Regulation of Gene Expression Through Nucleic Acid Binding Proteins: New Paradigms, Perspectives, and Tools

by

Michael B. Wolfe

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biological Chemistry)
in The University of Michigan
2019

Doctoral Committee:

Assistant Professor Peter L. Freddolino, Chair
Associate Professor Mark A. Saper
Associate Professor Maureen A. Sartor
Assistant Professor Evan S. Snitkin
Associate Professor David L. Turner

Michael B. Wolfe
mbwolfe@umich.edu
ORCID iD: 0000-0002-0276-0551

Dedication

For Casey (1987-2009), who always pushed me to be the best I could be. You are always missed and never forgotten.

Acknowledgments

I want to thank the members of my thesis committee for their valuable support and advice throughout the course of my graduate work. I want to thank my mentor, Peter, for instilling a life long passion for computer science, statistics, data analysis, and bacteria in me. I have been incredibly fortunate to have you as a mentor and friend. I want to thank the members of the Freddolino lab, both old and new, for their friendship, advice, patience, and support throughout the course of my graduate career. Specifically, I would like to thank Grace Kroner for everything she has done to help me get to this stage in my career. I would also like to thank all of the co-authors on the papers that I have presented here. In particular, I would like to thank Aaron Goldstrohm for his efforts to help me develop as a scientist even after moving institutions. In addition, I would like to thank both Scott Scholz and Rucheng Diao for their fantastic work on the genome profiling project and inumerable brainstorming sessions on a variety of topics. I would also like to thank additional members of the Biological Chemistry Department, including Amanda Howard and Beth Goodwin, for their constant support especially in times of stress. Thank you to Bruce Palfey and Pat O'Brien for stimulating conversations, even at late hours and into the morning. They both always knew how to challenge me. I am indebted to Jane Jackman and members of the Jackman Lab for initiating my development as a scientist and igniting my passion for basic research. I would not be writing this dissertation today without their influence. I want to thank my Dad for always believing in me and pushing me to pursue my passions, no matter how hard it was to do so. I want to thank my Mom for constantly encouraging my curiosity and always helping me find the best environment to thrive. I want to thank Kelly, Brad, and their two (soon to be three!) wonderful children for supporting me and always providing me with laughter. I want to thank my in-laws for their constant support and love. Finally and most importantly, I want to thank my wonderful wife, Lindsay, for all her love and dedication. Her support has been a constant all the way back to the very beginning of my research career. She has been the greatest motivator. From my lowest lows to my highest highs, she has been there by my side to push and encourage me. She has always made me feel like anything is possible and I would not be writing this today without her constant support.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

More than 50 years ago, Jacob and Monod first laid out a model for the regulation of gene expression through the interaction of proteins and nucleic acids. With the advent of high-throughput sequencing and the unprecedented ability to take millions of simultaneous measurements of the same biological system, better understanding of the full complexity of gene regulation is finally being unraveled. Here, I present my graduate work investigating the control of gene expression at the transcriptional and post-transcriptional levels through the analysis of high-throughput biological measurements in bacterial and human cell culture. In the realm of bacterial transcriptional control, I explore the impact of a global regulator, Lrp, and its regulation of up to one third of the genes in *E. coli*. We identify a prevalent mode of "poised" binding, where Lrp is bound at a given promoter but only appears to have a regulatory effect under certain conditions. We find that Lrp appears to change its binding mode from a non-specific A/T-rich preference in early growth phases to a more sequence specific preference in later growth phases. I also look at promoter-independent effects on transcriptional regulation in *E. coli*. I describe key features of the bacterial genome that predict the position-dependent effect of transcription on a randomly integrated uniform reporter gene. We find that binding signal from the highly abundant nucleoid associated proteins H-NS and Fis act as key predictors for low and high transcription, respectively. We also find that integration proximity to a ribosomal RNA operon appears to be the single greatest contributor to position-dependent transcriptional activation. Absent of ribosomal RNA operon effects, we find that recent maps of *E. coli* chromosomal structure do not help us explain the periodic transcriptional signal obtained from our library. In eukaryotic systems, I focus on post-transcriptional control through the regulation of mRNA decay. I review methods that can be used to measure the decay of mRNA in a high-throughput manner. I highlight the importance of spike-in controls and demonstrate strategies to determine relative mRNA decay between experimental conditions at minimal cost. Finally, I present my work identifying key sequence features that allow for the prediction of mRNA decay mediated by the human Pumilio proteins. I find that contextual sequence features around predicted PUM binding sites contribute meaningful information to the prediction of PUM-mediated post-transcriptional regulation. I also demonstrate that human Pumilio proteins primarily modulate RNA abundance through controlling mRNA decay and not through control of transcription. Taken together, my graduate work provides a comprehensive view of the regulation of gene expression at both transcriptional and post-transcriptional levels.

CHAPTER 1

Introduction: The use of information in biological systems

1.1 The Central Dogma of Biology

Information, through its efficient transfer, encoding, use, and storage, is fundamental to biology and the processes of life. Inherent at the beginning of life, and in the critical process of reproduction, is the transfer of information. This key notion was initially proposed in Ancient Greece by Pythagoras and later refined by Aristotle [1]. In discussing animal reproduction in *De Generatione Animalium*, Aristotle observed that:

> Just [as] no material part comes from the carpenter to the material, i.e. the wood in which he works, nor does any part of the carpenter's art exist within what he makes, but the shape and form are imparted from him to the material by means of the motion he sets up ... it is his knowledge of his art, and his soul, in which is the form ... In like manner, ... Nature uses the semen as a tool ... [2]

Here, Aristotle is suggesting that, just like a carpenter imparts their knowledge to create form out of raw materials through the use of tools, Nature is using some biological material as a tool to *transfer* the information needed to create the form of another organism. Although Aristotle goes on to incorrectly characterize aspects of reproduction in inherently sexist language, the key idea that *information* is transferred in the act of reproduction was born. It was not until the 1800's that experiments shedding light on the nature of this informational tool began in earnest with the Augustinian friar, Gregor Mendel, and his pea experiments. Gregor was able to show, with careful breeding of different pea hybrids and fortunate choices of observable biological traits, that the offspring of pea hybrids tended to have a trait that took after one or the other of the pea plant's parents. These experiments suggested that, not only was information transferred from parent to offspring, but those pieces of information were being transferred in an apparently discrete manner [1, 3]. Further work in the early 1900's by Thomas Hunt Morgan, this time using the now model organism *Drosophila Melanogaster*, a species of fruit fly, demonstrated that some of these units of heredity for certain traits tended to transfer to offspring together as if they were physically “linked.” He suggested that it was the physical proximity of these discrete units of information on chromosomes—large structures which were observable under the microscope in dividing cells—that allowed for them to be linked together during the process of reproduction [4]. However, the chemical nature of the heritable material was still unknown.

1.1.1 DNA as the information carrier

A critical set of experiments in the 1920’s by Frederick Griffith showed that material from a heat-treated and killed virulent strain of bacteria could be used to “transform” a living and non-virulent strain into a virulent strain. Moreover, the previously non-virulent strain now remained virulent over many generations. This suggested that information was being transferred between the dead virulent strain of bacteria to the living non-virulent strain. It also suggested that the material nature of this information could stay intact even after the cell was dead, and, once transferred to the living bacteria, could continue to pass on to progeny [5]. This led Oswald Avery and colleagues to determine, through careful purification of nucleic acids and treatment with crude preparations of enzymes to degrade the nucleic acids, that deoxyribonucleic acid (DNA) was the “transforming agent” that was necessary for Griffith’s observations [6]. It was known at the time that the chromosomes that Morgan was observing were made up of both protein and nucleic acid; however, it was widely believed that protein, not nucleic acid, was the key information carrier. Through the use of radioactively labeled P^{32}-DNA and S^{35}-protein, Hershey and Chase demonstrated that primarily P^{32}, and not S^{35}, was transferred into bacteria by a bacteriophage upon infection. This suggested that DNA, and not protein, was the information carrier consistent with Avery’s results [7].

1.1.2 The structure of DNA

Once the chemical identity of the molecule carrying genetic information was known, the stage was set for Watson and Crick, together with work from Franklin, Gosling, and Wilkins, to create the first model of the structure of DNA. In 1953, Watson and Crick first described the structure of DNA, an elegant two stranded structure with basepairs between adenine and thymine, and guanine and cytosine, which “immediately suggests a possible copying mechanism for the genetic material” [8]. The discovery of DNA—central to information processing within every cell—is one of the most ground-breaking scientific advances of the 20th century. However, although the essential nature of the information carrier had been determined, questions still remained about how that information is used within the cell, (i.e., how does the essential structure of DNA give rise to the *form* of all living things?).

1.1.3 Information is stored in the DNA sequence

Immediately after the discovery of the structure of DNA, work began to test if the "copying mechanism" that the DNA structure so evocatively suggested (i.e., one strand can be used as a template to reproduce the other strand) was indeed how DNA was reproduced. This idea of "semi-conservative" replication was confirmed through careful experiments conducted by Meselson and Stahl in which they grew *E. coli* in media containing N^{15} until DNA was fully labeled and then shifted growth into media containing N^{14}. By tracking the density of the labeled DNA over time, they found that the DNA only existed at densities consistent with either all N^{14}, all N^{15}, or an intermediate density consistent with half of both species, thus confirming semi-conservative replication [9]. This, together with the structural model of DNA, also suggested that since the order of the bases was kept intact under replication, then some sort of code may be stored within the actual sequence of bases [10].

1.1.4 Information is copied from DNA into RNA

Broad clues as to how information could be stored and used by organisms were first described by Beadle and Tatum, who showed that individual biological functions could be disrupted by inducing mutations in the bread mold *Neurospora crassa* while it grows on a "complete" media and then observing the mutated bread mold's growth on minimal media. Mutants deficient in synthesizing a particular metabolite could then be discovered by adding in individual components present in the complete media but absent in the minimal media until the newly essential metabolite was found [11]. This led to the idea that a single gene could encode for a single biological function or a single enzyme in a given pathway and was thus named the "one gene-one enzyme" hypothesis. However, the nature of how information was transferred from DNA to protein enzymes was still unclear. It was known that ribosomes, largely made of ribonucleic acid (RNA), were the locations of protein synthesis and it had been proposed that genes each encoded a special type of ribosome that could then create a single protein. However, a key set of work performed simultaneously by both Brenner et al. [12] and Gros et al. [13], discovered a new class of metabolically unstable RNA that was associated with ribosomes in bacteria during an active phage infection. Using radioactively labeled precursors to RNA, they showed that newly synthesized RNA with nucleotide base ratios that matched the DNA of infecting bacterial phage (with the appropriate substitution of Ts for Us) were associated with metabolically stable ribosomes. This suggested that: 1. the phages were not producing new ribosomes to create proteins needed for their function and 2. the highly unstable RNA—which they called "messenger" RNA (mRNA)—was synthesized as a copy of the genetic information contained within the phage DNA and was being used as a template by the ribosome to create a protein.

1.1.5 Cracking the genetic code

Following these experiments, the essential nature of this code, the genetic code, was still not known. How could a sequence of four nucleotide bases be used to specify a different sequence of 20 amino acids? Based solely on the number of amino acids and nucleotide bases, initial proposals suggested a triplet code—three nucleotides for every one amino acid—was probable as it was the smallest number of bases that would allow for the encoding of all 20 amino acids [14, 15]. Using mutations that introduced a single nucleotide insertion or deletion to a gene, Crick et al. [16] showed that the genetic code was a non-overlapping triplet code. However, it would take Nirenberg and Matthei's work to finally crack the code that mapped any given nucleotide triplet to its corresponding amino acid. Using poly-U RNA in a cell-free translation system, they first determined that the triplet UUU coded for the amino acid phenylalanine [17]. A highly competitive race to uncover the genetic code followed this discovery and a flurry of papers from Nirenberg and additional colleagues at the National Institutes of Health, as well as work from their main competitors Severo Ochoa and Gobind Khorana, resulted in a full solution for the genetic code, including the three codons that encoded for the end of a message [18–21].

1.1.6 tRNAs as a molecular Rosetta Stone

The final piece of the puzzle, the adapter molecule that allows for "translation" from the mRNA nucleotide code to the amino acid protein code, was first proposed by Crick [22, 23] and the structure of the key molecule, "transfer" RNA, was elucidated by Robert Holley using a series of nucleic acid digests to determine the first structure of a nucleic acid [24, 25]. Transfer RNAs (tRNA) are a special and highly abundant class of RNAs that form into cloverleaf-like shapes of three stem loops. The middle stem loop contains a three nucleotide anticodon that base-pairs with the corresponding codon on an mRNA molecule. The amino acid that corresponds to the matching codon for a given tRNA is added to the acceptor stem at the 3' end of the molecule by specialized enzymes in a process called aminoacylation [26]. During protein synthesis, the ribosome matches the mRNA codons with the tRNA anticodons, thereby stringing together amino acids in the sequence determined by the triplet genetic code. Thus, the tRNA represents a biological "Rosetta stone," serving as the translator between the nucleic acid and amino acid codes [27, 28].

1.1.7 The Central Dogma and beyond

Taken together, these discoveries represent a framework for describing how biological information is stored and inherited, as well as how that stored information is subsequently used to create functional biological molecules in the form of proteins.[1] These ideas culminated in what is now called "The Central Dogma of Biology." Represented in Figure 1.1, the term Central Dogma represents the flow of information in biological systems and was coined by Crick [22] when he first referred to the idea that once information is passed to the protein, information cannot go backwards to nucleic acid. In this model, the DNA acts as a storage molecule with the instructions needed to make active biological agents. Through the process of *transcription*, one strand of the DNA is used as a template to create a nucleic acid copy of a short stretch of DNA that encodes for a particular protein. The copy— a short stretch of RNA, called mRNA (described above)—is subsequently *translated* into protein through the use of the ribosome and tRNA adapter molecules. However, in direct contrast to Crick's original meaning for Central Dogma, I have drawn two additional arrows in red representing the flow of a *higher level* of information from protein back into the processes of transcription and translation. These arrows are no longer representing the flow of information at the level of direct manipulations of the same underlying symbols of the genetic code itself, either through transcription (nearly direct copies) or translation (conversion from one set of symbols to another). Instead, they represent a higher level of information transfer that acts on the biological mechanisms used to interpret the genetic code. Thus, in these cases, the act of translating a message gives the system information about how to translate future messages, much like a hypothetical record player that uses the sound waves from a previously played song to determine which song to play next.[2] Through gene regulation (the information transfer represented by these arrows) the DNA becomes more than just a passive information carrier; rather, through the very act of self-reference—changes in the *process* of expressing genes brought on by the expression of genes *themselves*—the system is able to control its own fate and act directly upon itself. It is this process of gene regulation, the control of the information flow in biology (as stored in nucleic acids) through the action of the end product of that information flow (i.e., proteins), that is the central focus of this thesis and my graduate work.

For the rest of this section I will focus on a detailed look at two overall mechanisms of gene regulation on which my graduate work has focused: transcriptional regulation (control of the synthesis of mRNA) and post-transcriptional regulation (specifically, control of the degradation of mature mRNA). Subsequent chapters will give detailed overviews of specific regulatory systems involved in each of these processes and references will be made to each chapter as they relate to the concepts being presented.

[1] For a more comprehensive overview on this subject, including details on the lives of each of the key scientists, see Siddhartha Mukherjee's *The Gene*[1].

[2] My thoughts on self-reference in biological systems are heavily influenced by Douglas Hofstadter's *Gödel, Escher, Bach*[29], in which he uses similar analogies to describe information transfer in the Central Dogma.

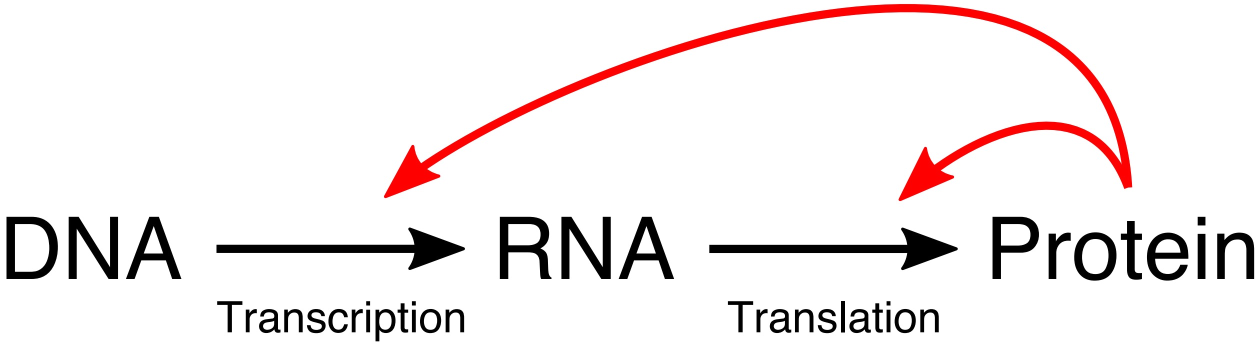

Figure 1.1: The Central Dogma of Biology. Red arrows represent gene regulation resulting from information flow at a higher level than strict symbol manipulation of the genetic code.

1.2 Transcriptional control of gene expression

Strong evidence for the existence of gene regulation first came from work describing how β-galactosidase activity was "induced," or turned on, in *E. coli* upon growth in lactose media. It had been observed that when growing *E. coli* in a mixture of glucose and lactose acting as a carbon source, the bacteria first consume the glucose exclusively. Once the glucose was completely consumed, a short lag time in growth was observed until the metabolic enzymes required for consuming lactose were created [30]. Thus, the activity required for lactose digestion, β-galactosidase, was induced by the presence of lactose in the absence of glucose. Using a series of mutants that were deficient either in the enzyme responsible for the β-galactosidase activity, or involved in controlling inducible vs. constitutive expression of the system, Jacob & Monod proposed that the inducibility of the β-galactosidase activity was controlled by a separate genetic element that acted as a "repressor" by preventing the synthesis of the gene involved in producing the β-galactosidase enzyme. In their landmark 1961 paper, Jacob & Monod synthesized their ideas about this system and other systems that were being identified at the time. They proposed that information flow from gene to protein must go through some intermediate messenger, that control of this messenger must occur in regions of the genetic sequence that do not code for protein (regions they termed "operators" to distinguish them from "structural" genes like the β-galactosidase enzyme), and that the structural genes required for a particular biological function are linearly connected in the sequence both to each other and to an operator. This final idea they termed an "operon" and they suggested that the operons could be controlled through their operators [31]. This influential paper inspired the discovery of messenger RNA (as discussed in Section 1.1.4), and began the study of mechanisms that control the expression of genes.

1.2.1 Major steps in transcription

The synthesis of RNA is universal to all domains of life and is carried out by a specialized enzyme complex known as RNA polymerase. Although the specific details of transcription and the exact identity of the molecular players differ between organisms, the general process is the same. Transcription occurs in three major steps: 1. initiation, in which the RNA polymerase binds upstream of the gene to be expressed and opens up the DNA helix to access the correct template strand, 2. elongation, where the RNA polymerase tracks along the DNA and adds the nucleotide base complementary to the template strand on to the 3' end of the growing nascent RNA, and 3. termination, in which the RNA polymerase halts synthesis of the nascent RNA and is released through various mechanisms. Control can be exerted at each step in transcription, but for the purposes of this thesis, I am primarily concerned with the processes of initiation and, to a lesser extent, elongation (termination will not be discussed).

1.2.2 σ factors and their role in initiation

In bacteria, particularly in the well studied model organism *E. coli*, the core RNA polymerase (RNAP) complex alone is capable of initiating transcription with low efficiency, but it requires a series of proteins called σ factors to increase affinity for regions of the DNA upstream of the transcription start site (TSS) and allow for specificity in the initiation of transcription [32]. These regions, called promoters, allow for the coordination of genes with similar functions to be expressed at the same time. For example, σ^{70}—the most abundant σ factor—directs RNAP to genes involved in general housekeeping processes that are constitutively expressed [33]. In contrast, σ^{54}, directs RNAP to genes involved in nitrogen metabolism and often requires an activator protein, in addition to the σ^{54}-RNAP complex, to initiate transcription at these genes [34–36]. Upon the transition from initiation to elongation, the σ factor is displaced from the RNAP and replaced by the elongation factor protein NusA. This allows σ factors to be rapidly reused by additional RNAP complexes for subsequent rounds of initiation [37]. Thus, key to predicting which genes are initiated by which σ-RNAP complex is the determination of the particular promoter *sequence* that the σ-RNAP complex recognizes.

1.2.3 Sequence elements controlling bacterial initiation

The first attempts at determining a "consensus" sequence for bacterial promoters were conducted with the σ^{70}-RNAP complex, due to its high abundance in the cell. Work by two separate groups characterized two sites that were specifically recognized by this complex: TATRATG, at 10 basepairs upstream from the start of RNA synthesis (-10 site), and GTTGACACTTTA, at 35 basepairs upstream (-35 site) [38–40]. Further work established the identity of the UP element—a large AT rich element upstream of the -35 site that stimulates transcription by σ^{70}-RNAP *in vitro* for some promoters [41]. However, simple recognition of a general promoter sequence does not explain how specific genes are regulated. For example, how do systems that turn off gene expression, such as the mysterious repressor discovered by Jacob and Monod, work?

