178(23), 6930-6936.
- Le, T. B. K., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. *Science (New York, N.Y.)*, 342(6159), 731-734. doi: 10.1126/science.1242059

- Li Q, B. J., Huang H, Bickel PJ. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), 289-293. doi: 10.1126/science.1181369
- Lin, R., D'Ari, R., & Newman, E. B. (1992). Lambda placMu insertions in genes of the leucine regulon: extension of the regulon to genes not regulated by leucine. *J Bacteriol*, 174(6), 1948-1955.
- Lin, W., Kovacicova, G., & Skorupski, K. (2007). The quorum sensing regulator HapR downregulates the expression of the virulence gene transcription factor AphA in *Vibrio cholerae* by antagonizing Lrp- and VpsR-mediated activation. *Molecular Microbiology*, 64(4), 953-967. doi: 10.1111/j.1365-2958.2007.05693.x
- Lintner, R. E., Mishra, P. K., Srivastava, P., Martinez-Vaz, B. M., Khodursky, A. B., & Blumenthal, R. M. (2008). Limited functional conservation of a global regulator among related bacterial genera: Lrp in *Escherichia*, *Proteus* and *Vibrio*. *BMC Microbiol*, 8, 60. doi: 10.1186/1471-2180-8-60
- Lioy, V. S., Cournac, A., Marbouty, M., Duigou, S., Mozziconacci, J., Espeli, O., . . . Koszul, R. (2018). Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell*, 172(4), 771-783.e718. doi: 10.1016/j.cell.2017.12.027
- Lu, P., Vogel, C., Wang, R., Yao, X., & Marcotte, E. M. (2006). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25, 117. doi: 10.1038/nbt1270
- Luijsterburg, M. S., Noom, M. C., Wuite, G. J. L., & Dame, R. T. (2006). The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: A molecular perspective. *Journal of Structural Biology*, 156(2), 262-272. doi: <http://dx.doi.org/10.1016/j.jsb.2006.05.006>
- Ma, H.-W., Buer, J., & Zeng, A.-P. (2004). Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, 5, 199-199. doi: 10.1186/1471-2105-5-199
- Marbouty, M., Le Gall, A., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J. B., . . . Nollmann, M. (2015). Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol Cell*, 59(4), 588-602. doi: 10.1016/j.molcel.2015.07.020

- Martínez-Antonio, A., & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6(5), 482-489. doi: <http://dx.doi.org/10.1016/j.mib.2003.09.002>
- Mathew, E., Zhi, J., & Freundlich, M. (1996). Lrp is a direct repressor of the dad operon in *Escherichia coli*. *J Bacteriol*, 178(24), 7234-7240.
- Mercier, R., Petit, M. A., Schbath, S., Robin, S., El Karoui, M., Boccard, F., & Espeli, O. (2008). The MatP/matS site-specific system organizes the terminus region of the *E. coli* chromosome into a macrodomain. *Cell*, 135(3), 475-485. doi: 10.1016/j.cell.2008.08.031
- Neidhardt, F. C., Bloch, P. L., & Smith, D. F. (1974). Culture Medium for Enterobacteria. *J Bacteriol*, 119(3), 736-747.
- Newman, E. B., D'Ari, R., & Lin, R. T. (1992). The leucine-Lrp regulon in *E. coli*: A global response in search of a raison d'Être. *Cell*, 68(4), 617-619. doi: [http://dx.doi.org/10.1016/0092-8674\(92\)90135-Y](http://dx.doi.org/10.1016/0092-8674(92)90135-Y)
- Nielsen, H. J., Ottesen, J. R., Youngren, B., Austin, S. J., & Hansen, F. G. (2006). The *Escherichia coli* chromosome is organized with the left and right chromosome arms in separate cell halves. *Molecular Microbiology*, 62(2), 331-338. doi: 10.1111/j.1365-2958.2006.05346.x
- Oshima, T., Ito, K., Kabayama, H., & Nakamura, Y. (1995). Regulation of lrp gene expression by H-NS and Lrp proteins in *Escherichia coli*: dominant negative mutations in lrp. *Mol Gen Genet*, 247(5), 521-528.
- Pachkov, M., Erb, I., Molina, N., & van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue), D127-131. doi: 10.1093/nar/gkl857
- Papantonis, A., & Cook, P. R. (2013). Transcription Factories: Genome Organization and Gene Regulation. *Chemical reviews*, 113(11), 8683-8705. doi: 10.1021/cr300513p
- Parti, R. P. S., Shrivastava, R., Srivastava, S., Subramanian, A. R., Roy, R., Srivastava, B. S., & Srivastava, R. (2008). A transposon insertion mutant of *Mycobacterium fortuitum* attenuated in virulence and persistence in a murine infection model that is complemented by Rv3291c of *Mycobacterium tuberculosis*. *Microbial Pathogenesis*, 45(5-6), 370-376. doi: <http://dx.doi.org/10.1016/j.micpath.2008.08.008>
- Perona, S. d. I. R. a. J. J. (2007). Structure of the *Escherichia coli* leucine-responsive regulatory protein Lrp reveals a novel octameric assembly. *Journal of Molecular Biology*, 366(5), 1589-1602.

- Peterson, S. N., Dahlquist, F. W., & Reich, N. O. (2007). The role of high affinity non-specific DNA binding by Lrp in transcriptional regulation and DNA organization. *J Mol Biol*, 369(5), 1307-1317. doi: 10.1016/j.jmb.2007.04.023
- Peterson, S. N., & Reich, N. O. (2008). Competitive Lrp and Dam Assembly at the pap Regulatory Region: Implications for Mechanisms of Epigenetic Regulation. *Journal of Molecular Biology*, 383(1), 92-105. doi: <http://dx.doi.org/10.1016/j.jmb.2008.07.086>
- Platko, J. V., & Calvo, J. M. (1993). Mutations affecting the ability of Escherichia coli Lrp to bind DNA, activate transcription, or respond to leucine. *J Bacteriol*, 175(4), 1110-1117.
- Platko, J. V., Willins, D. A., & Calvo, J. M. (1990). The ilvIH operon of Escherichia coli is positively regulated. *J Bacteriol*, 172(8), 4563-4570.
- Polaczek, P., Kwan, K., & Campbell, J. L. (1998). GATC motifs may alter the conformation of DNA depending on sequence context and N6-adenine methylation status: possible implications for DNA-protein recognition. *Mol Gen Genet*, 258(5), 488-493.
- Potel, C. M., Lin, M.-H., Heck, A. J. R., & Lemeer, S. (2018). Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. *Nature Methods*, 15, 187. doi: 10.1038/nmeth.4580
- Prieto, A. I., Kahramanoglou, C., Ali, R. M., Fraser, G. M., Seshasayee, A. S., & Luscombe, N. M. (2012). Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res*, 40(8), 3524-3537. doi: 10.1093/nar/gkr1236
- Pul, Ü., Wurm, R., Lux, B., Meltzer, M., Menzel, A., & Wagner, R. (2005). LRP and H-NS – cooperative partners for transcription regulation at Escherichia coli rRNA promoters. *Molecular Microbiology*, 58(3), 864-876. doi: 10.1111/j.1365-2958.2005.04873.x
- Pul, U., Wurm, R., & Wagner, R. (2007). The role of LRP and H-NS in transcription regulation: involvement of synergism, allostery and macromolecular crowding. *J Mol Biol*, 366(3), 900-915. doi: 10.1016/j.jmb.2006.11.067
- Ramani, V., Cusanovich, D. A., Hause, R. J., Ma, W., Qiu, R., Deng, X., . . . Duan, Z. (2016). Mapping three-dimensional genome architecture through in situ DNase Hi-C. *Nature protocols*, 11(11), 2104-2121. doi: 10.1038/nprot.2016.126
- Reitzer, L. (2003). Nitrogen assimilation and global regulation in Escherichia coli. *Annu Rev Microbiol*, 57, 155-176. doi: 10.1146/annurev.micro.57.030502.090820
- Reitzer, L., & Schneider, B. L. (2001). Metabolic Context and Possible Physiological Themes of  $\sigma(54)$ -Dependent Genes in Escherichia coli. *Microbiology and Molecular Biology Reviews*, 65(3), 422-444. doi: 10.1128/MMBR.65.3.422-444.2001

- Rex, J. H., Aronson, B. D., & Somerville, R. L. (1991). The *tdh* and *serA* operons of *Escherichia coli*: mutational analysis of the regulatory elements of leucine-responsive genes. *J Bacteriol*, *173*(19), 5944-5953.