1.2.4 The Lac repressor, a bacterial transcription factor

Continued work on the β-galactosidase operon, or the *lac* operon, allowed for a detailed understanding of single operon repressors. The Lac repressor attenuates transcription of the *lac* operon by binding to two operator sites, one directly downstream and one directly upstream of the promoter region for RNAP and mediating the formation of a loop in the DNA [42]. It has been shown, through a combination of mathematical modeling and targeted *in vivo* experiments, that the primary role of this looping is to occlude the RNAP from binding to the promoter region [43]. Thus, the Lac repressor is acting as a transcription factor—defined here as a protein that modulates the activity of RNAP at a promoter region—to control the expression of the *lac* operon. In order for de-repression, or induction, of the *lac* operon to occur, lactose binds to the Lac repressor in an allosteric site, which induces a conformational change in the protein that disrupts its affinity for the operator sites [44]. Thus, in the presence of high lactose concentrations, the Lac repressor no longer binds to its operator site, thereby freeing up the promoter to be bound by RNAP and the initiation of transcription of the *lac* operon to begin. However, repression of expression through the Lac repressor alone does not fully explain regulation of the *lac* operon. How does the bacterium decide to only metabolize glucose and not lactose when both are present in the media?

1.2.5 Combinatorial control from the global regulator CRP

The current model explained thus far suggests that the *lac* operon should be expressed whenever lactose is present, regardless of the concentration of glucose in the media. It turns out that another transcription factor, the cyclic AMP (cAMP) receptor protein (CRP), allows for the fine-tuning of the *lac* operon by activating *lac* operon expression only when glucose levels are low [45]. cAMP is created from ATP by the enzyme cAMP phosphodiesterase and serves as a molecular signal for low glucose levels. Upon binding of cAMP to CRP, CRP binds to DNA in a sequence specific manner [46]. Activation of transcription by the CRP protein occurs through interactions with the RNAP that promote RNAP binding to the promoter and, in some cases, facilitate conformational changes in the RNAP-promoter complex that allow for the completion of transcriptional initiation [47]. Thus, in the case of media that is rich in both lactose and glucose, the Lac-repressor is not blocking the promoter of the *lac* operon due to the excess lactose in media. However, due to the high concentration of glucose, the cAMP-CRP complex is not formed. Thus, the absence of cAMP-CRP complex binding upstream of the *lac* operon, results in repression of expression due to lack of activation of the RNAP by CRP. Once the glucose is fully utilized by the *E. coli* cells, the CRP transcription factor can then promote transcription of the *lac* operon. However, unlike the Lac repressor's specificity for a single promoter region, CRP proteins bind hundreds of different promoters in the *E. coli* genome and can either activate or repress transcription at those promoters [48]. In addition, CRP regulates the expression of both σ factors and other identified transcription factors, making it a regulator of regulators—a global regulator [49].

1.2.6 Global regulators coordinate large regulatory networks

In contrast to single promoter regulators, like the Lac repressor, global regulators influence the actions of many promoters. They frequently work together with other co-regulators to exert control over those promoters, and also tend to work together with other global regulators. In addition, global regulators work together with promoters from different σ factor classes and tend to sense and respond to a large number of growth conditions. Finally, global regulators tend to occur in their own isolated regions of the genome and their expression is controlled by feedback from themselves [49]. The Leucine-responsive regulatory protein (Lrp) is one key global regulator in *E. coli* that is well-studied, but still not well-understood. Lrp was first identified by its activation of the *ilvIH* promoter which is attenuated in media containing leucine [50]. The *ilvIH* operon is one of three operons that encodes an aceto hydroxyacid synthase which performs the first common step in the biosynthesis of the branched chain amino acids [51]. In addition, it was found that Lrp also acted as a negative regulator of the *oppABCDF* operon—an operon involved in the transport of oligopeptides—whose repression was attenuated in the presence of leucine [52]. Thus, Lrp was identified as a dual-regulator involved in controlling amino acid synthesis and transport in response to leucine levels. Early experiments indicated that leucine could disrupt the equilibrium of Lrp binding at the *ilvIH* promoter [53]. Identification of additional promoters that are regulated by Lrp demonstrated that for some promoters—such as *livK*—Lrp needed leucine for negative regulation [54]. This was inconsistent with a simple model of leucine disrupting Lrp binding in all cases. Consistent with its role as a global regulator, microarray experiments have indicated that Lrp regulates about 70% of the 215 genes with differential expression upon entrance to stationary phase [55]. Tani et al. [55] also identified a subclass of genes whose regulation did not depend on presence or absence of leucine in the media, suggesting that leucine is not the only signal Lrp responds to. Lrp itself is thought to exist primarily as an octamer *in vivo* and dynamic light scattering experiments have demonstrated that leucine disrupts the formation of higher order hexadecamer oligomeric states—forcing the equilibrium towards the octameric form [56]. Furthermore, additional amino acid co-regulators have been shown to modulate the activity of Lrp at the *livK* promoter, resulting in everything from activation to repression depending on the amino acid present [57]. Recently, Cho et al. [58] used high-throughput assays to explore the full regulatory network of Lrp in the presence and absence of leucine. However, this study is confounded by the use of a C-terminal 8xMyc tag that is on the order of half the size of a single Lrp monomer. The C-terminus has been shown to be important for the oligomerization of the Lrp protein [59] and the presence of a large tag on the C-terminus likely disrupts native oligomerization. In Chapter 2, I present our published work on Lrp. To alleviate the need to use a tagged protein and remove concerns about disruptions to native oligomerization, we developed a monoclonal antibody to Lrp. We then used modern high-throughput sequencing techniques to identify both the binding locations and RNA expression in a combination of three different media conditions and three different time points. In that chapter, we provide evidence that Lrp regulates up to one third of the *E. coli* genome. We find that Lrp appears to bind at promoters even in conditions where no Lrp-dependent regulation is observed, suggesting that Lrp is sitting in a poised position in preparation for regulatory activity under the correct condition. Further, we hypothesize that like the Lac repressor, Lrp may cause looping in the DNA to control access to a given promoter.

1.2.7 Technologies enabling the resolution of regulatory networks

The full complexity of gene regulation mediated at promoter regions in *E. coli*, as described above, conjures up a tangled web of regulation and co-regulation resulting from interactions between many different factors. Current understanding, through targeted biochemical experiments, has allowed for the determination of the key regulators and, for some, mechanistic detail about how they work. However, one needs to understand the *connections* between the regulators and their targets to fully understand gene regulation and model biological systems. With the advent of high-throughput sequencing, key techniques have been developed that allow for the simultaneous measurement of transcription factor binding sites and steady-state RNA levels across the entire genome and transcriptome.

The first of these techniques, RNA sequencing (RNA-Seq) allows for the measurement of the relative abundance of RNA within a pool of purified RNA [60]. The second of these techniques, chromatin immunoprecipitation and sequencing (ChIP-Seq), allows for the empirical determination of the locations of binding for a specific transcription factor within the genome [61]. ChIP-Seq is performed by using a chemical, such as formaldehyde, to cross-link protein to the DNA *in vivo*. After cell lysis and sonication of the cellular debris, an antibody specific to the protein of interest is used to pull-down DNA specifically cross-linked to the protein. This is compared to an input sample which either consists of the cellular debris before pull-down or mock pull-down with a non-specific antibody. Both the input and the specific pull-down sample are then sequenced and, through computational analysis, regions of DNA that were enriched in the specific protein sample can be identified. Many different computational techniques have been developed to analyze ChIP-Seq data [62–65]. However, in Chapter 2, I present an analysis pipeline I developed to specifically address our unique experimental setup.

Together these technologies, and their older iterations using microarrays instead of high-throughput sequencing, have allowed for the resolution of regulatory networks for each of the σ-RNAP complexes [66], the CRP regulon [67, 68], and the regulons of many transcription factors in *E. coli*. These regulons are curated and maintained in a large database, RegulonDB, which is free for academic use [48]. With this database and other databases in hand, it appeared that the original call to mathematically model an *E. coli* cell [69] would finally be realized. However, recent attempts to use current databases to do this fell well short of the mark [70], suggesting that our current set of regulatory interactions is incomplete.

1.2.8 Nucleoid-associated proteins and their role in transcription

The current understanding of transcriptional regulation presented thus far has focused around control at the stage of initiation through direct interactions at promoter regions. These interactions can be thought of as local and highly specific interactions between a given transcription factor and the DNA. However, larger regions of protein-DNA complexes exist within the bacterial nucleoid. Work focused on isolating proteins that share similar properties with eukaryotic histones, illuminated a set of proteins named "Nucleoid-associated proteins" (NAPs) for their high abundance, general propensity to bind non-specifically to DNA [71], and role in controlling chromosomal structure [72]. Although these proteins share similar properties to eukaryotic histones, they appear to be specific to bacteria [73]. However, homologues of the most highly conserved NAP, HU, have been seen in both yeast mitochondria [74] and spinach chloroplasts [75], suggesting a broad functional role for this class of proteins. The need for constraining and controlling DNA structure and accessibility is likely a universal problem for biological systems and it perhaps not surprising that the different domains of life have come up with unique solutions. Analysis of the expression patterns of several identified NAPs revealed large changes in expression across growth-phases for twelve of the identified NAPs, suggesting functional roles for NAPs at different stages of growth [76]. Furthermore, visualization of different NAPs' locations throughout the nucleoid allowed for separation of the proteins into two groups: 1. those that formed localized clusters and 2. those that seemed to bind diffusely throughout the DNA [77]. Further evidence for large regions of protein occupancy was found using a technique to isolate protein-DNA complexes in a non-specific manner. Using this technique, Vora et al. describe large transcriptionally silent regions of extended protein occupancy (tsEPODs) that suggested a silencing role for the complexes [78].

Silencing of transcription through general protein occupancy was also consistent with the known role of the Histone-like nucleoid-structuring protein (H-NS) in silencing AT rich DNA by blocking RNAP access to the promoter region [79–82]. ChIP-chip (ChIP-Seq with using a microarray instead of sequencing) and ChIP-Seq studies have indicated that H-NS does indeed form large regions of protein occupancy across the bacterial chromosome [83, 84]. It has been suggested that H-NS's primary biochemical role is to mediate DNA-DNA contacts by bridging together two strands of DNA [85, 86]. In fact, *in vitro* studies using H-NS and its partner proteins StpA and Hha have shown that, not only can H-NS prevent transcriptional initiation, but it can also promote the pausing of RNAP during elongation, presumably by trapping the RNAP in a constrained topoisometric state [87]. However, H-NS does not appear to be the only NAP bound up in tsEPODs, as experiments with a Δ*hns E. coli* strain have shown that tsEPODs remain even in the absence of H-NS (Peter Freddolino, personal communication).

Additionally, NAPs do not strictly silence gene expression by blocking initiation or disrupting transcriptional elongation. The NAP, factor of inversion stimulation (Fis), has important roles in activating the transcription of several genes. Like H-NS, Fis binds throughout the bacterial genome, but it has more distinct peaks around promoter regions, albeit with a more diffuse background binding signal [84]. Fis has been shown to increase the expression of a *lacZ* gene fused to the *rrnB* promoter up to 20-fold *in vitro* [88]. It is thought that Fis directly interacts with RNAP upstream of the UP site to promote transcription in this manner [89]. For some promoters, it has been proposed that Fis mediates the local supercoiling of the DNA to facilitate transcriptional initiation, a mechanism that suggests a more general role for Fis activation beyond specific interactions with RNAP [90]. The targeted nature of Fis regulation makes Fis more akin to a global regulator than a NAP [91], but it also highlights the fluidity of these assignments and the difficulty of drawing distinct boundaries for proteins with multiple avenues of regulatory control. The promoter-independent aspects of transcriptional control have been best studied by using reporter constructs with identical promoters and inserting these reporters throughout the *E. coli* genome. Initial experiments of this type suggested that position-dependent effects on transcription could be explained by a simple gene-dosage effects—rapidly dividing *E. coli* cells have multiple copies of genomic sequence around the origin compared to a single copy around the terminus. Thus, reporters inserted closer to the origin of replication had higher measured transcription levels than reporters inserted near the terminus [92]. However, a later study found that reporters integrated in regions coinciding with tsEPODs had lower expression than one would expect strictly from gene dosage effects [93]. For both of these studies, observations were made from only a handful of sites, limiting the ability to make any sort of broad conclusions about position-dependent transcription in *E. coli*. In Chapter 3, I present our published work on position-dependent effects on transcription at an unprecedented resolution of an average of one unique integration site every 47 bp in the *E. coli* genome. We find that both Fis and H-NS are highly correlated with position-dependent expression levels of a reporter gene with an identical promoter in each integration site. Consistent with the previously identified functional roles of each NAP, we find that H-NS occupancy is negatively correlated and Fis occupancy is positively correlated with transcription levels from our reporter library. Taken together, we find that position-dependent and promoter-independent effects represent an additional layer of regulatory control beyond the gene-specific impact of particular promoters.

1.3 Post-transcriptional control of gene expression

The process of transcription is not the only process where an "operator," as proposed by Jacob and Monod [31], could be used to control the expression of a gene. After the message has been created, it still must be translated from the nucleic acid code into the amino acid code through the actions of the ribosome. Thus, the mRNA represents a second chance to intercept or enhance the expression of a gene through post-transcriptional control. In bacteria, the process of transcription and translation are coupled and there is less room for control at this level [94]. Although several mechanisms for post-transcriptional control in bacteria do exist, particularly the well studied Hfq-sRNA system which uses base-pairing between a class of small RNAs (sRNAs) and target mRNAs to either occlude the ribosome from initiating translation or recruit RNAses to degrade the mRNA [95, 96], these systems are not the focus of this thesis. I will instead focus on mechanisms for post-transcriptional control in eukaryotic systems.

Unlike bacteria, eukaryotes have a distinct nucleus that is separated from the cytosol by a nuclear membrane, decoupling the process of transcription from translation. The separation of transcription and translation allows for a larger role for post-transcriptional regulation, as the nascent mRNA message can be targeted at many steps before and after it is exported from the nucleus and into the cytosol for translation [97].

1.3.1 Mature mRNA formation in eukaryotes involves several steps

Messenger RNAs in eukaryotes require processing at several stages to mature from a pre-mRNA into a fully translatable mRNA. Although my thesis work is primarily focused on controlling the decay of a particular message, an understanding of the complementary reactions that create the key elements of a messenger RNA in eukaryotes is needed for putting the subsequent decay reactions and their post-transcriptional control in context. The stages of mRNA formation include 5' end capping, splicing, 3' end cleavage, and polyadenylation. Many of these processes happen simultaneously and most are coordinated through the actions of RNA-binding proteins (RBPs). Mature mRNAs are capped on the 5' end by an N^7 methylated guanosine residue that is connected by a 5' to 5' connection to the 5' end of the nascent RNA. Capping happens co-transcriptionally early during the synthesis of mRNA and is added by a complex of proteins that differs from organism to organism [98]. Once created, the 5' cap is bound by the Cap Binding Complex (CBC), which plays various roles in controlling splicing, export, and ultimately, the initiation of translation [99].

Unlike bacterial genes, eukaryotic genes do not occur in continuous linear regions of the genome; rather, they contain regions of coding sequence (exons) interspersed with regions of non-coding sequence (introns). Thus, in order to create a mature mRNA molecule that will ultimately code for a protein, the introns must be cut out and the exons must be stitched together. This process is called splicing and is carried out by a specialized complex made of both RNA and protein called the spliceosome. Through binding to the pre-mRNA, RBPs can mediate different choices of exons to include or exclude for a given message, which leads to a substantial expansion of the possible isoforms that can be encoded for from a single gene [100]. Splicing can also be mediated through self-splicing introns, which were the basis for the discovery of catalytic RNA [101]; however, splicing mediated through the spliceosome, like capping, often occurs co-transcriptionally [102].

After establishing the 5'end through capping and the central region of the mRNA through splicing, the final aspect of the pre-mRNA must be processed to be converted into a mature mRNA. Present in the 3' un-translated (3'UTR) region of the nascent RNA is the sequence of AAUAAA, a U-rich upstream element, and a G/U-rich downstream element that, together, serve as a signal to cleave the nascent mRNA and subsequently, act as a substrate for non-templated addition of adenosine residues (polyA tail) to the 3' end. Like the other steps described in this section, cleavage and polyadenylation also occur co-transcriptionally and serve as the signal for transcriptional termination [103]. However, many transcripts have several separate alternative cleavage and polyadenylation sites, the choice of which can influence the identity of the 3'UTR and lead to differential regulation downstream [104]. The polyA tail plays an important role in the stability of the mRNA, and transcripts with shorter polyA tails tend to be degraded faster [105, 106]. Additionally, interaction of the 3' polyA tail with the polyA binding protein PABP, together with the 5' cap, form a loop that is stimulatory for the initiation of translation [107]. Thus, the polyA tail plays a vital role in controlling the fate of the mature mRNA transcript.

Together, each of these steps define the anatomy of the mature eukaryotic mRNA. Starting from the 5' end the transcript consists of a 5' cap, a 5' untranslated region (5'UTR), a coding region made up of spliced together exons, a 3'UTR, and finally the polyA tail. Each of these regions of the mature RNA have a role to play in its life-cycle and ultimate translation, but the 3'UTR serves as one of the key operators that Jacob & Monod had postulated over fifty years ago [108].

1.3.2 Control of mRNA stability allows for a key locus of control

Like the production of mRNA, the decay of a eukaryotic mRNA is a multistep process involving large enzyme complexes that act on each end, as well as the middle, of the mRNA. Each of these steps represents a locus of putative post-transcriptional control and several different regulatory factors take advantage of each stage of mRNA decay. Just as the cap and polyA tail were added to the mRNA, they are also removed to allow access for exonucleases to facilitate mRNA decay. Alternatively, endonucleases can cleave at internal sites in the mRNA to allow for access of free 5' and 3' ends to be degraded by exonucleases.

RNA decay in the cytoplasm is typically initiated through disruption of the translational initiation complex involving the loop-like structure formed between the interactions of PABP and the translation initiation factor 4F—a replacement for CBC in the cytosol after the pioneer round of translation [109]. This process is thought to start primarily through deadenylation of the polyA tail at the 3' end of the mRNA [110]. The primary enzyme complex involved in deadenylation is the Ccr4-Not complex, a multi-component enzyme complex with two catalytic members, Ccr4 and Caf1. Ccr4 (CNOT6 and CNOT6L in humans) and Caf1 (CNOT7 and CNOT8 in humans) both act together as catalytic subunits and each have 3'→5' exonuclease activity that is polyA-specific [111, 112]. The scaffold protein NOT1, the only subunit of the Ccr4-Not complex that is essential in yeast, is involved in facilitating interactions between Ccr4, Caf1, and a menagarie of additional proteins involved in making up the Ccr4-Not complex [112]. A secondary deadenylase complex, the Pan2-Pan3 complex, also specifically degrades polyA tails in the 3'→5' direction; however, it appears to only degrade up to the last PABP protein [111]. Thus, a model has emerged suggesting that the Pan2-Pan3 complex has a 3' end trimming activity, whereas the Ccr4-Not complex is involved in a second phase of deadenylation and polyA shortening [111, 113].

Deadenylation of an mRNA transcript results in the dissociation of PABP and the disruption of the translation initiation complex. This disruption allows for the decapping complex, Dcp1-Dcp2, to remove the 5' m^{7}G cap from the mRNA. Outside of deadenylation, additional enhancer proteins can help facilitate cap removal, even during translation elongation [114].

The removal of the 5' cap and 3' polyA tail opens up the mRNA for attack by both 5'→3' and 3'→5' exonucleases for the full degradation of the mRNA transcripts. The key 3'→5' exonuclease in mammalian cells is an enzyme complex known as the exosome. After deadenylation, the RNA exosome, together with the Ski proteins, proceeds to degrade the mRNA through the 3'UTR and beyond [115, 116]. On the 5' end of the molecule, XRN1 and XRN2 act as the 5'→3' exonucleases to degrade the mRNA after decapping [117]. By acting together, either simultaneously or alone, these two pathways serve to completely degrade mRNA within the cytoplasm.

1.3.3 Quality control and surveillance pathways promote decay through endonucleases

Decapping and deadenylation are not the only pathways that can lead to free 5' or 3' ends for XRN1, XRN2, and the RNA exosome to act. Endonuclease activity is involved in both surveillance and quality control mechanisms to target mRNAs for degradation. Three pathways involved in the quality control of mRNAs through detection of the improper translation of a message include Nonsense-mediated decay (NMD), nonstop decay (NSD), and no-go decay (NGD). NMD involves the detection and degradation of mRNA transcripts with a premature stop codon, NSD detects and degrades mRNAs with no stop codon, and NGD degrades mRNAs with a stalled ribosome [118]. NMD is the best studied of these pathways and occurs when the UPF1-SMG1 complex, associated with the terminating ribosome, interacts with a UPF2-bound exon junction complex downstream of the stop codon during the first round of translation [109]. A phosphorylation event occurs on the UPF1 protein that then inhibits additional rounds of translation and promotes RNA decay, perhaps through the recruitment of DCP1 and XRN1 [119]. Additional evidence suggests that NMD mediated decay can also occur through an endonucleolytic pathway through the catalytic activities of the SMG6 protein. Additionally, both NGD and NSD are also thought to occur through endonucleolytic pathways[118]. Thus, these pathways represent a last line of defense for detecting mistakes in messages and allow for rapid acceleration of the decay pathways through the exonucleases by quickly creating free 3' and 5' ends through an endonucleolytic cleavage event [118].