- Roesch, P. L., & Blomfield, I. C. (1998). Leucine alters the interaction of the leucine-responsive regulatory protein (Lrp) with the *fim* switch to stimulate site-specific recombination in *Escherichia coli*. *Molecular Microbiology*, *27*(4), 751-761. doi: 10.1046/j.1365-2958.1998.00720.x
- Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., . . . Heinemann, M. (2015). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature biotechnology*, *34*, 104. doi: 10.1038/nbt.3418
- Schroeder, U., Henrich, B., Fink, J., & Plapp, R. (1994). Peptidase D of *Escherichia coli* K-12, a metallopeptidase of low substrate specificity. *FEMS Microbiol Lett*, *123*(1-2), 153-159.
- Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with python*. Paper presented at the Proceedings of the 9th Python in Science Conference.
- Shimada, T., Saito, N., Maeda, M., Tanaka, K., & Ishihama, A. (2015). Expanded roles of leucine-responsive regulatory protein in transcription regulation of the *Escherichia coli* genome: Genomic SELEX screening of the regulation targets. *Microbial Genomics*, *1*(1). doi: 10.1099/mgen.0.000001
- Shingler, V. (1996). Signal sensing by  $\sigma_{54}$ -dependent regulators: derepression as a control mechanism. *Molecular Microbiology*, *19*(3), 409-416. doi: 10.1046/j.1365-2958.1996.388920.x
- Simko, T. W. a. V. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Retrieved from <https://github.com/taiyun/corrplot>
- Skene, P. J., & Henikoff, S. (2015). A simple method for generating high-resolution maps of genome-wide protein binding. *eLife*, *4*, e09225. doi: 10.7554/eLife.09225
- Sutton, V. R., Mettert, E. L., Beinert, H., & Kiley, P. J. (2004). Kinetic Analysis of the Oxidative Conversion of the [4Fe-4S]<sup>2+</sup> Cluster of FNR to a [2Fe-2S]<sup>2+</sup> Cluster. *J Bacteriol*, *186*(23), 8018-8025. doi: 10.1128/jb.186.23.8018-8025.2004
- Tani, T. H., Khodursky, A., Blumenthal, R. M., Brown, P. O., & Matthews, R. G. (2002). Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis. *Proc Natl Acad Sci U S A*, *99*(21), 13471-13476. doi: 10.1073/pnas.212510999
- Thao, S., Chen, C.-S., Zhu, H., & Escalante-Semerena, J. C. (2010). N( $\epsilon$ )-Lysine Acetylation of a Bacterial Transcription Factor Inhibits Its DNA-Binding Activity. *PLoS ONE*, *5*(12), e15123. doi: 10.1371/journal.pone.0015123

- Traxler, M. F., Zacharia, V. M., Marquardt, S., Summers, S. M., Nguyen, H.-T., Stark, S. E., & Conway, T. (2011). Discretely calibrated regulatory loops controlled by ppGpp partition gene induction across the 'feast to famine' gradient in *Escherichia coli*. *Molecular Microbiology*, *79*(4), 830-845. doi: 10.1111/j.1365-2958.2010.07498.x
- Umbarger, M. A., Toro, E., Wright, M. A., Porreca, G. J., Baù, D., Hong, S.-H., . . . Church, G. M. (2011). The Three-Dimensional Architecture of a Bacterial Genome. *Molecular Cell*, *44*(2), 10.1016/j.molcel.2011.1009.1010. doi: 10.1016/j.molcel.2011.09.010
- Val, M.-E., Marbouty, M., de Lemos Martins, F., Kennedy, S. P., Kemble, H., Bland, M. J., . . . Mazel, D. (2016). A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Science Advances*, *2*(4), e1501914. doi: 10.1126/sciadv.1501914
- Valens, M., Penaud, S., Rossignol, M., Cornet, F., & Boccard, F. (2004). Macrodomain organization of the *Escherichia coli* chromosome. *The EMBO Journal*, *23*(21), 4330-4341. doi: 10.1038/sj.emboj.7600434
- Valens, M., Thiel, A., & Boccard, F. (2016). The MaoP/maoS Site-Specific System Organizes the Ori Region of the *E. coli* Chromosome into a Macrodomain. *PLoS Genetics*, *12*(9), e1006309. doi: 10.1371/journal.pgen.1006309
- Vassilyev, D. G., Tomitori, H., Kashiwagi, K., Morikawa, K., & Igarashi, K. (1998). Crystal Structure and Mutational Analysis of the *Escherichia coli* Putrescine Receptor: STRUCTURAL BASIS FOR SUBSTRATE SPECIFICITY. *Journal of Biological Chemistry*, *273*(28), 17604-17609. doi: 10.1074/jbc.273.28.17604
- Wang, D., Calla, B., Vimolmangkang, S., Wu, X., Korban, S. S., Huber, S. C., . . . Zhao, Y. (2011). The Orphan Gene ybjN Conveys Pleiotropic Effects on Multicellular Behavior and Survival of *Escherichia coli*. *PLoS ONE*, *6*(9), e25293. doi: 10.1371/journal.pone.0025293
- Wang, Q., & Calvo, J. M. (1993). Lrp, a major regulatory protein in *Escherichia coli*, bends DNA and can organize the assembly of a higher-order nucleoprotein structure. *The EMBO Journal*, *12*(6), 2495-2501.