In contrast to a general affinity for disrupted messages through the quality control pathways, targeted control of specific mRNAs is achieved through the use of small non-coding RNAs—called micro-RNAs (miRNA)—that, together with their protein partners, inhibit translation and promote decay. Mature miRNAs are single stranded RNAs with an average length of 22 bases that are bound by the Argonaute proteins to together make a RNA-induced silencing complex (RISC) [120]. RISC complexes use the sequence of their bound miRNAs to recognize target RNAs through complementary base-pairing, typically in the 3'UTR of target transcripts [121]. In one mechanism of post-transcriptional control by miRNAs, the Ago2 protein of the RISC complex cleaves the target mRNA in an endonucleolytic fashion after site specific recognition using the miRNA [122]. After cleavage, the mRNA then becomes a substrate for the typical exonucleases to fully degrade the mRNA. However, miRNA-induced silencing of gene expression has also been shown to recruit factors involved in both deadenylation and decapping, and this is now thought to be the primary mechanism of silencing by RISC complexes [123]. Due to the sequence specific manner of miRNA binding, several groups have bioinformatically determined putative miRNA binding sites for known miRNAs to varying degrees of accuracy [124]. Similar to global regulators in bacterial transcriptional control, miRNAs target a large number of mRNAs and are involved in many biological processes [125].

1.3.4 RNA binding proteins promote decay through binding to the 3'UTR

Central to the control of mRNA metabolism is the role of RNA binding proteins. RNA binding proteins regulate diverse processes from the control of splicing through the splicesome, to direct involvement in mediating RNA decay by recruitment of deadenylases. A census of 1,542 human RNA binding proteins indicated that, of the different classes of RNA binding proteins studied thus far, the RNA binding proteins that are bound directly to mRNAs are by far the most abundant. Furthermore, RNA binding proteins are more highly expressed than transcription factors across diverse tissue types and they have substantial expression dynamics during tissue development, particularly in neuronal tissue [126].

The most prototypical example of post-transcriptional control of gene expression through RNA binding proteins revolves around the establishment of body segmentation in the developing *Drosophila* embryo. In this process, a gradient of expression of the maternal RNA *hunchback* is established through the action of the Nos and Pum proteins binding to a nos-response element in the 3'UTR of the gene [127, 128]. After binding of the Pum-Nos complex, translation initiation is blocked and the *hunchback* gene is silenced [129]. Disruption of the activity of either the Pum or Nos proteins results in incorrect patterning and development of the embryo [130]. Further work with Pum proteins in both *Drosphila* and *Homo sapiens* has unraveled key aspects of Pum function, such as its recruitment of the Ccr4-Not complex to facilitate deadenylation [131], and the structural determinants of its high sequence specificity [132]. Additionally, the activity of Pumilio proteins have been implicated in a number of developmental functions, such as neurological development and cancer [133]. Recent measurements of the effect of the human members of the Pumilio family of proteins on steady state RNA abundances have identified over 1000 transcripts that change in steady state abundance when the Pumilio proteins are knocked down [134]. Models using a simple count of Pumilio sequence motifs in the 3'UTR of target transcripts fail to account for the full variance of Pum-mediated control of RNA abundances [134]. Several groups have suggested the Pum proteins may interact with the miRNA system due to overlaps in binding sites between the two systems [135, 136] and human Pum proteins have been shown to remodel RNA secondary structure during coregulation with miRNAs for some targets [137]. Key to understanding regulation by Pumilio proteins is the ability to measure RNA decay at a global level. In Chapter 4, I review high throughput sequencing techniques that allow for the measurement of RNA decay at a global level both within and between samples. Additionally, I discuss some of the standard computational methods used to process data from these experiments as well as suggesting new ways to maximize the utility of sparse decay data. In Chapter 5 we use these techniques to measure changes in RNA decay mediated by PUM1 and PUM2, the two human PUM homologues. There, I define sequence determinants of PUM action in human cell culture, revealing a rulebook for the ideal PUM binding site, based on contextual features around predicted binding sites, to facilitate regulation by the protein.

1.4 Interactions between protein and nucleic acid at the center of gene regulation

Throughout this introduction, I have discussed several examples of protein-mediated control of the processes of transcription and translation through interactions with specific sequences of DNA or RNA, respectively. In the coming chapters, I will give more detail on exactly how specific proteins mediate gene regulation in both bacteria and human cells as measured through large high-throughput experiments. Although the details differ, understanding the interactions between proteins and nucleic acids is fundamental to predicting the functioning of biological systems. In Chapter 6, I give my perspective on some of the key findings from the particular systems under study. I also highlight gaps in our knowledge, focusing on the complex role that higher order structure and context play in determining how genes are regulated. Thus, my graduate work represents a conceptual bridge between the work defining the detailed molecular mechanisms of the key components driving biological systems, as presented in this introduction, and the emergent properties resulting from the higher order interactions between said components into which we are just now beginning to gain mechanistic insight.

CHAPTER 2

Escherichia coli Lrp regulates one-third of the genome via direct, cooperative, and indirect routes

2.1 Contribution details

This work was reproduced from its published form, with permission, from Kroner et al. [138]. I am a co-first author on this manuscript and I developed the entire computational pipeline for analyzing the ChIP-seq and RNA-seq data used in this study. In addition, I created my own ChIP-seq peak caller that is described in detail in the methods section and is available freely online on Github (link in the methods section). The figures in this manuscript are a mix of those created by both Grace Kroner and me. The intellectual content and models about Lrp action come from many conversations between Grace, Peter, and me as we strove to make sense of this complicated system. Likewise, the text is also a combination of work between Grace, Peter, and me.

2.2 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-seq under nine physiological conditions to comprehensively 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 to enable combinatorial interactions with other regulators, and second, that nutrient levels induce a global shift in the equilibrium between less sequence-specific and more sequence-specific DNA binding. The overall regulatory behavior of Lrp, which as we now show extends to 38% 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.

2.3 Introduction

Regulation in response to changing nutrient conditions is a vital characteristic for free-living microbes, which must rapidly sense and respond to their environment in order to optimize fitness. The frequently-studied model microbe *Escherichia coli* (*E. coli*) uses a hierarchical regulatory architecture to coordinate responses to environmental changes, with the activity and actions of dozens of specific transcription factors organized by seven global regulators: ArcA, FNR, Fis, CRP, IHF, H-NS and Lrp [49]. *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 [55]. It influences a variety of cellular processes: amino acid synthesis, degradation and transport, porin expression, and pilus formation [53, 54]. The latter represents an example of how Lrp homologues have recently been tied to expression of virulence genes [139–144].

Lrp itself is an 18 kD protein containing a helix-turn-helix DNA binding domain and a regulator of amino acid metabolism (RAM) domain [59]. *In vivo*, it is thought to exist in an equilibrium between octameric and hexadecameric states [145]. Binding of leucine to the RAM domain is known to favor formation of octamers over hexadecamers [56] and to increase the nonspecific DNA binding affinity of Lrp [146]. In addition, the presence of leucine can affect Lrp's regulatory role. Depending on the target, Lrp either activates or represses transcription, and in turn, leucine binding to Lrp either potentiates, inhibits, or has no effect on Lrp function [58]. Recent studies also indicate that Lrp may respond to other amino acids, including alanine, methionine, isoleucine, histidine and threonine [57]. Cho et al. [58] performed chromatin-immunoprecipitation (ChIP) using epitope-tagged Lrp under three conditions, resulting in expansion of the known Lrp regulon to 138 binding-sites. However, based on estimates about the levels of Lrp and the percentage found free of the nucleoid [146], 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.

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, representing 38% of genes in *E. coli* (roughly one-sixth of which are direct targets of Lrp) compared to the 2.3% currently documented in RegulonDB [48]. The fact that many of the newly identified Lrp targets are only apparent under physiological conditions which had not been included in prior studies of the Lrp regulon underscores the importance of considering a wide range of conditions in any survey of transcription factor activity, and also highlights the physiological role of Lrp in balancing foraging and biosynthetic strategies as nutrient conditions change.

We also 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 [147, 148]. 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, σ^{54}, as well as to changes in Lrp binding occupancy that are consistent with changes in Lrp's oligomerization state. 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 [149] argues for its critical role in organismal survival, and here we provide 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.

2.4 Results

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); representative growth curves for each condition are shown in Figure 2.1. 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 (Figure 2.2, Table 2.1). Care must be taken while interpreting these results, as knocking out a global regulator such as Lrp may induce some regulatory re-wiring that is perhaps distantly related to Lrp's normal biological function. For ChIP-seq analysis we are only using the knockout strain samples as a control to filter out peaks resulting from non-specific interactions of the antibody with complexes other than Lrp. For the RNA sequencing results we are looking at transcriptional changes between the WT and *lrp::kanR* strains, and it is possible that some of the changes in transcription may be a result of this regulatory rewiring. Therefore, some false positives are unavoidable. Nevertheless, we find this data to be a valuable resource for exploring the full scope of Lrp regulatory activity. Thus, when we observe, and state, that a given gene is 'regulated' by Lrp, we mean by this that its transcript level changes substantially in the complete absence of Lrp—we find this to be an appropriate definition to reflect the extremely broad impacts of a global regulator such as Lrp on cellular regulation and physiology. For the more specific definition of genes that are regulated directly by Lrp binding to their promoters, we introduce below the concept of a 'direct target' of Lrp, which defines a more narrowly construed Lrp regulon based only on effects in cis at specific target promoters.

Many well-studied Lrp targets are reproduced in our data. For example, IlvI (b0077) is an enzyme critical for valine and isoleucine biosynthesis that is known to be activated by Lrp [150]. Consistent with prior work, we see a strong Lrp binding signal at the *ilvI* transcription start site (Figure 2.2B, top panel), and a Lrp-dependent activation of *ilvI* transcription in several media conditions (Figure 2.2B, bottom 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* [50]. Similarly, we also see strong binding at the promoter of *sdaA* (b1814), a serine deaminase that has been previously shown to be repressed by Lrp in minimal media [58] (Figure 2.2C). Consistent with prior reports, this repression and binding is relieved in the presence of exogenous leucine. Due to the higher resolution of our ChIP-seq data compared to that of prior ChIP-chip studies, we are able to resolve an additional peak in the MIN conditions at the 3' end of the *sdaA* coding region that may play a role in the repression seen exclusively in these conditions. Note that the lack of a unique transcription start site for *pdeD* precluded classification of *pdeD* in our analysis pipeline.

2.4.1 ChIP-seq identifies hundreds of novel Lrp binding sites

While the level of Lrp protein remained fairly stable across the conditions that we tested (Figure 2.3), we observed a ten-fold range (between 61 and 638) 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 (summarized in Figure 2.2A), in agreement with previously published Lrp ChIP data [58] and with Lrp's known role as a regulator which responds to decreasing nutrient levels. However, our data identifies between 1.8-fold (at LIV_Log) and 4.8-fold (at MIN_Stat) more binding sites overall than previous studies [58]. In general, we document more Lrp binding sites at later time points (Trans and

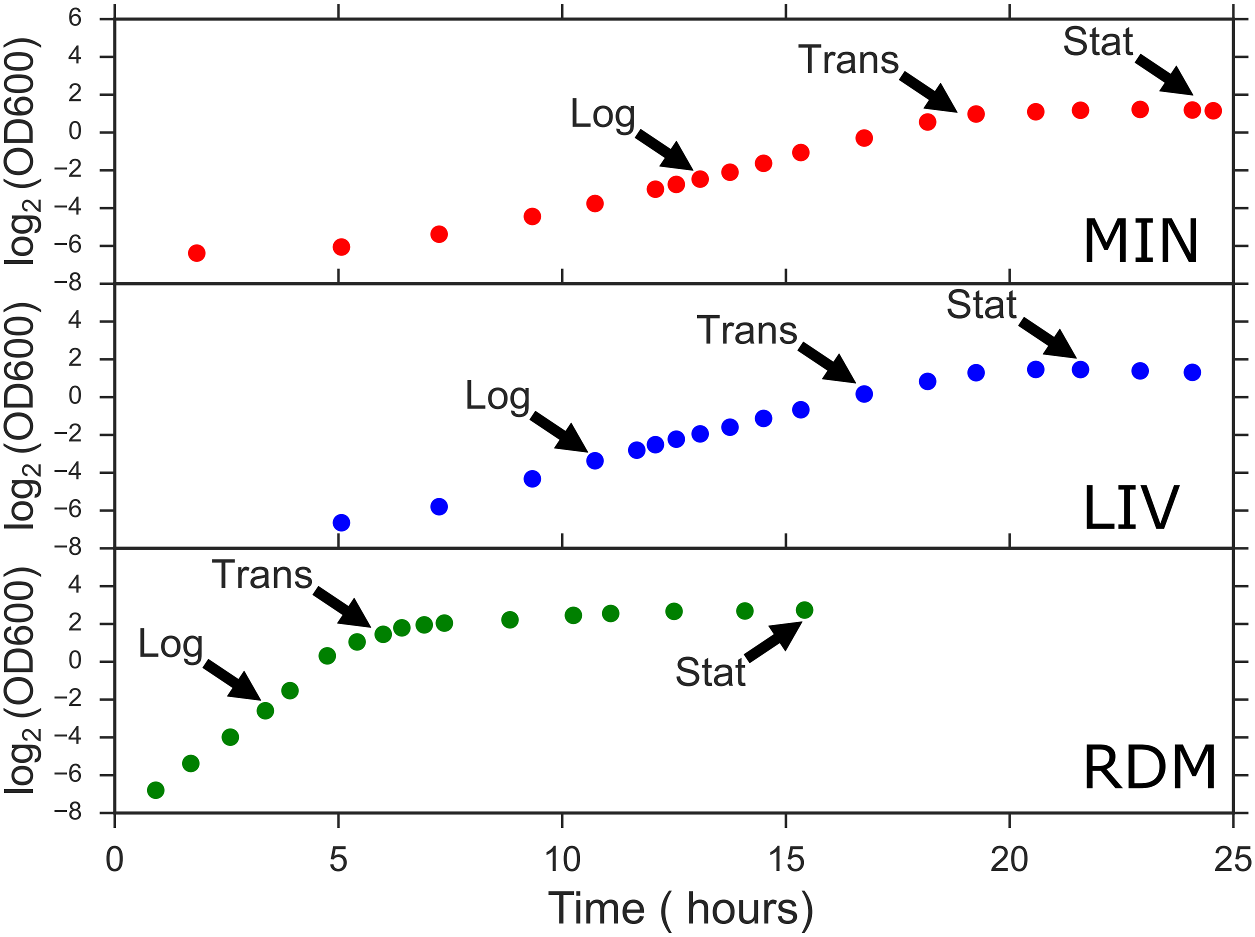

Figure 2.1: Representative growth curves for each condition. Arrows indicate approximate position on the growth curve for each timepoint.

	No. (%) of total genes significantly[a]:	
Condition	Upregulated by Lrp	Downregulated by Lrp
MIN_Log	251 (5.39)	227 (4.87)
MIN_Trans	453 (9.73)	635 (13.63)
MIN_Stat	90 (1.93)	71 (1.52)
LIV_Log	147 (3.16)	258 (5.54)
LIV_Trans	105 (2.25)	138 (2.96)
LIV_Stat	99 (2.13)	71 (1.52)
RDM_Log	41 (0.88)	90 (1.93)
RDM_Trans	58 (1.25)	21 (0.45)
RDM_Stat	728 (15.63)	670 (14.38)

[a] Percentage is out of the total number of genes in *E. coli* (4,658).

Table 2.1: Genes with significant Lrp-dependent changes in expression

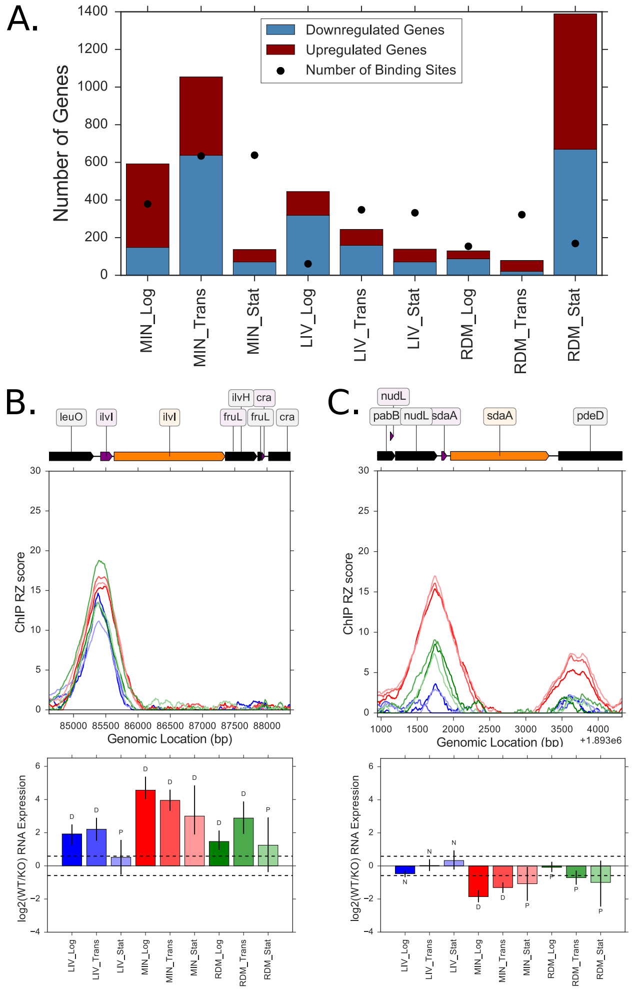

Figure 2.2: ChIP-seq data shows agreement with previous data and reveals novel Lrp binding sites. (A) Total number of high-confidence Lrp binding sites identified in each condition (black circles), and the number of genes upregulated and downregulated by Lrp in each condition based on RNA-seq. (B) ChIP robust Z-score (top) and RNA-seq expression change ($\log_2$(WT/KO); bottom) for known Lrp activated target *ilvI*. Error bars for the RNA-seq data indicate a percentile based 95% confidence interval from 100 bootstrap replicates of expression levels, with conservative pooling of replicate information (see Section 2.6.15 for details). Labels above each bar indicate classification of the gene based on combining RNA-Seq and ChIP-Seq results (D-Direct Lrp target, I-Indirect Lrp target, P-Poised Lrp binding site with no regulatory effect under that condition, N-No Lrp Link; see Figure 2.6A and accompanying text for details of the classification). Dashed lines in RNAseq plots indicate a 1.5 fold cutoff for the ratio between WT and KO strains needed for biological significance (see Section 2.6.15 for details). In the genomic diagram above the plots, open reading frames (ORFs) are shown in black, regulatory regions (as defined in our Methods but without the 250 bp padding) in purple, and the particular gene of interest in orange; we follow this color scheme throughout the text. (C) ChIP robust Z-score (top) and RNA-seq expression change ($\log_2$(WT/KO); bottom) for known Lrp repressed target *sdaA*, panels as in B.

Stat) relative to Log (Figure 2.2A); again in agreement with previously published Lrp ChIP data [58] and with the known role of Lrp as being a critical regulator at the transition to stationary phase. As would naturally be expected, we saw strong enrichment of Lrp binding sites among regulatory regions of the genome (see Section 2.4.8). Comparing our data to previously published ChIP-chip studies [58], 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 compared to a null distribution of randomly shuffled peaks of identical lengths; $p < 0.001$, permutation test, $r = 1000$; here and throughout the chapter we use r to refer to the number of replicates used for resampling tests), 44% at LIV_Log (123.0 fold enrichment compared to a null distribution, $p < 0.001$, permutation test, $r = 1000$) and 84% at MIN_Stat (15.5 fold enrichment compared to a null distribution, $p < 0.001$, permutation test, $r = 1000$). The larger disparity at LIV_Log is likely due to differences in metabolic responses upon addition of leucine alone (as in prior studies) versus supplementation with leucine, isoleucine and valine as in our study. Overall, across the conditions in our study, we identified over 730 novel Lrp binding sites, 198 only occurring under conditions that had not previously been tested and 532 occurring under conditions similar to those tested previously, highlighting both the enhanced resolution given by modern sequencing techniques [151] and the general need to explore a broad range of condition space to understand the complete regulon of any transcriptional regulator (Binding sites under each condition in the regulatory regions of genes are enumerated in Supplementary Data File 1; data on all identified binding sites can be accessed at GEO Dataset GSE11874). In fact, our newly identified binding sites found in conditions that have been previously studied with ChIP-chip [58] have, on average, lower ChIP-signal at their peak summits than the peaks that overlap with the peaks found in the previous ChIP-chip study (Figure 2.4). However, the newly identified direct Lrp targets in this study have a similar distribution of magnitude $\log_2$ fold change in RNA expression as genes that were previously annotated as Lrp targets in RegulonDB (Figure 2.5), showing far more overlap with the effect sizes shown for previously annotated targets than did the ChIP signals in Figure 2.4. Taken together with our stringent data analysis pipeline, these data suggest that the novel sites we are identifying in this study represent functional Lrp binding sites that are revealed by our more sensitive experimental methods.

2.4.2 Lrp's regulatory effects are broad and highly condition specific

Through our use of parallel RNA-seq experiments in WT and *lrp::kanR* cells, we were able to identify the full range of transcripts showing Lrp-dependent regulation across the conditions in our study. Based on our RNA-seq data, we find that the number of genes regulated by Lrp varies across conditions from 1.7% to 30.0% of all known *E. coli* genes (Figure 2.2A, Table 2.1); in all, 2,459

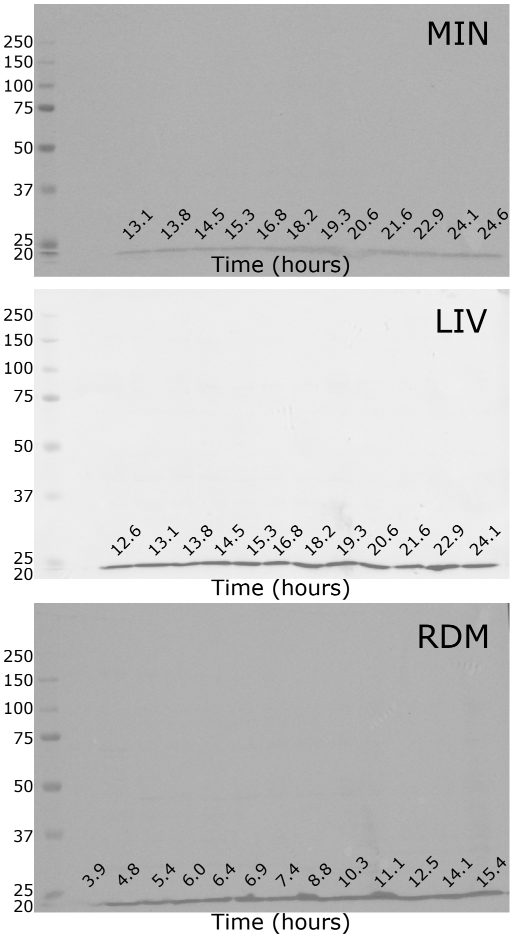

Figure 2.3: Lrp levels are consistent within each condition. Western blots of whole cell lysate harvested at indicated times across the experimental time course (times are given in hours; compare with Figure 2.1 for representative growth curves covering the same time range). Lysate was quantified by Bradford assay, and 15 μg of total protein was loaded for the MIN and LIV time courses (except for samples at 13.1 ($\sim$7.9 μg), 13.8 ($\sim$6.3 μg), 15.3 ($\sim$13.1 μg), and 18.2 hours ($\sim$5.4 μg) in the MIN time course which were too dilute to reach 15 μg); 20 μg of total protein was loaded for the RDM time course (except for the samples at 3.9 ($\sim$2.3 μg) and 4.8 hours ($\sim$13.2 μg) which were too dilute to reach 20 μg). Molecular weight ladder markers are indicated in the leftmost lane.

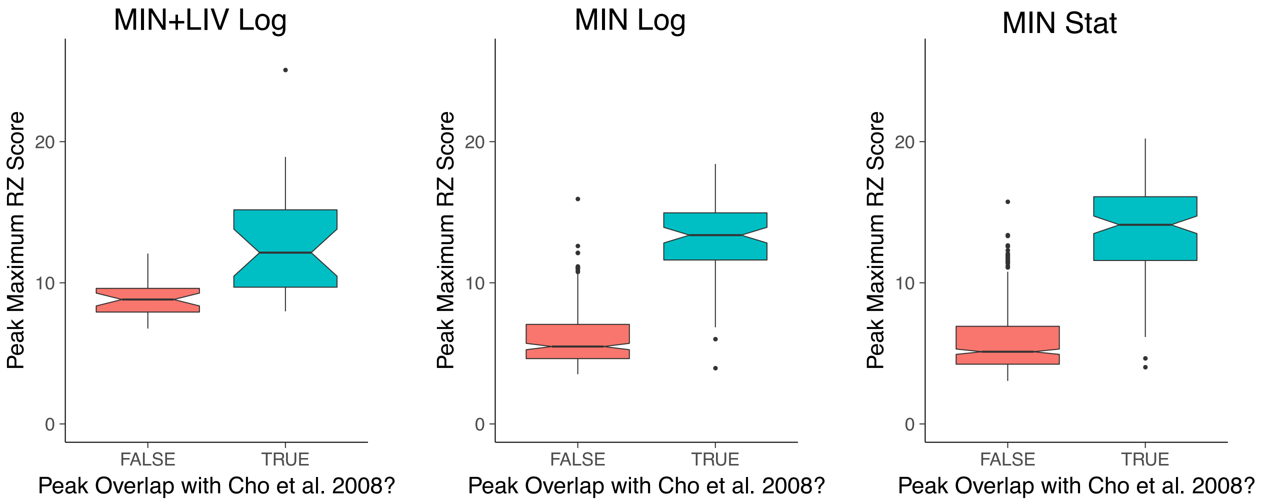

Figure 2.4: Newly identified Lrp peaks have a lower overall ChIP-seq signal than those previously identified by ChIP-chip. Box plots representing the distribution of the peak summit RZ score for peaks identified in this study. The data is split into novel peaks identified only in this study and peaks identified in this study that overlap with previous ChIP-chip data from Cho et al. [58]. Only the conditions that were similar between this study and the Cho et al. [58] study are shown.

genes (52.8% of all genes in *E. coli*) show Lrp-dependent changes in transcript levels under at least one condition (see Section 2.6.15 for details). Lrp-dependent RNA expression changes for each categorized gene can be found in Supplementary Data File 1. Overall RNA expression changes for all annotated genes in both WT and *lrp::kanR* cells can be accessed at GEO dataset GSE11874.

Comparing all genes with a Lrp-dependent change in expression in our RNA-seq data to genes previously identified as Lrp targets, our data set overlaps with 73% of the known targets in RegulonDB (1.38 fold enrichment compared to a null distribution of randomly shuffled gene names, $p < 0.001$, permutation test, $r = 1000$), 81% of the previously identified ChIP-chip targets (1.53 fold enrichment compared to a null distribution, $p < 0.001$, permutation test, $r = 1000$) (15), and 89% of the previously identified microarray targets (1.68 fold enrichment compared to a null distribution, $p < 0.001$, permutation test, $r = 1000$) [55], showing good agreement across the variety of strains and media conditions present in the compared studies, despite some variations in precise experimental conditions (it is important to note that the large fraction of the genome that is regulated by Lrp imposes a fairly low upper limit on the amount of enrichment possible when comparing with prior lists of targets). Our data also reveals 2,241 genes with previously undocumented Lrp-dependent expression.

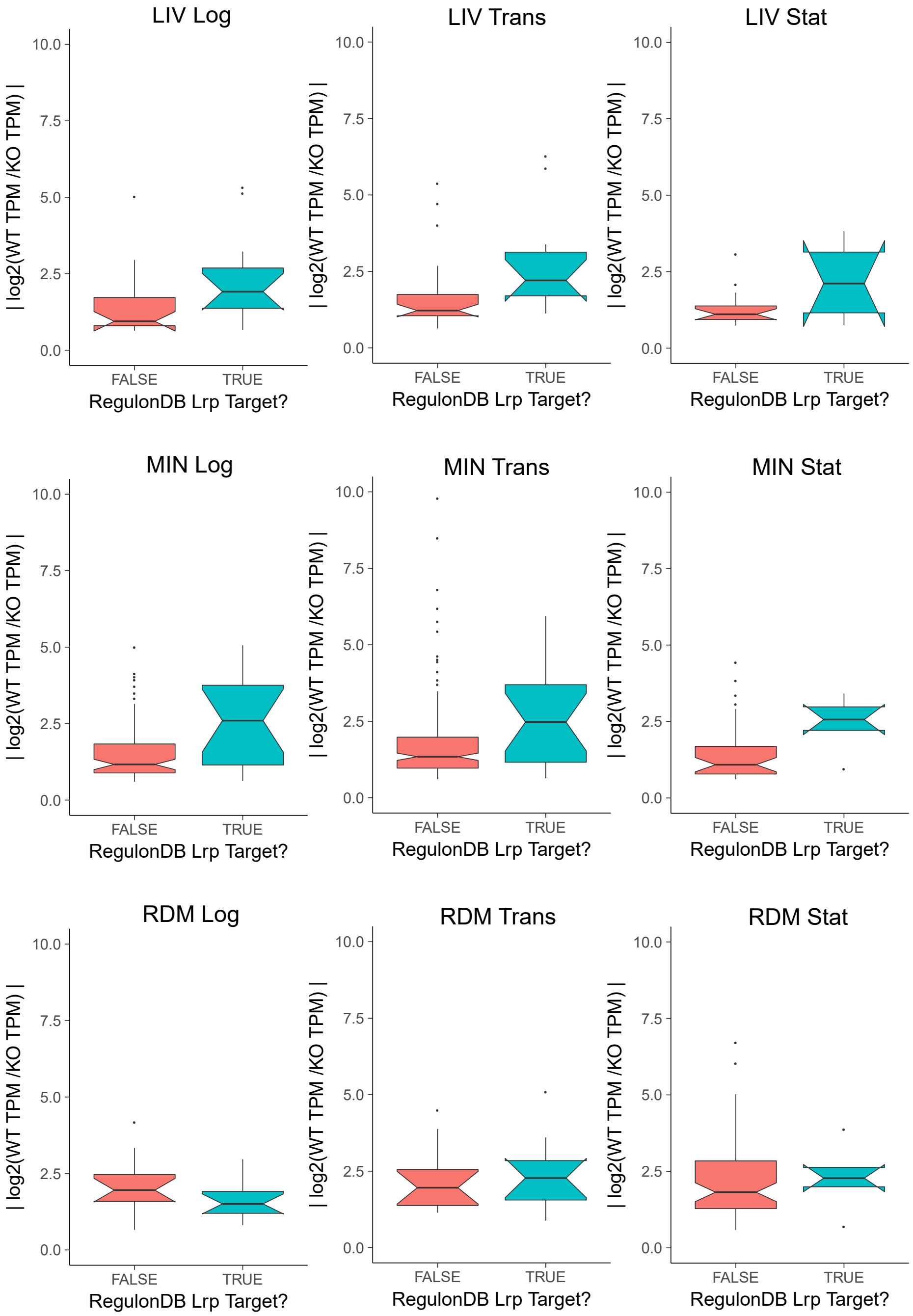

Figure 2.5: Newly identified Lrp direct targets have a similar magnitude change in RNA expression as known Lrp targets. Box plots representing the distribution of magnitude $\log_2$ fold change in RNA expression as compared between WT and *lrp::kanR* strains and calculated as described in Section 2.6.7. The data is split into newly identified direct Lrp targets and Lrp targets that were previously identified as indicated by RegulonDB. All nine conditions are shown here.

2.4.3 The majority of Lrp-dependent regulation occurs via indirect effects

Global regulators are known to act both directly, by binding target sites and modulating transcription levels, and indirectly, by modulating the expression of other transcription factors or regulatory RNAs which have their own targets [49]. 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 (see Section 2.6.17). Direct and indirect targets are both characterized by Lrp-dependent changes in transcript level, but only direct targets have a Lrp binding signal in their regulatory regions, defined as 250 bp upstream and downstream of all annotated transcription start sites in RegulonDB (Figure 2.6A; annotations from RegulonDB; see Section 2.6.17 for details).

In order to allow cross-referencing of our binding and expression data, we restrict our analysis here to the set of genes for which an annotated transcription start site exists in the PromoterSet dataset in RegulonDB version 9.4 [48], or an unannotated transcription start site exists in the same direction within 500 base pairs upstream of the start of the coding region (covering 2908 genes out of the possible total of 4658). Typically, this categorizable subset includes the first gene in each transcriptional unit, as well as any genes with additional internal promoters. A heatmap of classifications for all categorizable genes across conditions is shown in Figure 2.6B. Additionally, the total number of direct and indirect targets which were identified in previous studies are tabulated in Table 2.3.

From our analysis of that categorizable subset, we note that 37.8% of all *E. coli* genes are regulated by Lrp, either directly or indirectly, in at least one condition. Out of those, about 10% are only ever regulated directly across the experimental conditions, 84% are only ever regulated indirectly, and 6% are regulated directly and indirectly in different conditions. Due to the restriction on categorizing genes noted above, the counts given here are an underestimate. Even so, we also observe a dramatic increase in the number of indirect targets at MIN_Trans and RDM_Stat, going from 237 to 610 indirect targets between MIN_Log and MIN_Trans, and from 36 to 985 indirect targets from RDM_Trans to RDM_Stat.

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 (Figure 2.6B), 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 its indirect targets at each condition were annotated targets of transcriptional regulators categorized as direct Lrp targets under that condition (all transcription factor-gene interactions were taken from RegulonDB [48]; see Section 2.6.17 for details). 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 ($q < 0.05$, permutation test), albeit small, enrichment of explainable indirect targets at LIV_Log and RDM_Stat; a maximum of 5.2% of indirect targets can be explained by the currently known targets of direct Lrp targets (Table 2.2). Several key transcription factors that are direct Lrp targets are responsible for explaining the identified indirect Lrp targets across conditions: Nac, LrhA, LeuO, ArgR, QseB, CysB, SlyA, SoxS, and GadW (Table 2.2). Many of these transcription factors have been previously identified as Lrp targets [152]. 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 (recent large-scale studies based on current annotations of the *E. coli* transcriptional regulatory network demonstrate that our current enumeration of regulatory interactions is incomplete [70]). In addition, regulatory RNAs that are direct Lrp targets are also likely important in mediating indirect regulation. However, based on current RegulonDB annotations, we are unable to account for any indirect targets in this manner. Since our classification scheme allows for some genes within operons to be separately annotated from transcription of the full operon due to existing internal TSSs, it is possible that some indirect targets could be accounted for by direct Lrp regulation of the gene's parent operon. Additionally, it is also possible that Lrp may be mediating transcription by binding within a particular gene's coding region, rather than its TSS. We therefore subclassified our indirect targets into these possibilities and noted that both cases represent a small fraction of indirect targets in all conditions (Figure 2.7B). Furthermore, Lrp binding within the coding region of genes does not appear to meaningfully impact raw RNA-Seq coverage as would be expected if Lrp was interfering with transcription by binding within the coding region of genes (Figure 2.7C-E).

Investigating at a local as opposed to global scale provides several informative examples of indirect regulation by Lrp. At LIV_Log, LIV_Trans and RDM_Log, the dual regulator LrhA is a direct Lrp-activated target gene (Figure 2.6C). LrhA represses *flhC* and *flhD* (Figure 2.6D). At LIV_Trans, *flhC* is indirectly repressed, and at RDM_Log, both *flhC* and *flhD* are indirectly repressed (Figure 2.6C). While this pattern does not show activity at all LrhA targets in each condition (for example, *fimE* is known to be activated by LrhA in some conditions, but does not show a Lrp-dependent response under conditions tested here), overall it suggests that indirect regulation of *flhCD* by Lrp may be explained in some cases by direct LrhA activation by Lrp. All three target genes (*fimE*, *flhC*, and *flhD*) 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 2.6E). CysB is known to activate *tcyP* and *cysI*, among other genes (Figure 2.6F). Both *tcyP* and *cysI* were categorized as indirect Lrp-repressed targets, supporting the hypothesis that Lrp repression of *cysB* is what leads to repression

	Total	Number (Percentage)	Number of direct			
	indirect	of indirect targets	targets implicated in			
	targets	explained by direct targets	indirect regulation	Fold-enrichment[a]	p-value	q-value
MIN_Log	237	6 (2.53%)	2 (LeuO, GadW)	2.73	0.121	0.218
MIN_Trans	610	29 (4.75%)	6 (Nac, ArgR, CysB, SlyA, SoxS, GadW)	1.20	0.086	0.193
MIN_Stat	76	2 (2.63%)	1 (ArgR)	1.56	0.621	0.799
LIV_Log	252	13 (5.16%)	2 (LeuO, GadW)	4.84	0.001	0.009
LIV_Trans	121	1 (0.83%)	1 (LrhA)	6.01	0.206	0.309
LIV_Stat	104	0 (0%)	0	0	1	1
RDM_Log	56	2 (3.57%)	1 (LrhA)	12.98	0.026	0.078
RDM_Trans	36	0 (0%)	0	NA	1	1
RDM_Stat	985	12 (1.22%)	3 (QseB, GadW, SlyA)	2.08	0.002	0.009

[a] 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.

Table 2.2: 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.

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 classified 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. However, it is important to note that due to the interconnected regulatory network of *E. coli*, compensatory and interconnecting mechanisms likely also contribute to the regulation of these targets; Lrp is unlikely to be the sole regulator responsible for the observed behavior, and other Lrp-dependent pathways may also act in parallel with those suggested.

2.4.4 The majority of Lrp binding reflects poised, rather than active, regulatory sites

In addition to the indirect regulation discussed above, 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 (Figure 2.6A). In fact, these sites comprise as little as 53% (at LIV_Log) to as much as 92% (at MIN_Stat, LIV_Stat, and RDM_Trans of all instances of Lrp binding (Figure 2.7A). We refer to such cases as poised targets of Lrp, since they suggest that Lrp is bound in preparation for regulatory activity under changed conditions. The regulatory potential of the identified poised sites is apparent from the fact that across the set of nine experimental conditions in our study, 40% of genes that are poised targets in at least one condition become direct targets in a different condition, and conversely, 93% of direct Lrp targets are poised targets under at least one other condition (several such cases are discussed in the following section; for further dis-

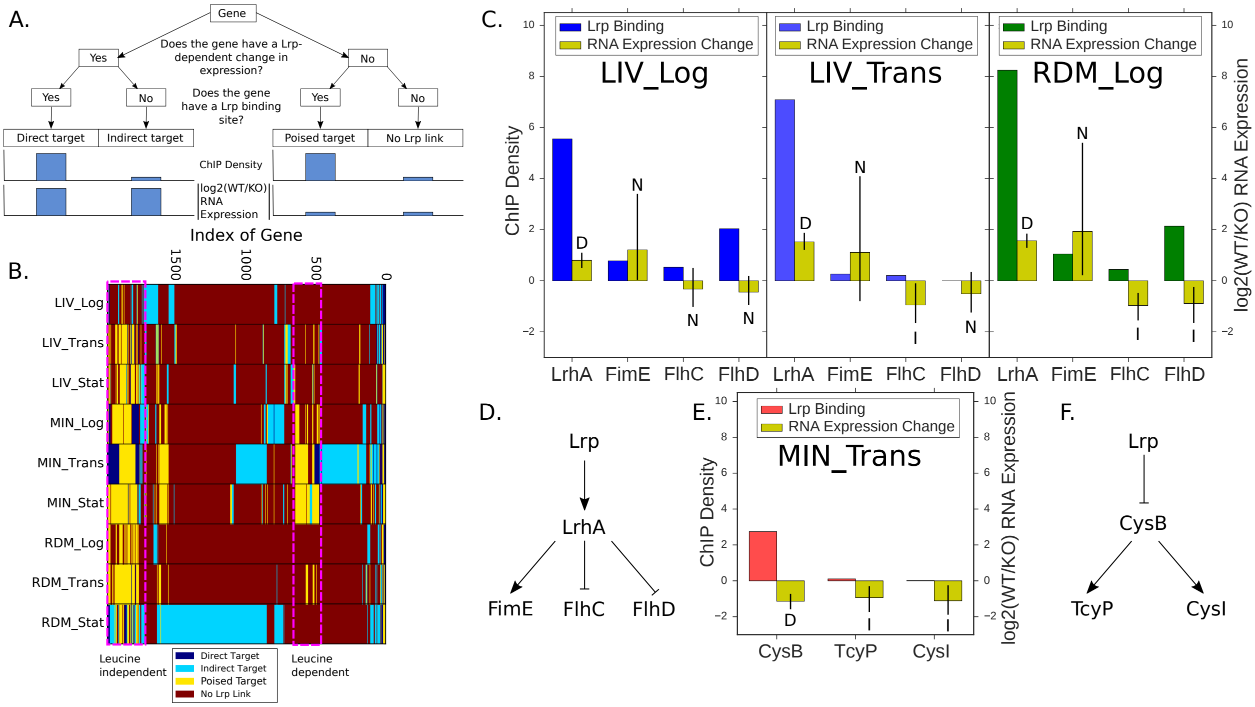

Figure 2.6: Lrp regulates genes both directly and indirectly. (A) Schematic showing how genes were categorized: direct targets of Lrp (Lrp-bound regulatory region and with a significant RNA expression change between WT and *lrp::kanR* cells), indirect targets (not bound but with a significant RNA expression change), poised 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) 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. Pink boxes mark out notable clusters of genes: those with leucine-dependent or -independent binding. (C) ChIP density and RNA-seq expression change ($\log_2$(WT/KO)) for direct Lrp target LrhA and its known target genes, FimE, FlhC and FlhD (56). Error bars for the RNA-seq data indicate a percentile based 95% confidence interval from 100 bootstrap replicates of expression levels. Labels above each bar indicate classification of the gene based on combining RNA-Seq and ChIP-Seq results (D-Direct Lrp target, I-Indirect Lrp target, N-No Lrp Link) (D) Proposed model of Lrp/LrhA mediated regulation of LrhA targets. (E) ChIP density and RNA-seq expression change ($\log_2$(WT/KO)) for direct Lrp target CysB and some of its known target genes, TcyP and CysI (56), as in C. (F) Proposed model of Lrp/CysB mediated regulation of CysB targets. In both (C) and (E), only conditions where LrhA or CysB was a direct target are shown.