- Wang, Q., Wu, J., Friedberg, D., Plakto, J., & Calvo, J. M. (1994). Regulation of the *Escherichia coli* lrp gene. *J Bacteriol*, *176*(7), 1831-1839.
- Wang, X., Le, T. B. K., Lajoie, B. R., Dekker, J., Laub, M. T., & Rudner, D. Z. (2015). Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes & Development*, *29*(15), 1661-1675. doi: 10.1101/gad.265876.115
- Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*: Springer.

- Willins, D. A., & Calvo, J. M. (1992). In vitro transcription from the Escherichia coli ilvIH promoter. *J Bacteriol*, *174*(23), 7648-7655.
- Willins, D. A., Ryan, C. W., Platko, J. V., & Calvo, J. M. (1991). Characterization of Lrp, and Escherichia coli regulatory protein that mediates a global response to leucine. *Journal of Biological Chemistry*, *266*(17), 10768-10774.
- Wu, F., Japaridze, A., Zheng, X., Kerssemakers, J. W. J., & Dekker, C. (2018). Direct Imaging of the circular chromosome of a live bacterium. *bioRxiv*.
- Xiao, G., White, D., & Bargonetti, J. (1998). p53 binds to a constitutively nucleosome free region of the mdm2 gene. *Oncogene*, *16*, 1171. doi: 10.1038/sj.onc.1201631
- Xie, T., Fu, L.-Y., Yang, Q.-Y., Xiong, H., Xu, H., Ma, B.-G., & Zhang, H.-Y. (2015). Spatial features for Escherichia coli genome organization. *BMC Genomics*, *16*(1), 37. doi: 10.1186/s12864-015-1258-1
- Yokoyama, K., Ishijima, S. A., Koike, H., Kurihara, C., Shimowasa, A., Kabasawa, M., . . . Suzuki, M. (2007). Feast/Famine Regulation by Transcription Factor FL11 for the Survival of the Hyperthermophilic Archaeon Pyrococcus OT3. *Structure*, *15*(12), 1542-1554. doi: <http://dx.doi.org/10.1016/j.str.2007.10.015>
- Youngren, B., Nielsen, H. J., Jun, S., & Austin, S. (2014). The multifork Escherichia coli chromosome is a self-duplicating and self-segregating thermodynamic ring polymer. *Genes & Development*, *28*(1), 71-84. doi: 10.1101/gad.231050.113
- Yuhai Cui, Q. W., Gary D. Stormo, Joseph M. Calvo. (1995). A Consensus Sequence for Binding of Lrp to DNA. *J Bacteriol*, *177*(17), 4872-4880.
- Zhang, N., & Buck, M. (2015). A Perspective on the Enhancer Dependent Bacterial RNA Polymerase. *Biomolecules*, *5*(2), 1012-1019. doi: 10.3390/biom5021012
- Zhi, J., Mathew, E., & Freundlich, M. (1999). Lrp binds to two regions in the dadAX promoter region of Escherichia coli to repress and activate transcription directly. *Molecular Microbiology*, *32*(1), 29-40. doi: 10.1046/j.1365-2958.1999.01314.x