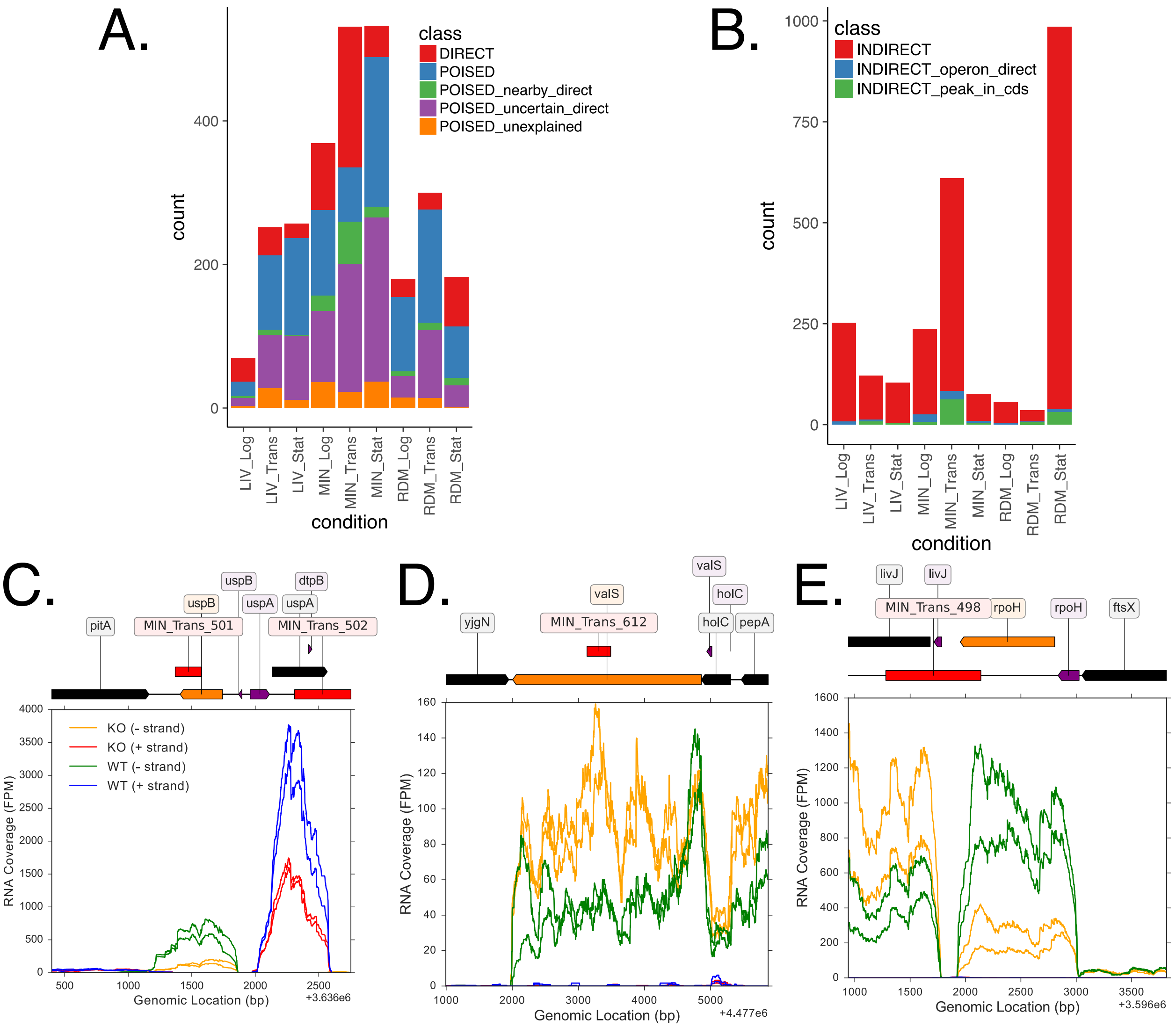

Figure 2.7: Subclassification of Poised and Indirect targets. (A) Subclassification of poised Lrp targets into additional categories per condition. Direct is the same classification as Figure 2.6A. Poised are genes that transition to direct targets in at least one condition. Poised_nearby_direct are genes that were classified as poised but have a gene within 1000 bp that was classified as direct in the same condition. Poised_uncertain_direct represent genes that we do not have strong enough evidence from our RNA-seq data to decisively conclude that they are not transcriptionally regulated by Lrp in that condition. Poised_unexplained represent poised genes in which the conservative 95% confidence interval for the RNA-seq data falls entirely within the region of practical equivalence. (B) Subclassification of indirect Lrp targets into additional categories per condition. Indirect_operon_direct represent indirect targets that are part of an operon that is a direct target in the same condition. Indirect_peak_in_cds represent targets that have a Lrp peak overlapping their coding region but not their regulatory region. Indirects represent all indirects that do not fall into either of the other two categories. For both subclassifications more details can be found in the methods. (C-E) Selected examples of genes subcategorized as indirect_peak_in_cds. Here, RNA coverage in MIN_Trans is plotted as Fragments per Million (FPM) for each replicate of both WT and *lrp::kanR* strains. Peaks in MIN_Trans are represented above each plot as a red bar and labeled as MIN_Trans_#, where the # represents the peak number in the associated GEO narrowPeak files.

cussion on the classification of poised targets, including analysis of the ∼60% of poised targets that do not become direct targets, see Section 2.4.9). Among genes that undergo a transition between being a poised target and a direct target, 37.8% become activated, 45.9% become repressed, and 16.2% become both activated and repressed in different conditions. The gene *brnQ*, for example, shows a strong Lrp binding site in its regulatory region under all conditions that we studied, but is only Lrp-repressed under a subset of those conditions, and under other conditions is clearly unaffected by Lrp (Figure 2.8A). On the other hand, consistently poised binding is apparent at the *mog* gene; its promoter region is bound by Lrp in 7 of the 9 conditions we studied, but never exhibits a significant change in expression, thus making it a poised target in all of those conditions (Figure 2.8B). Interestingly, *mog* plays a role in the synthesis of molybdenum-containing cofactors [153] and is non-essential in the conditions we are studying here [48]. The consistent binding of *mog*'s promoter by Lrp suggests that Lrp is poised to regulate *mog*, and that it may be a direct target of Lrp under conditions not tested here (perhaps those involving changing molybdenum concentration, as all of our conditions supply abundant molybdenum as part of the MOPS micronutrient mixture [154]). In our consideration of poised targets, it is important to note that some fraction of targets that we assign as poised may actually represent direct targets that our differential expression analysis lacks the power to detect, although this scenario likely accounts for only a minority of cases (see Section 2.4.9). At a system-wide level, it is particularly apparent from the highlighted blocks of leucine-dependent and leucine-independent binding sites in Figure 2.6B, that many genes are bound by Lrp under a far broader range of conditions than the set under which they are regulated by Lrp (or at least by Lrp alone). These findings suggest more broadly that Lrp is often poised at a particular gene under many conditions, but may act combinatorially with some other factor or environmental stimulus in order to actually alter expression. The total number of poised genes which overlap with those identified in previous studies are enumerated in Table 2.3.

2.4.5 Poised Lrp targets enable condition-specific combinatorial regulation

The abundance of genes that shift between direct and poised Lrp targets across conditions 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, or acting completely redundantly with a second factor to repress or activate transcription from a particular target. That second regulator could in principle be a sigma factor, another classical transcription factor, or even a nearby condition-dependent Lrp binding site; we in fact observe examples of all three such scenarios in our data.

Lrp binding at the *potF* promoter region represents a case of additional nearby Lrp binding being associated with conversion of a poised to a direct target. A strong Lrp binding signal is seen

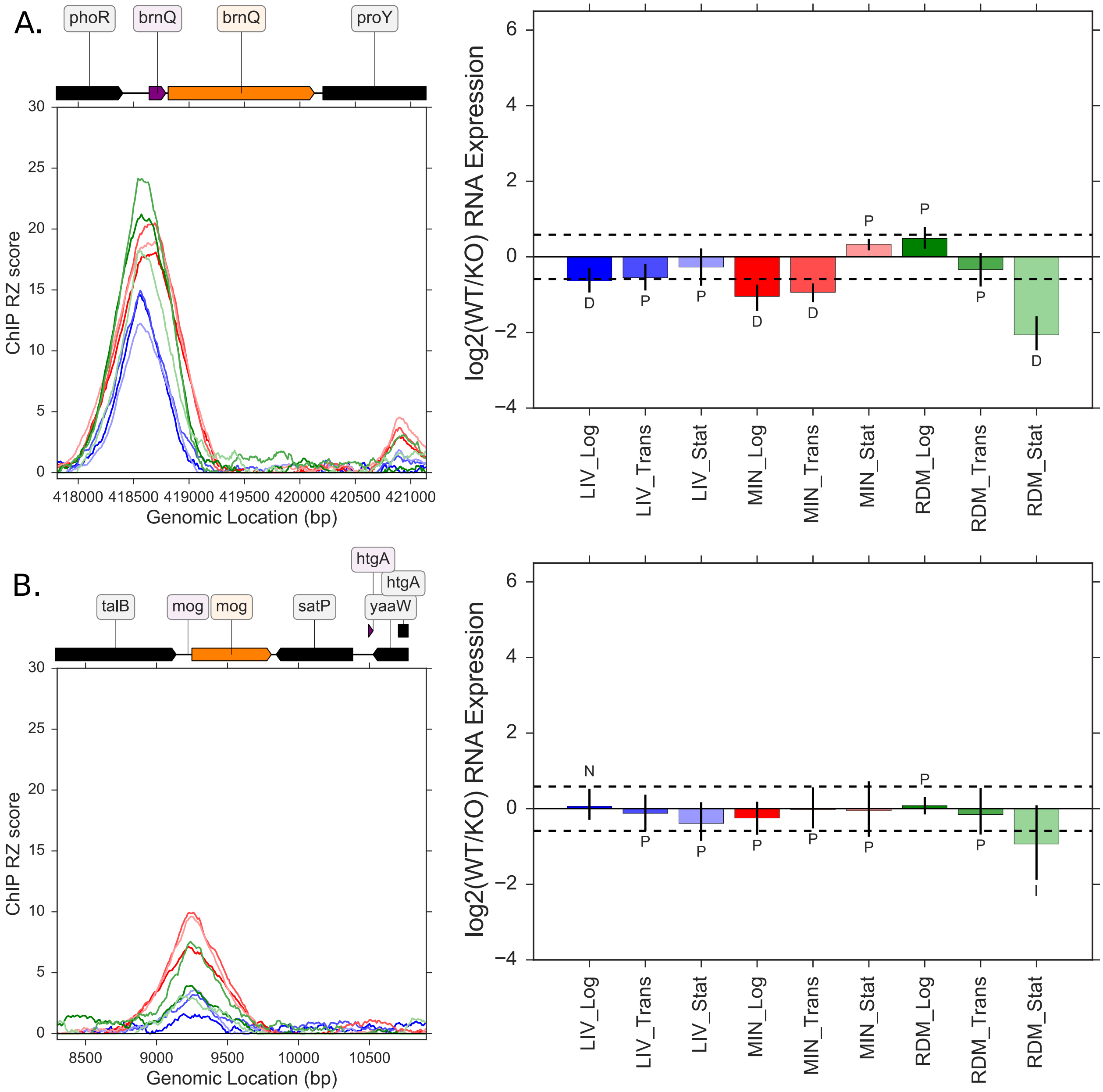

Figure 2.8: Poised targets show condition dependent Lrp regulation separate from their Lrp binding profiles.(A) ChIP robust Z-scores (left) and Lrp dependent RNA-seq expression change ($\log_2$(WT/KO); right) for Lrp poised target *brnQ*, which shows condition invariant binding but is clearly Lrp regulated only under some conditions (MIN_Log, MIN_Trans, RDM_Stat), whereas in other conditions it can be confidently said to have no substantial Lrp-dependent effect (LIV_Stat, MIN_Stat). Coloring as in Figure 2.2B. (B) As in panel A, but for the *mog* gene, which shows constitutive Lrp binding but no direct Lrp-dependent regulation under the conditions studied.

	LIV_Log	LIV_Trans	LIV_Stat	MIN_Log	MIN_Trans	MIN_Stat	RDM_Log	RDM_Trans	RDM_Stat
Total Direct Targets	33	39	20	93	196	43	25	23	69
Total Indirect Targets	252	121	104	237	610	76	56	36	985
Total Poised Targets	37	213	237	276	335	489	155	277	114
RegulonDB Overlaps (103 annotated targets)									
Direct	12	11	5	16	18	6	12	11	5
Indirect	9	4	3	12	12	5	2	2	16
Poised	4	8	18	9	11	23	9	13	17
No Link	34	36	33	22	18	25	36	33	21
Microarray overlaps (53 annotated targets)									
Direct	3	2	0	4	7	4	2	1	3
Indirect	29	5	2	19	9	3	3	0	37
Poised	0	3	5	2	2	4	2	4	1
No Link	13	35	38	20	27	34	38	40	4
ChIP-chip overlaps (185 annotated targets)									
Direct	20	22	9	49	48	22	20	16	34
Indirect	13	9	3	22	17	4	5	2	22
Poised	14	40	62	30	33	59	53	62	41
No Link	72	48	45	18	21	34	41	39	22

Table 2.3: Total classified genes overlapping with previous studies

directly at the *potF* promoter in all nine conditions measured in our data, but *potF* expression is only activated by Lrp in six of the conditions that we studied (Figure 2.9A). In contrast with the variable Lrp-dependent RNA expression levels, Lrp binding directly at the *potF* promoter is very similar across conditions, spanning a similar length of DNA, and showing maximal signal at the same point. However, an adjacent upstream Lrp peak at the *ybjN* promoter shows nearly monotonically increasing occupancy with the strength of Lrp-dependent regulation. Interestingly, this secondary peak does not appear to modulate expression of the *ybjN* gene in any of the conditions in our study suggesting that the primary role of this secondary peak under these conditions is in the modulation of *potF*. This secondary condition-specific binding site may represent an interaction between a weak and strong Lrp binding site or it may represent the formation of a hexadecamer through the bridging of these two sites (Figure 2.9B). Future studies will be needed to differentiate these possibilities. Additional examples of secondary peaks appearing under the condition where we can detect Lrp-dependent regulation can be seen clearly in *sdaA* (Figure 2.2C), *lrhA* and *alaA* (Figure 2.10A), and *dadA* and *ycgB* (Figure 2.10B).

In contrast to *potF*, *pepD* has relatively invariant Lrp binding signal at its promoter across each condition (Figure 2.9C). Although a small secondary peak can be seen in each MIN condition, only the MIN_Trans condition shows direct Lrp regulation. Additionally, the RDM_Stat condition does not show this secondary peak but is similarly Lrp-repressed, suggesting that the secondary peak is not sufficient to explain Lrp regulation at this locus. Thus, *pepD*, likely represents a case where Lrp is interacting with an additional factor; for example, the activity at *pepD* could be explained by Lrp's presence blocking a transcriptional activator from binding (Figure 2.9D), although other scenarios are also possible. An additional example of invariant Lrp binding with differential RNA regulation can be seen at *ilvI* (Figure 2.2A). As detailed in Section 2.4.10, systematic analysis of all genes with Lrp bound at their regulatory regions reveals that both the transcription factor NtrC and σ^{54} explain a subset of these transitions from poised to direct targets, but additional factors also likely interact with Lrp in similar ways.

2.4.6 Lrp directs distinct survival strategies across changing nutrient conditions

By applying iPAGE [155] to search for gene ontology (GO) terms that that show significant patterns in Lrp binding and regulation across conditions, we identified several key patterns in Lrp's regulatory logic (summarized in Figure 2.11A and detailed in Figure 2.12). Consistent with its previously established physiological roles, and the fitness effects of *lrp* loss of function mutations (e.g., [152, 156]), the most prominent pathways regulated by Lrp involve the synthesis and uptake of amino acids, as well as nutrient foraging (via regulation of flagellar motility). In particular, Lrp tends to directly activate amino acid biosynthetic pathways during logarithmic and transition phase growth, particularly in nutrient poor media, and at the same time, to directly repress amino acid uptake pathways under the same conditions, presumably responding to a lack of available substrates in the environment (Figure 2.12B). Regulation of leucine transport itself represents a special case, where Lrp appears to activate a subset of leucine transporters and repress others (Figure 2.11B). A similar switch is apparent in Lrp's regulation of flagellar motility in rich media, where Lrp acts as a global repressor of motility in log phase (presumably keeping cells static in conditions of optimal nutrition) but lifts repression and activates a small set of flagellar genes when nutrients are depleted during stationary phase (Figure 2.11C). Thus, Lrp directly governs a shift between strategies of synthesizing or foraging for critical nutrients, depending on their availability in the cell's surroundings. A very different Lrp-dependent regulatory program is apparent under conditions of slowing growth (typified by our MIN_Trans and RDM_Stat conditions), where Lrp additionally acts to inhibit translation, through indirect repression of ribosomal components and tRNA synthetases (Figure 2.12B).

It is intriguing to note that several of Lrp's pathway-level activities, such as the aforementioned switching between amino acid biosynthesis vs. transport, do not appear to depend solely on the presence of leucine, as similar regulatory behaviors are observed in both our MIN and LIV conditions. Indeed, the same behavior is also suggested by the leucine-independent cluster of Lrp targets noted in Figure 2.6B. To assess if there is any class of genes that Lrp binds in a Leucine-dependent manner, we identified GO terms showing informative patterns of leucine-dependent occupancy at their promoters (Figure 2.11D). Aside from being surprisingly short, this list is especially notable for the fact that Lrp binds genes associated with several GO terms involved in amino acid

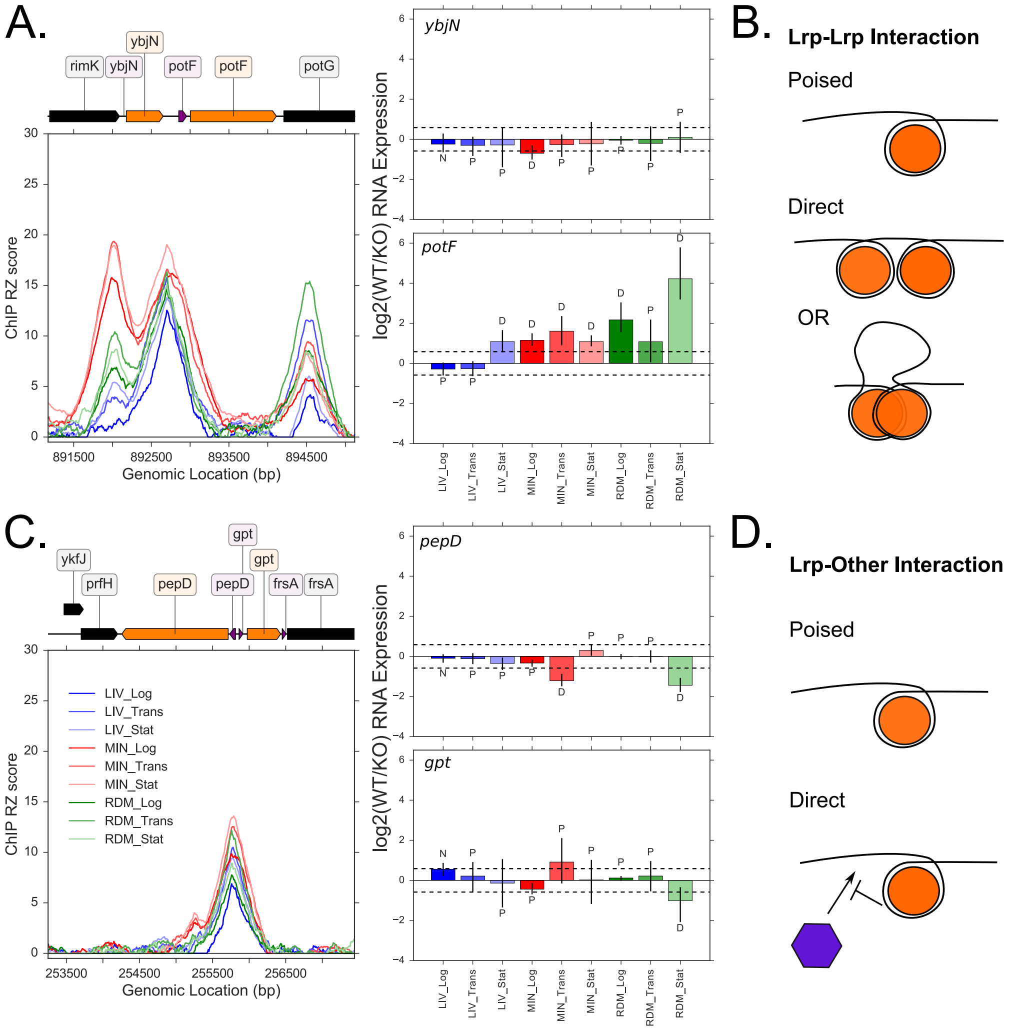

Figure 2.9: Lrp sits at genes in poised position in preparation for regulatory activity. ChIP robust Z-score (left) and RNA-seq expression change ($\log_2$(WT/KO); middle) for two Lrp targets. (A) *potF* represents a case where a secondary Lrp peak is seen only in the conditions where it is a direct target. *ybjN* is clearly not transcriptionally regulated by Lrp in these conditions. Here and in panel C, error bars for the RNA-seq data indicate worst-case percentile based 95% confidence interval from 100 bootstrap replicates of expression estimates across different biological replicates (see Section 2.6.15 for details). Labels above each bar indicates classification of the gene based on combining RNA-Seq and ChIP-Seq results. (D-Direct, I-Indirect, P-Poised, N-No Lrp Link, see Figure 2.6A for details). Dashed lines in RNA-Seq plots indicate a 1.5 fold cutoff for the ratio between WT and KO strains needed for biological significance (see Section 2.6.15 for details). (B) Model suggested by panel A, in which Lrp primarily interacts at the promoter site in most conditions (poised) but upon changes in condition, Lrp binds adjacent sites as a separate octamer (direct top) or together as a hexadecamer through potential looping of the DNA (direct bottom). (C) *pepD* represents a case where no obvious changes in Lrp binding signal occur, but differences in transcriptional regulation are clear. (D) Model suggested by the data in panel C, where Lrp is always bound at the site (poised) but upon conditions where a secondary factor is needed for expression, Lrp is present to interact with the factor (direct) and block or enhance its activity. Here, *gpt* displays similar RNA expression patterns to *pepD*.

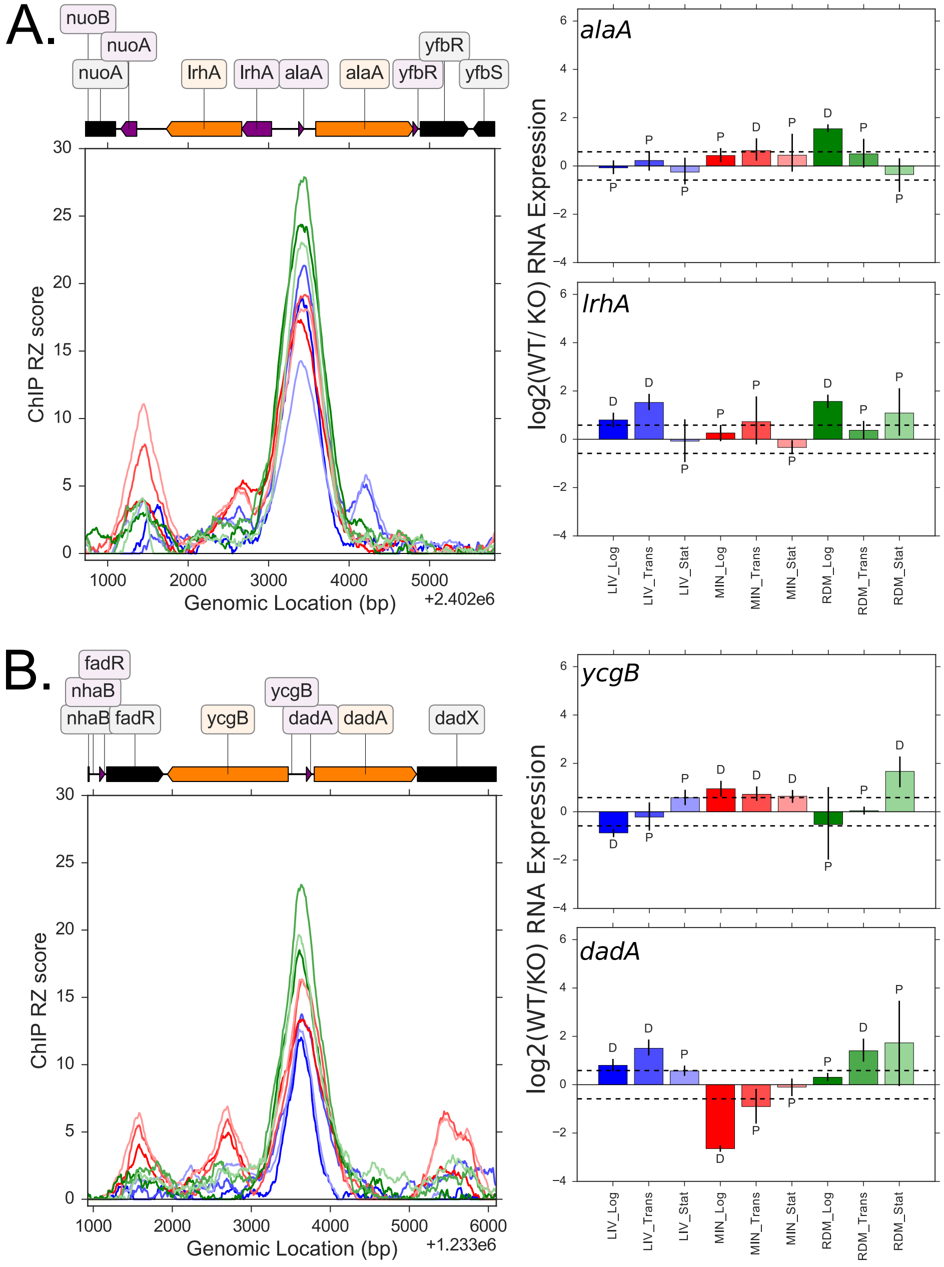

Figure 2.10: Additional examples of poised targets showing evidence for secondary Lrp binding sites with regulatory activity. ChIP robust Z-score (left) and RNA-seq expression change ($\log_2$(WT/KO); right) for several Lrp targets. (A) *alaA* and *lrhA* represent cases where a secondary peak appears under conditions where the gene is a direct target. (B) *ycgB* and *dadA* represent cases where secondary Lrp peaks appear under conditions for which transcriptional regulation by Lrp is strongest, with opposite effects for each gene. Error bars for the RNA-seq data indicate worst-case percentile based 95% confidence intervals from 100 bootstrap replicates of expression estimates across different biological replicates (see Section 2.6.15 for details). Labels above each bar indicate classification of the gene based on combining RNA-Seq and ChIP-Seq results. (D-Direct, I-Indirect, P-Poised, N-No Lrp Link, see Figure 2.6A for details). Dashed lines in RNA-seq plots indicate a 1.5 fold cutoff for the ratio between WT and KO strains needed for biological significance (see 2.6.15 for details)

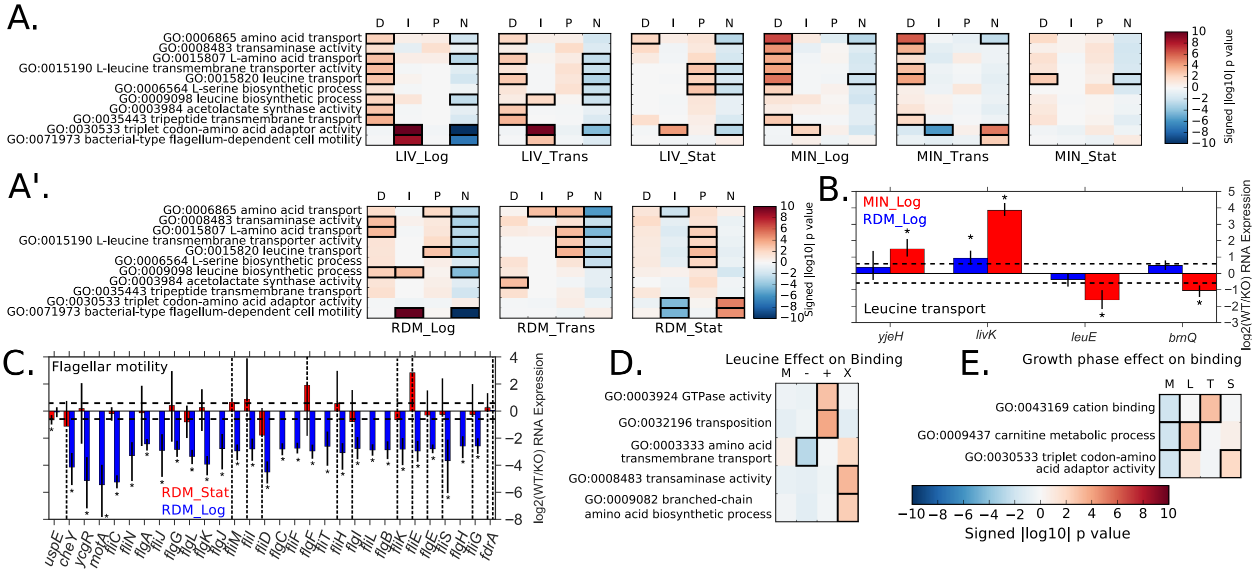

Figure 2.11: Pathway analysis of genes regulated and bound by Lrp. (A) Pathway analysis using iPAGE identifying GO terms that show significant mutual information with gene classification (Direct, Indirect, Poised, or No Lrp link). Color indicates magnitude of the log_{10} p value, with positive values indicating enrichment and negative values indicating depletion of members of a given GO term among genes in that class. Boxes indicate particularly outstanding cells ($p < 0.01$). (B) Comparison of Lrp-dependent effects on gene expression for genes annotated with GO:0015820 (leucine transport). Stars indicate significant Lrp-dependent expression changes (following our standard criteria), with error bars indicating bootstrap-based 95% confidence intervals. (C) As in panel B, for genes annotated with GO:0071973 (bacterial-type flagellum-dependent cell motility). Error bars that pass to infinity under our bootstrap-based 95% confidence intervals are indicated with dashed lines. Bars where WT/KO ratio could not be determined are not plotted. (D) iPAGE plots (as in panel A) showing genes with significant leucine dependents of nearby Lrp binding sites. Categories are: -: Lrp binds target genes only in low leucine conditions (MIN); +: Lrp binds target genes only in high-leucine conditions (RDM, LIV); X: leucine independent; Lrp binds in at least 8 of 9 conditions; M: genes with any other pattern of Lrp binding. Several additional transposition-related terms with similar expression profiles to GO:0032196 are omitted for clarity. (E) As in panel D, showing dependence of binding on growth phase. Categories are: L: Lrp binds target genes only in log phase (in one or more media types, but at no other growth phase); T: Lrp binds target genes only in transition phase; S: Lrp binds target genes only in stationary phase; M: all other genes with Lrp binding across multiple growth phases.

metabolism independently of leucine, suggesting an important role for poised regulation in mediating the key metabolic functions of Lrp. A similarly small set of GO terms shows consistent occupancy patterns at different phases of growth (Figure 2.11E), although the inclusion of tRNAs among those targets is notable. These findings highlight the important role played by Lrp in repressing the translational apparatus during stationary phase, but at the same time, demonstrate the importance of poised regulation and local regulatory interactions in setting the effects of Lrp at each bound promoter.

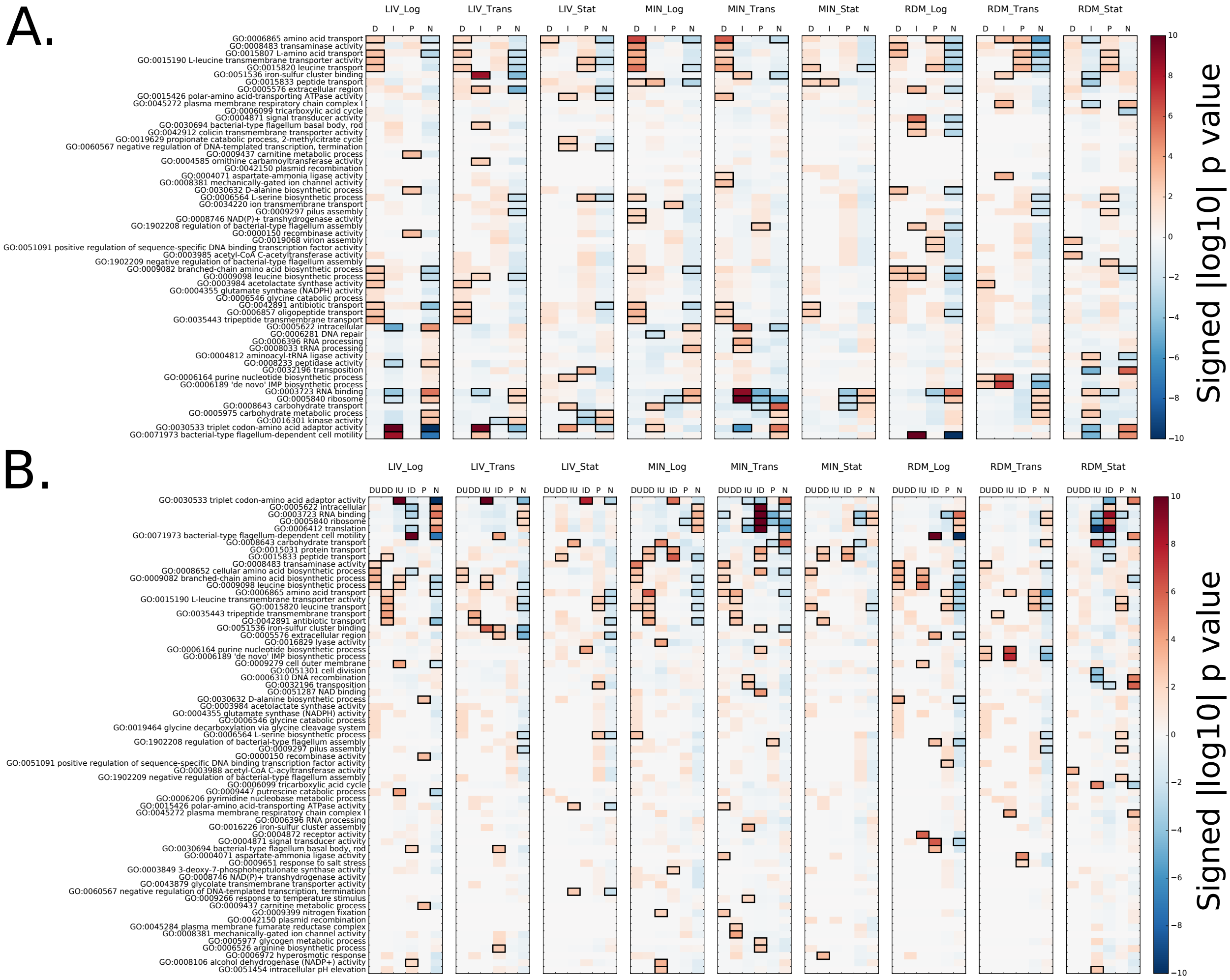

Figure 2.12: Full GO-term enrichment results for general target classification and sub-classification by direction of Lrp regulatory change. (A) All GO-terms identified by iPAGE as having significant mutual information with our target classification within various conditions are listed to the left. Abbreviations are as follows: D - Direct targets, I - Indirect targets, P - Poised targets, N - 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, P - Poised targets, N - No Lrp link genes.

2.4.7 Lrp shows condition-dependent changes in DNA sequence specificity

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 has been previously identified for Lrp [58, 157]. To determine how well previously identified Lrp motifs could predict the binding sites identified in our study, we used a logistic regression model to classify 500 bp windows of the genome as either containing a Lrp binding site 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 2.13A; 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 2.14). 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 general preference for low information content AT-rich regions at Log points and a preference for higher information content sequence motifs at later time points across all conditions (Figure 2.13A). Here we are referring to information content in the information-theoretic sense, i.e. a motif with higher information content indicates that protein has higher specificity for more positions within the motif and we consider a motif with higher information content to indicate a higher sequence specificity. In each condition, from early to late time points, there is a decrease in how predictive the general AT-content is in terms of differentiating between Lrp binding sites and background genomic locations. 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 less influenced by higher information content sequence motifs in earlier phases of growth, and only gains preference for these higher information content sequence-motifs 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 preference for higher information content sequence motifs 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 derived models perform relatively well; the receiver operator curves, which show the re-

Condition	MCC[a]	Specificity[a]	Sensitivity[a]	ROC-AUC[a]
LIV_Log	0.61 (0.45-0.77)	0.85 (0.81-0.87)	0.80 (0.66-1.00)	0.86 (0.73-1.00)
LIV_Trans	0.38 (0.26-0.48)	0.77 (0.72-0.80)	0.64 (0.49-0.76)	0.78 (0.71-0.83)
LIV_Stat	0.40 (0.30-0.47)	0.77 (0.73-0.81)	0.66 (0.54-0.71)	0.79 (0.75-0.84)
MIN_Log	0.34 (0.29-0.38)	0.74 (0.70-0.77)	0.64 (0.56-0.72)	0.77 (0.70-0.81)
MIN_Trans	0.25 (0.16-0.30)	0.71 (0.68-0.76)	0.57 (0.49-0.62)	0.70 (0.65-0.72)
MIN_Stat	0.24 (0.19-0.34)	0.72 (0.69-0.74)	0.55 (0.47-0.63)	0.70 (0.65-0.74)
RDM_Log	0.49 (0.38-0.54)	0.80 (0.72-0.85)	0.74 (0.65-0.83)	0.84 (0.81-0.86)
RDM_Trans	0.31 (0.22-0.41)	0.74 (0.67-0.78)	0.60 (0.54-0.69)	0.73 (0.70-0.77)
RDM_Stat	0.34 (0.24-0.46)	0.74 (0.69-0.79)	0.64 (0.54-0.74)	0.77 (0.73-0.82)

[a] Values in parentheses show the minimum and maximum values from 5-fold cross-validation for each metric.

Table 2.4: Performance of Lrp binding site prediction models

call for every potential false positive rate, trend toward the upper left corner where a perfect model would be (Figure 2.13B; quantified by area under the curve, ROC-AUC, in Table 2.4). In addition, the Matthews correlation coefficient (MCC), a combined measure of precision and recall which has potential values from -1 to 1, ranges from 0.24 to 0.61 (Table 2.4). 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 parenthesis in Table 2.4). 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 higher information content 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, as well as efforts to determine any sequence elements that differentiated activated from repressed targets (data not shown). This could simply indicate that sequence independent mechanisms, such as the well-established observation of Lrp cooperativity in binding [158], or recruitment of Lrp by binding of additional factors, could play a role in determining Lrp binding locations.

2.4.8 Lrp binding is enriched among regulatory regions of the genome

As detailed in Section 2.6.17, our process for categorizing genes as Lrp targets involved testing whether there was a called Lrp peak overlapping anywhere within 250 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 250 bp upstream of the most distal TSS (relative to the start of the gene itself) and 250 bp downstream of the most proximal TSS. We classified those approximately 500 bp windows as regulatory regions, and tested whether Lrp binding was sig-

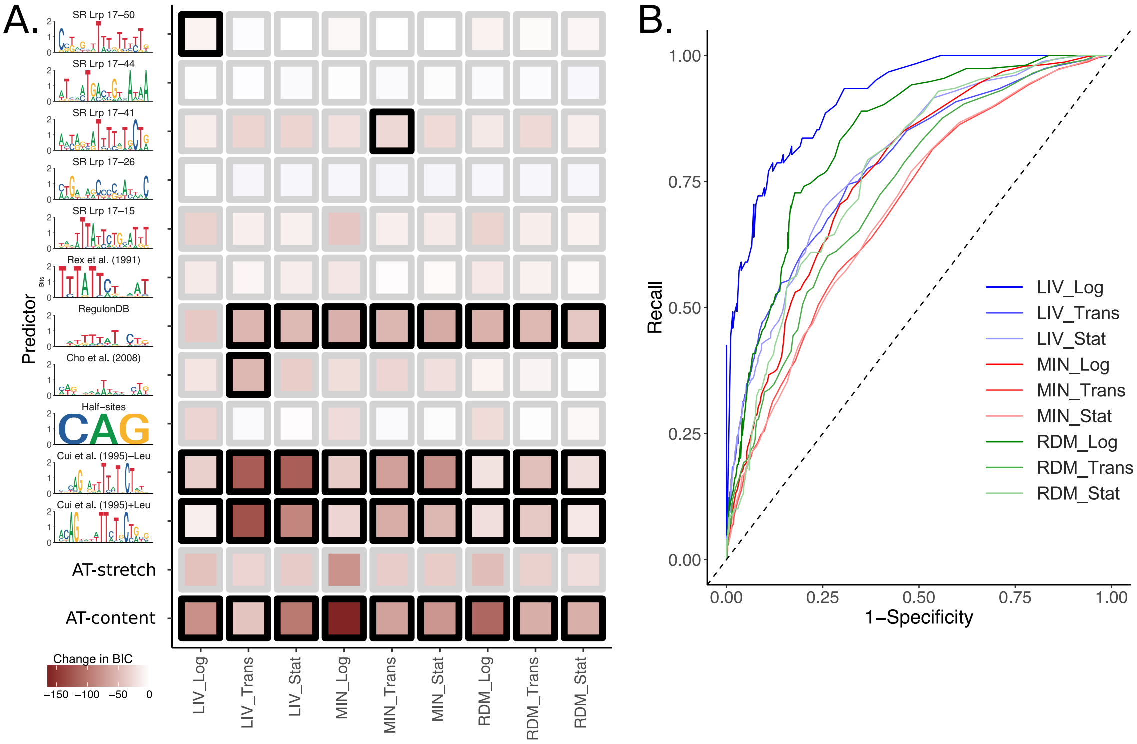

Figure 2.13: Lrp exhibits condition-dependent sequence-preference. (A) Change in BIC for add-one-in logistic regression models. The y-axis displays the Position Weight Matrix (PWM) used to create a particular feature. PWMs were obtained from the publication indicated above the PWM [66, 157, 159], RegulonDB [48] or, in the case of SR motifs, the SwissRegulon [160]. Features were created from a given PWM by dividing the count of matches within a sequence (as obtained by FIMO [161] with q-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, moving from dark red (negative BIC, added term is favored in the model) to light blue (positive BIC, added term is disfavored in the model), 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 cut off from the logistic regression. Full statistics including five-fold cross-validation are included in Table 2.4.

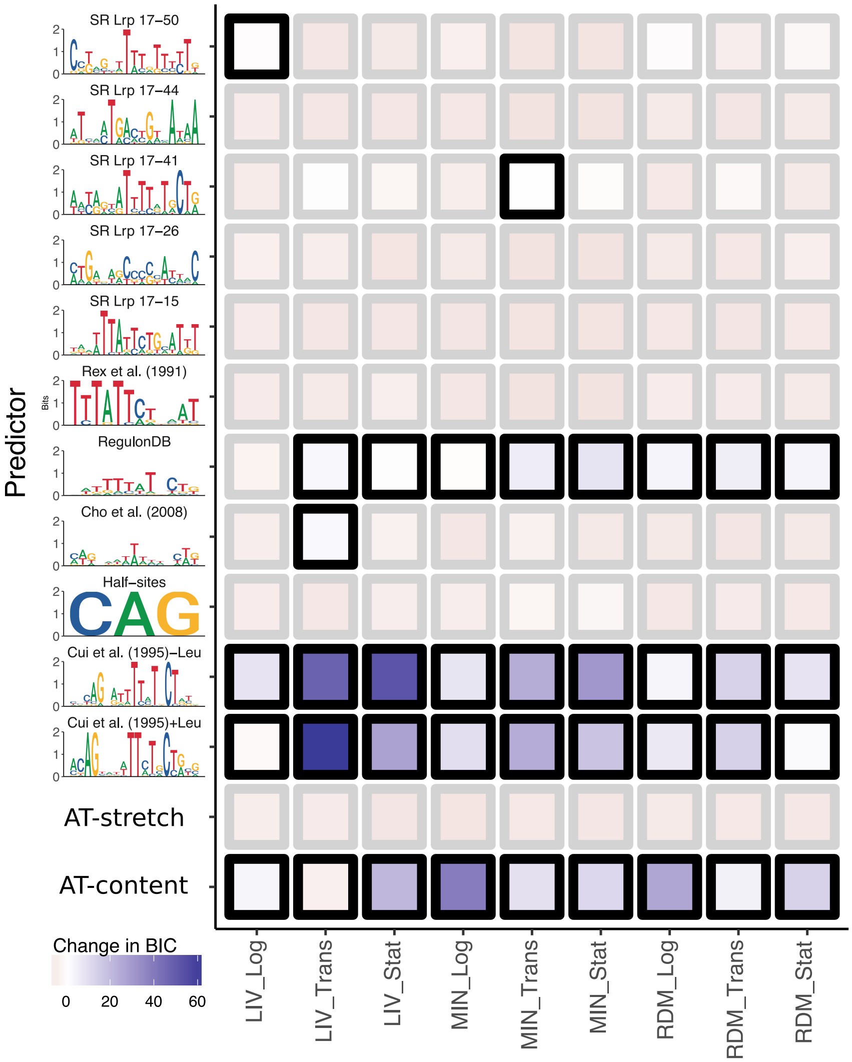

Figure 2.14: Changes in BIC for leave-one-out logistic regression models. PWMs were obtained from the publication indicated above the PWM [66, 157, 159], RegulonDB [48] or, in the case of SR motifs, the SwissRegulon [160]. Features were created from a given PWM by dividing the count of matches within a sequence (as identified by FIMO [161] 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, moving from light red (negative BIC, removed term is disfavored in the model) to dark blue (positive BIC, removed term is favored in the model), indicate 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.

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	$< 1.0 \times 10^{-3}$
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 2.5: Results of permutation test for enrichment of Lrp binding in regulatory regions

nificantly enriched anywhere within those regions. Overall, 29% of the *E. coli* genome falls into these regulatory regions. However, we observe between 53% and 85% of Lrp peaks overlapping with 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 2.5). 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 2.6); the only exception is at LIV_Log, which only has 61 Lrp peaks, thus leading to some skewing of expected percentages. 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, did not have a Lrp peak within the promoter region but did have a Lrp peak overlapping the gene coding region. We then 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* (Figure 2.2B), we again note that Lrp binding does not guarantee a regulatory effect. Genes that have an internal Lrp binding site do not evince a Lrp dependent change in expression, and Lrp binding sites within an operon do not, in general, appear to hamper transcription (Figure 2.7C-E). 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 appear to have a constant effect.

2.4.9 Uncertainties in identification of poised Lrp targets

An important caveat to consider in the analysis of 'poised' targets is that some of the identified poised sites may represent misannotation due to Lrp binding that is specific to a nearby gene such

Condition	% in gene region	% in transcription-unit	% in intergenic region[b]
Genome-wide[a]	95.6	0.6	3.8
MIN_Log	97.6	0	2.4
MIN_Trans	99.2	0	0.8
MIN_Stat	100	0	0
LIV_Log	66.7	11.1	22.2
LIV_Trans	98.2	0.6	1.2
LIV_Stat	97.9	0	2.1
RDM_Log	93.9	3.0	3.0
RDM_Trans	98.3	0	1.7
RDM_Stat	97.8	0	2.2

[a] Percentages genome-wide were determined at a 1 bp resolution.

[b] Intergenic defined as region neither in regulatory region, gene or transcription-unit.

Table 2.6: Percentages of non-regulatory region peaks that annotate to other mutually exclusive regions of the genome

as the *ybjN*/*potF* case above (Figure 2.9A). Systematic analysis of all poised sites showed that targets that we classified as poised but had a direct target within 1000 bp of their regulatory region in the same condition represent a small fraction of all total sites (Figure 2.7A, POISED_nearby_direct class). However, of the poised sites that never transition to direct targets in any condition, we must distinguish between those that clearly have no Lrp-dependent regulation, and those for which our RNA-seq analysis lacks the statistical power to say with certainty that there is no Lrp dependent change. To distinguish between those cases, we defined a region of practical equivalence (ROPE) as a Lrp-dependent change in expression of less than 1.5-fold, and account as clearly Poised the subset of putative Poised targets for which a conservative 95% confidence interval falls entirely within the ROPE (Figure 2.7A, POISED_unexplained). For the remainder of sites, due to the uncertainty in our experimental measurements, we do not have strong enough evidence from our RNA-seq results to decisively conclude that there is no Lrp dependent effect on the target gene (Figure 2.7A, POISED_uncertain_direct). However, for the vast majority of these cases, the maximum likelihood estimate of the Lrp-dependent change falls within the ROPE, and thus in all probability only a small fraction of the POISED_uncertain_direct class are actually mis-annotated direct targets; the majority are most likely true Poised targets with no Lrp-dependent change in expression.

It is also useful to consider the question of what fraction of poised peaks will become active under at least one other physiological condition. By its nature this question requires extrapolation from our data set to hypothetical other, unobserved conditions, but we can at least provide an approximate answer by considering the rate of discovery of targets that are always poised, vs. targets that are poised in some conditions and direct targets of Lrp in others, as we expand among the conditions in our study. We plot the results of analyzing these discovery rates in Figure 2.15, which leads us to three main conclusions. First, our discovery of direct targets is slower than our discovery of poised targets, indicating that a given target is likely poised under more conditions than it is a direct target. Thus, consideration of a broader range of conditions is necessary to exhaustively identify the set of poised targets which become direct under at least one condition. Second, the discovery of new direct targets is by no means saturated among the set of conditions that we have so far identified, whereas the discovery of poised targets is nearing saturation (compare the slopes of the Poised vs. Poised and Direct vs. Direct curves in Figure 2.15). Thus, we expect that consideration of additional conditions would bring discovery of relatively fewer exclusively poised targets, while a larger fraction of the poised targets would be found to become direct under at least one condition. Attempting to press on to estimate the precise fraction of poised targets which would be direct under at least one conceivable condition seems to us unduly speculative at this point, as we do not know how conditions highly dissimilar to those considered here would affect our discovery of new targets, but at the very least, based on the data in Figure 2.15 it seems likely that the fraction of poised targets which are direct under at least one condition (and thus represent functional Lrp sites rather than sites that play no regulatory role) would substantially increase beyond what we have observed here.

2.4.10 Lrp connects with other regulatory factors

The phenomenon of poised targets—at which Lrp frequently binds to a promoter under many conditions but only shows regulatory activity under a few—suggests that other regulatory factors, such as σ factors or transcription factors, may be important in triggering an activating or repressive effect secondary to Lrp binding. If a σ factor and Lrp co-regulate some set of targets, we expect to see enrichment for direct targets relative to poised targets within the σ factor's regulon, especially at conditions when the σ factor is most active. To establish relative σ factor activity, we determined the average expression of all known σ factor target genes (taken as the union of the Sigma-gene interactions annotated in RegulonDB and the factor associated with each annotated TSS to each gene (Factor column in Supplementary Data File 1) at each of our nine experimental conditions (Figure 2.16A) [48]. This allows us to estimate in which conditions the σ factor is most active; for example, σ^{38} activity peaks at LIV_Stat, MIN_Trans and RDM_Stat in agreement with its role as the general stress response σ factor. 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 σ factor all genes that are classified are regulated. Subject to these constraints, our analysis in this section included 2885 genes (out of the total 4685 genes in *E. coli*). In addition, in some

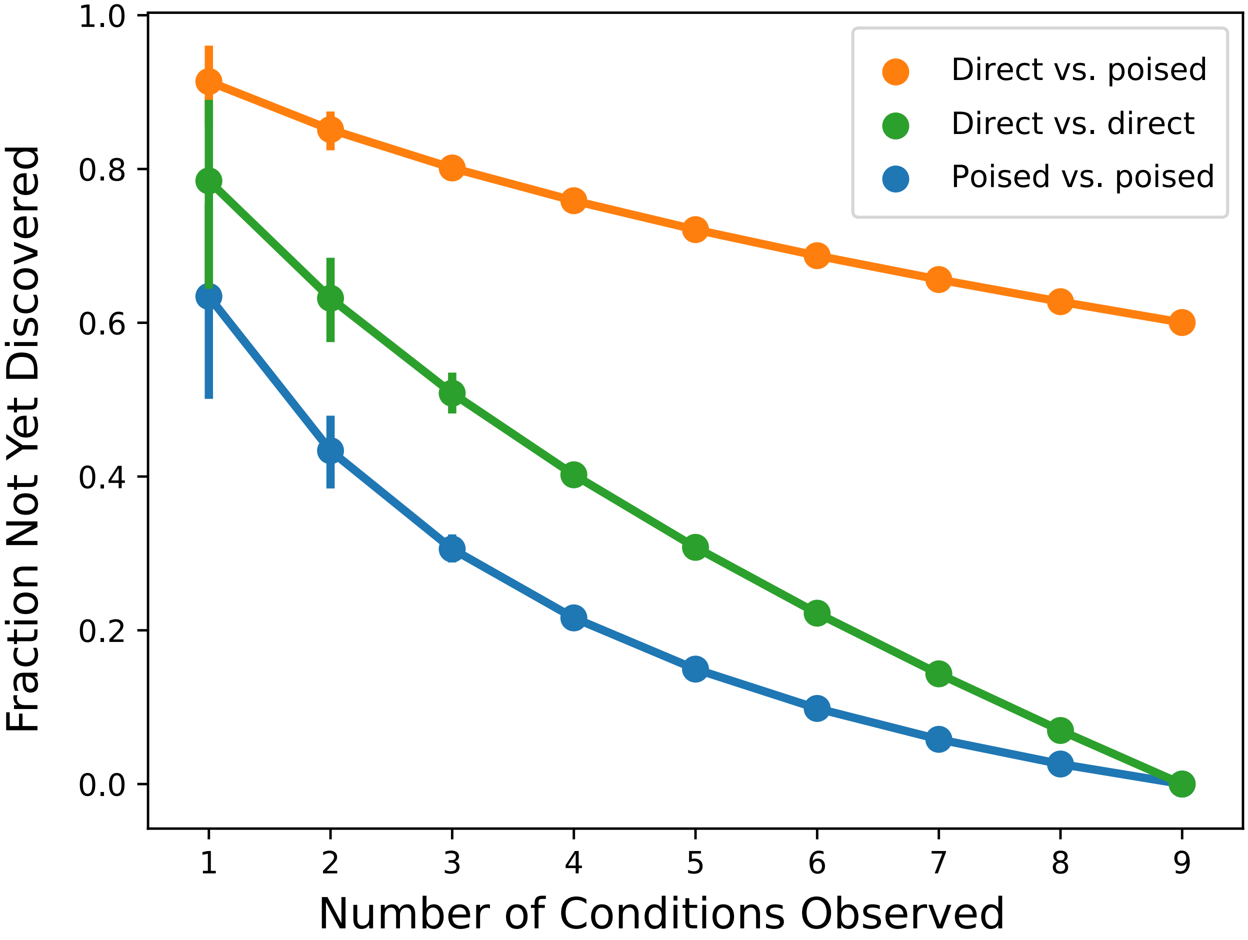

Figure 2.15: Rates of discovery of new poised and direct targets as additional conditions are considered. Results from resampling calculations enumerating the fraction of direct and poised targets that remain undiscovered when considering only one of the nine conditions in our study, to considering all nine of them. In each case, points represent the mean across all possible orderings of acquisition of experimental knowledge, and error bars represent bootstrap-based 95% confidence interval for the mean of the observations across different possible condition orderings. 'Poised vs. poised' refers to the fraction of all poised targets known in the entire data set that have been discovered by a certain point, and likewise, 'Direct vs. direct' indicates the fraction of all direct targets from our entire data set that are discovered by a certain point (thus both reach zero when all nine conditions, which comprise our entire data set, are considered). 'Direct vs. poised' indicates the fraction of all poised targets (pooled across all experimental conditions in our study) that have been observed as a direct target at least once after consideration of a given number of conditions.

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 σ factors to activate target genes at specific conditions, there are a few possible explanations for why the poised to direct target transition occurs at those points: 1) the transition only occurs when the genes' controlling σ factor is active; 2) the nature or extent of Lrp binding itself changes at that condition; or 3) an accessory factor needed for Lrp-σ factor interaction is only present at that condition.

We applied a permutation test to identify any σ 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 2.7; results from the permutation test with only direct-activated targets are in Table 2.8, ($r = 10000$ for both). Only σ^{54} at MIN_Trans had significant overlaps ($q < 0.05$). Specifically, we document enrichment for direct Lrp targets with $\sigma^{54}(\sigma^{N})$ at MIN_Trans (1.9-fold enrichment, q-value: 0.038). At MIN_Trans, 37% of Lrp binding sites overall are direct targets, whereas 70% of σ^{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 σ factor, there is enrichment specifically for direct Lrp-activated target genes among σ^{54} targets at MIN_Trans (2.6-fold enrichment, q-value: 0.032). Overall, 19% of Lrp binding sites are direct activated targets at MIN_Trans, whereas Lrp-bound targets in the σ^{54} regulon are direct Lrp-activated targets 50% of the time, a 2.6-fold increase. σ^{54} regulates many genes involved in nitrogen assimilation [162], and these results indicate that Lrp is likely involved in co-activating some σ^{54} dependent genes, in agreement with Lrp's role in sensing and responding to nutrient levels. At MIN_Trans, Lrp actually also weakly represses σ^{54} itself directly; σ^{54} is not a direct or indirect target under any other conditions.

Average expression of σ^{54} targets reaches peak levels at MIN_Trans (Figure 2.16A), in agreement with when we see significant overlap with Lrp direct-activated genes (15.8% of the direct Lrp-activated targets at MIN_Trans are known σ^{54} targets, and conversely 13.9% of the classified σ^{54} targets are direct Lrp-activated targets at MIN_Trans). Twelve out of the fifteen overlapping target genes only become a direct Lrp-activated target at MIN_Trans. The remaining three genes (*astC*, *potF*, *yhdW*) are sometimes affected at conditions when there is a slight peak in σ^{54} activity, as measured by the overall expression of known target genes (Figure 2.16A), and could be subject to other regulatory control. For example, astC is also regulated by ArgR in some conditions [163, 164]. The fact that the shared regulated genes are generally only direct Lrp-activated targets when σ^{54} itself is most active supports the notion that σ^{54} may require Lrp binding to activate transcription of certain genes. At a molecular level, this suggests that while expression of σ^{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.

To investigate the possibility that Lrp binding itself changes to facilitate interaction with σ^{54}, we visualized the Lrp-ChIP binding signal at shared direct Lrp/σ^{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 very similar binding regardless of condition (see Figure 2.9A for potF Lrp-binding signal); thus, it is unlikely that changes in Lrp binding itself are in general responsible for the regulatory interaction with σ^{54}. Given that σ^{54} is known to require activating factors, it is likely that an accessory factor may facilitate Lrp/σ^{54} co-regulation.

To identify other candidates for co-regulators acting with Lrp, just as we tested for Lrp co-regulation with σ factors, we investigated whether Lrp has particular correlations with any of the other annotated transcription factors in *E. coli*. 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 ($r = 10000$): FlhDC, GadW, ModE, and NtrC. 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 2.16B); the transcription factor FlhDC did not pass this filter and was eliminated from further analysis. ModE likely represents an example of convergent regulation due to the existence of no or limited overlap between its targets and direct Lrp targets. As detailed in the main text, GadW likely represents an example of indirect Lrp regulation via direct regulation of a transcription factor (see Section 2.4.3).

The transcription factor NtrC is a notable exception to the above trends, as 31% of all its targets are also direct Lrp-activated targets (Figure 2.16C). 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, 63% of NtrC targets are also direct Lrp-activated targets. NtrC is one of the transcription factors which can serve as an activator of σ^{54}, so the intersection between Lrp, NtrC and σ^{54} is interesting to consider. Activators of σ^{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 σ^{54}; in previous studies, the bending has been documented as being intrinsic to the region or looping mediated by IHF [36]. In accordance with the possibility of intrinsic bending, the average AT content upstream of σ^{54} target genes is 70%, with the lowest being at 50% [162]. As previously reported and seen in our data, Lrp is known to bind AT-rich regions preferentially [165]. Lrp induces bending of 52° to 135° depending on the size of the binding sites [166]. Thus, we hypothesize that Lrp may play a role in bending DNA to coordinate NtrC-σ^{54} interaction at NtrC targets. Thus, while many instances of Lrp regulation appear to require co-regulation with as yet unidentified regulatory

Condition	Value	σ^{24}	σ^{28}	σ^{32}	σ^{38}	σ^{54}	σ^{70}
MIN_Log	q-value	1	0.671	1	0.661	0.450	0.784
	Fold change[a]	0.51	1.30	0.71	1.30	1.80	1.04
MIN_Trans	q-value	1	0.848	0.946	0.755	0.038	1
	Fold change[a]	0.75	1.01	0.97	1.12	1.89	0.94
MIN_Stat	q-value	1	1	0.848	0.644	0.644	0.946
	Fold change[a]	0.65	0.42	1.02	1.60	1.75	0.99
LIV_Log	q-value	1	1	1	0.644	1	0.802
	Fold change[a]	0.78	0.58	0.58	1.40	0	1.04
LIV_Trans	q-value	1	0.755	1	0.784	1	0.644
	Fold change[a]	0.61	1.32	0.68	1.20	0.41	1.14
LIV_Stat	q-value	0.848	0.848	1	0.792	0.644	1
	Fold change[a]	1.12	1.02	0	1.28	2.36	0.95
RDM_Log	q-value	1	0.671	1	1	0.661	0.450
	Fold change[a]	0.37	1.62	0.40	0	1.80	1.28
RDM_Trans	q-value	1	0.848	1	0.848	0.792	0.450
	Fold change[a]	0.31	1.01	0.34	0.97	1.31	1.30
RDM_Stat	q-value	1	1	0.165	0.644	0.661	1
	Fold change[a]	0.56	0.28	1.69	1.32	1.39	0.90

[a] Fold change is calculated by dividing the fraction of bound σ factor targets (either direct or poised) which are classified as direct targets, by the overall fraction of Lrp-bound targets which are direct targets.

Table 2.7: Results of permutation test for enrichment of direct Lrp targets relative to poised targets within the known σ factor regulons at each condition

factors, we are able to identify some likely possible mechanisms.

2.5 Discussion

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

By investigating both the binding and regulatory activity of Lrp under several media conditions and time points, we are able to present a broader 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, which appear in our experimental setup to correspond to times of high Lrp activity due to dropping nutrient conditions. Targets may appear selectively in certain conditions due to a

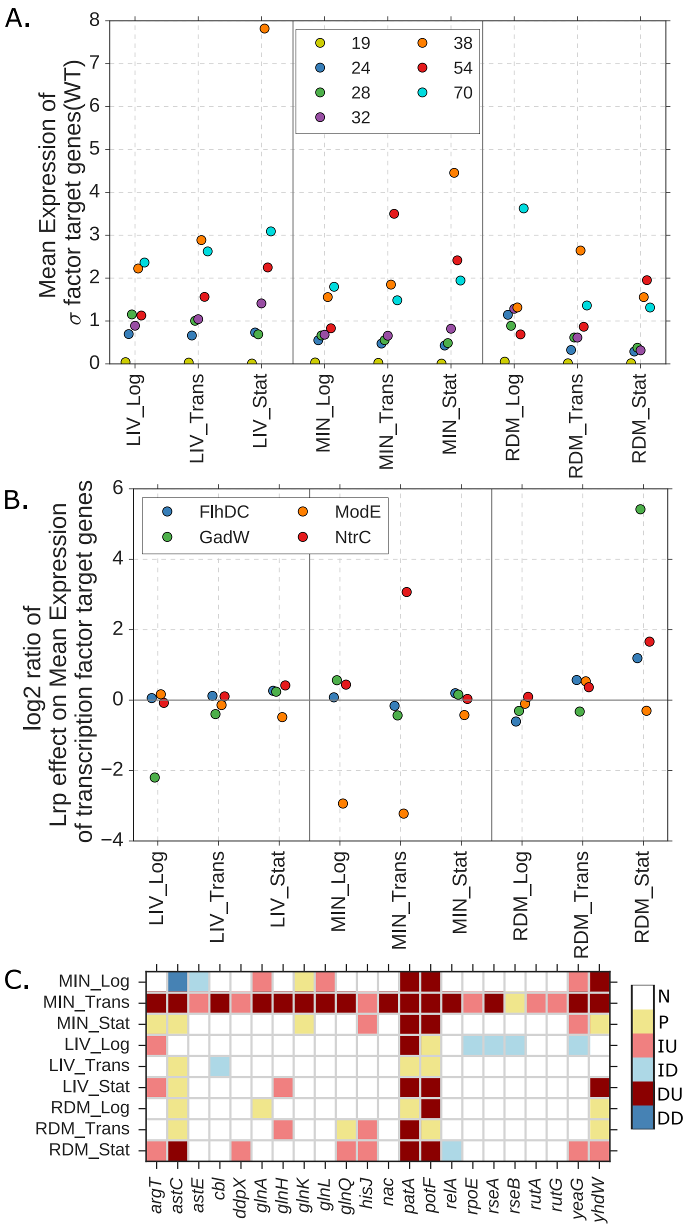

Figure 2.16: Lrp interacts with other regulatory factors to control some targets' expression. (A) Average expression of known targets of each σ factor in WT cells at each condition (calculated for each gene as normalized transcript abundance divided by gene length). (B) Average $\log_2$(WT/KO) expression ratio of known transcription factor targets for selected transcription factors at each condition. (C) Heatmap showing classification of those NtrC targets which have an annotated transcription start site and thus are classified in our analysis. Abbreviations on the color bar are as follows: DD - Direct Downregulated targets, DU - Direct Upregulated targets, ID - Indirect Downregulated targets, IU - Indirect Upregulated targets, P - Poised targets, N - No Lrp link.

Condition	Value	σ^{24}	σ^{28}	σ^{32}	σ^{38}	σ^{54}	σ^{70}
MIN_Log	q-value	1	0.846	1	0.837	0.592	0.837
	Fold change[a]	0.19	1.12	0.92	1.22	2.33	1.08
MIN_Trans	q-value	1	0.846	0.741	1	0.032	1
	Fold change[a]	0.81	1.07	1.29	0.90	2.57	0.82
MIN_Stat	q-value	1	1	0.846	0.741	0.837	0.846
	Fold change[a]	0.32	0	1.24	1.68	1.43	1.07
LIV_Log	q-value	1	1	1	0.741	1	0.837
	Fold change[a]	0.81	0	0	1.62	0	1.08
LIV_Trans	q-value	1	0.988	1	0.837	1	0.695
	Fold change[a]	0.75	1.01	0	1.47	0	1.29
LIV_Stat	q-value	0.741	1	1	0.741	0.592	1
	Fold change[a]	1.74	0	0	1.98	3.65	0.65
RDM_Log	q-value	1	1	1	1	0.837	0.479
	Fold change[a]	0	0	0.78	0	1.72	1.45
RDM_Trans	q-value	1	1	1	0.837	1	0.741
	Fold change[a]	0.48	0	0.53	1.53	0	1.27
RDM_Stat	q-value	1	1	0.894	0.286	0.741	1
	Fold change[a]	0.57	0.48	1.04	2.26	1.59	0.84

[a] Fold change is calculated as for Table 2.7 except that the number of direct targets is replaced with the number of direct Lrp-activated targets.

Table 2.8: Results of permutation tests for enrichment of direct Lrp-activated targets relative to direct Lrp-repressed and poised targets within the known σ factor regulons at each condition

number of potential influences, including: 1) variable levels of Lrp protein may dictate that only the strongest binding sites are occupied; 2) required co-regulators may only be expressed in certain conditions; 3) post-translational modifications of Lrp may influence its binding or interaction with co-regulators. In addition, the high number of indirect targets that are unique to one or two conditions are most easily explained by invoking one transcription factor that is regulated by Lrp as we discuss above, but we cannot know from this study alone how many levels of regulatory control are actually in play for many indirect targets. Given that, it is clear that there are many possibilities for condition-specific regulation.

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. Many of the identified GO-terms include genes previously implicated as Lrp targets, indicating agreement with previous work. However, newly identified targets and novel patterns of regulation (such as poised binding) suggest that further work on the mechanistic aspects of Lrp regulation is important.

2.5.2 Poised Lrp binding argues for interaction with co-regulatory 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 poised sites. If that was the case, we would expect Lrp binding sites (the majority of which are poised sites in any condition) to be distributed more evenly across the genome. The idea of poised regulation is not without precedent, as poised regulation has also been reported for some eukaryotic transcription factors such as the tumor suppressor p53 in binding to the *mdm2* gene [148]. 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) switching between the presence of a Lrp octamer or hexadecamer may control or influence the regulatory behavior of Lrp. We investigated the first possibility by analyzing if certain σ factors or transcription factors might be responsible for the condition-dependent regulation on a global scale (see Section 2.4.10). While many potential connections appear to be cases of convergent regulation, we do find that Lrp may facilitate NtrC/σ^{54} interaction by binding and bending DNA. This would agree with the connection between Lrp and nitrogen metabolism regulation seen previously in genome-wide studies [167]. Analogous interactions with other transcription or regulatory factors may explain other poised/direct target transitions. For example, Lrp interaction with H-NS is important for regulating rRNA promoters [168], and Lrp competition with DNA adenine methyltransferase is critical in regulating expression of the *pap* operon, which produces pili [169]. In addition, non-protein small molecules like ppGpp are known to affect some Lrp-regulated target genes [170]. Finally, although we do not see global evidence in our analysis, gene-level studies have previously implicated Lrp in interacting with σ^{38} [171, 172]. Further studies are needed to investigate Lrp's interactions with other regulatory factors and the alternate mechanisms proposed above. We must also acknowledge the possibility that some fraction of the always-poised sites present in our data set are in fact false positives; some false positive rate is essentially unavoidable in a high throughput experiment of this type, and thus the behavior of any particular site can only be resolved with certainty through a targeted follow-up experiment. However, several lines of evidence point to the majority of poised sites being genuine, and likely being cases where Lrp binding will play a regulatory function under an as-yet unstudied condition: our ChIP-seq peak calling pipeline is designed to err toward being conservative; even the low-intensity binding sites near direct targets called in our study have levels of regulatory activity similar to those identified in previous experiments (Figure 2.5); and extrapolation from our existing set of conditions suggests that while we have neared saturation in our discovery of poised targets, the fraction of poised targets that become direct under at least one condition is likely to increase upon consideration of additional conditions not studied here (Figure 2.15).

2.5.3 Lrp binding activity is partially predicted by known sequence motifs

While we detected a preference for Lrp binding at several previously-identified, related motifs and AT-rich regions, there are still a significant subset of peaks that are not predicted by these models. We were unable to improve Lrp binding prediction from additional sequence determinants 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 [146], 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 [57] or potentially post-translational modifications [173, 174] 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.

2.6 Materials and Methods

2.6.1 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 [175]. Primers used for strain construction and validation are listed in Table 2.9. 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 [154] and grown overnight. The cells were then back-diluted to OD_{600} = 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 0.2% 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 2.10.

The cells were grown at 37°C with shaking (200 rpm) until the OD_{600} 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),

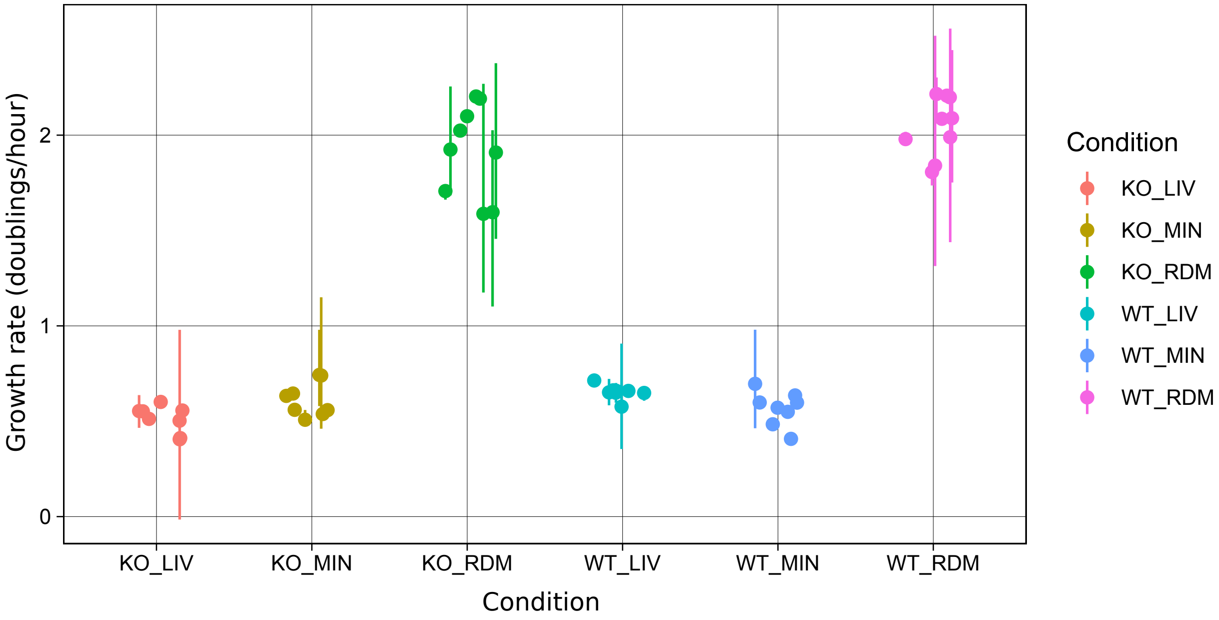

Figure 2.17: Exponential phase growth rates for all biological replicates used in the present study. Plotted are the biological replicate level log phase growth rates observed during growth of each sample used in the present study. Growth rates were calculated using a simple linear regression for each replicate (solid points), and error bars indicate a 95% confidence interval based on jackknife resampling at the level of individual optical density measurements. Time points to use in the calculation were selected from the overall growth curve by identifying the points most closely prior to an OD_{600} level of 0.15-0.25 which exhibited log linear growth kinetics.

between 2.3 and 2.7 (for transition point in RDM), or 12 hours past the log point (for stationary phase samples). The same incubator was used for all cell growth in order to limit variation in temperature or aeration. The OD_{600} 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 OD_{600} range during which exponential growth becomes non-linear when visualized on a log scale (Figure 2.1). Logarithmic growth rates for all samples are summarized in Figure 2.17. Associated data for growth rates can be found in Supplementary Data File 2.

2.6.2 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 min 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 min. The culture was then immediately centrifuged for 5 min at 5,500 $\times g$ 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 10,000 $\times g$ at 4°C and removal

Identifier	Sequence	Notes
P1582	TCAGACAGGAGTAGGGAAGGAATACAGAGAGACAATAATATGTGTAGGCTGGAGCTGCTTC	Generate Kan cassette to delete *lrp*
P1583	GAGTGTAATCAAAATACGCCGATTTTGCACCTGTTCCGTGCATATGAATATCCTCCTTA	
P965	GAACTTCGAAGCAGCTCCAG	Test *lrp::kanR* deletion
P1568	CAAGGCAACGGTCTTCTCAC	
P1569	CCTGGCTCAAGAAAGGCTCT	

Table 2.9: Primers used for *lrp::kanR* construction

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

[a] All MOPS media formulations are based on [154].

Table 2.10: Media conditions for cell growth

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 (phosphate-buffered saline [PBS], 0.1% Tween 20, 1 mM EDTA, 1× cOmplete Mini EDTA-free Protease Inhibitors (Roche; Basel, Switzerland), 0.6 mg lysozyme (Amresco; Solon, OH)), vortexed for 3 s, and incubated at 37°C for 30 min. The sample was then sonicated in 3 bursts of 10 s each at 25% power (Branson Digital Sonifier). Cellular debris was removed by centrifugation at 16,000 $\times g$ for 10 min at 4°C. To obtain an accurate representation of the isolated pool of DNA before the extraction procedure, 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)) to be the input sample. 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 hr at room temperature for pre-clearing. A separate aliquot of 100 μL of pre-washed SureBeads Protein G magnetic beads was incubated with 10 μg Lrp monoclonal antibody (Neoclone; Madison, WI) for 10 min 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 hr 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 μL Elution Buffer and incubated at 65°C for 20 min, vortexing every 5 min. 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 hrs at 37°C, then 0.2 mg Proteinase K (Thermo Fisher; Waltham, MA) for 2 hrs 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 NextSeq500, with 38×37 bp paired end reads. We obtained at least 3,000,000 reads that passed all filters and aligned properly to the genome per biological replicate with an average of 9,000,000 reads per replicate (Supplemental File 2). Input samples were treated identically to the ChIP extracted samples beginning at the overnight incubation to reverse the cross-links.

2.6.3 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 min at room temperature, and then centrifuged for 10 min at 5,000 $\times g$ 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 10 min at room temperature, vortexing every two min. 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 min 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 [176]. 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 NextSeq500 at the University of Michigan's DNA Sequencing Core Facility.

2.6.4 Preprocessing of ChIP-seq data

Sequencing adapters were removed from all sequences using CutAdapt version 1.8.1 [177] with parameters -a AGATCGGAAGAGC -A AGATCGGAAGAGC -n 3 -m 20 –mask-adapter –match-read-wildcards. Low quality reads were trimmed with Trimmomatic version 0.32 [178] using the parameters TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20. The quality of the raw and preprocessed fastq files was assessed using FastQC version 0.10.1 [179] and MultiQC version 1.2 [180]. The number of raw and surviving reads for each sample are described in Supplemental File 2.

2.6.5 Alignment of ChIP-seq data

All samples were aligned to the MG1655 U00096.2 genome modified to match the insertions and deletions for the ATCC 47076 variant of *E. coli* MG1655 as reported by [181]. Alignments were performed using bowtie version 2.1.0 [182] and arguments: -X 2000 -q –end-to-end –very-sensitive -p 5 –phred33 –dovetail in order to maximize the sensitivity of the alignment. Final alignment rates for each sample are described in Supplemental File 2.

2.6.6 Calculation of ChIP-seq summary signal

The amount of Lrp-mediated DNA enrichment in any given experimental condition or genotype is represented by two different sequencing reactions: An extracted sample, where the DNA crosslinked to Lrp is extracted and purified using a specific monoclonal antibody, and a matched input sample (taken from the same tube after lysis and digestion), where the total input DNA before the extraction procedure is sequenced. Throughout the following text references to the extracted and input samples will refer to the definitions above for any given pair of samples for each combination of experimental condition and genotype. To determine the raw enrichment for a set of paired extracted and input samples, the coverage c of paired-end reads at every tenth base pair n across the genome, was calculated from the alignments for the ChIP-extracted and input reads for each sample separately using samtools [183] and custom python scripts. The raw read coverage for extracted and input samples was then scaled using the median coverage across the genome for each individual track to account for differences in sequencing depth between the two samples. The median coverage was chosen as a scaling factor as it represents an estimator of the baseline read coverage in a given sample that is less impacted by the exact heights of the peaks within that sample. The raw enrichment (RE) was calculated using the $\log_2$ ratio of scaled extracted to input coverage separately for each pair of extracted and input samples as shown below:

$$RE_{(n)} = \log_2 \frac{c_{E(n)}}{\text{median}(c_E)} - \log_2 \frac{c_{I(n)}}{\text{median}(c_I)} \tag{2.1}$$

Where E and I denote the extracted and input samples, respectively. Thus, the RE represents a $\log_2$ transformed ratio of normalized extracted DNA abundance to normalized input DNA abundance. In addition, these $\log$ transformed ratios put each sample, for each condition, on the same scale and removes the need for additional normalization between samples, thereby allowing for direct comparison between REs from any given genotype or experimental condition.

Unique to this experimental set-up is the use of Lrp knockout samples as an additional control to account for any biases from the antibody-mediated extraction procedure. Possible biases could include antibody interactions with the DNA or low-level cross reactivity with other crosslinked proteins during the extraction procedure. To the best of the author's knowledge, no existing ChIP-seq analysis pipeline is able to use two separate sets of control information from both an input control sample and an entirely separate set of extracted and input samples under the knockout genotype for the protein of interest. Therefore, we set out to create our own pipeline tailored to this experimental design with the goal of minimizing the high rate of false positives commonly seen in ChIP-seq experiments. We use the *lrp::kanR* samples to remove enrichment that also exists in the absence of Lrp by subtracting the *lrp::kanR* RE signal from the Lrp WT RE signal to obtain a raw enrichment signal (RSE) for any combination of WT and *lrp::kanR* replicates within a single experimental condition (see Figure 2.18B and C. for an example of how this subtraction removes false positive peaks). The RSE is represented mathematically below:

$$RSE_{(n)} = RE_{WT(n)} - \text{max}(RE_{\textit{lrp::kanR(n)}}, 0) \tag{2.2}$$

The max function in the equation above ensures that the *lrp::kanR* signal is subtracted only if its RE was positive. Since both the Lrp WT RE signal and the *lrp::kanR* RE signal represent normalized log transformed ratios they can be directly subtracted without additional normalization between the samples from the two genotypes. The RSE can be interpreted as how much more enrichment with the monoclonal antibody over the purified DNA is obtained when Lrp is present in the WT genotype as compared to when Lrp is not present in the *lrp::kanR* genotype. For each experimental condition in this paper we generated two Lrp WT replicates and two *lrp::kanR* replicates. The Lrp WT and *lrp::kanR* samples are not paired; therefore, we took each combination of a WT Lrp RE replicate and a *lrp::kanR* RE replicate to generate a raw subtracted enrichment signal representing the Lrp WT - *lrp::kanR* signal. This results in four possible RSEs for each condition and time point (i.e. WT rep1 - KO rep1, WT rep2 - KO rep1, WT rep2 - KO rep1, WT rep2 - KO rep1). We next converted each of the RSE scores to a robust Z-score so that enrichments between different experimental conditions could more easily be interpreted on a universal scale. For each replicate pairing, the raw subtracted Lrp enrichment signals were converted to robust Z-score estimates (RZ) using the following formula:

$$RZ_{(n)} = \frac{RSE_{(n)} - \text{median}(RSE)}{\text{median}(\left|RSE_{(n)} - \text{median}(RSE)\right|) \times 1.4826} \tag{2.3}$$

Here, the 1.4826 is a standard scaling factor used to convert the Median Absolute Deviation (MAD) in the denominator into an estimator for the standard deviation under the assumption that the values follow a normal distribution [184, 185]. This allows the RZ to be treated as a proper Z-score. The RZ replicates were then averaged to generate a final occupancy signal for visualization and estimates of the ChIP signal at a peak summit. Reproducibility of both the RE and RSE for each replicate can be seen in Figure 2.18A.

2.6.7 Determination of high-confidence Lrp binding sites

In order to determine regions of high-confidence Lrp binding we required three criteria for Lrp enrichment to be satisfied: 1. The enrichment must be technically reproducible. 2. The enrichment must be above the input background. 3. The enrichment must be biologically reproducible. The following paragraphs detail how each of these criteria were determined.

2.6.8 Assessment of technical reproducibility of Lrp enrichment

To assess the technical reproducibility of the Lrp enrichment, we used custom python scripts to sample with replacement from the aligned reads separately for each paired extracted and input sample. The RSE for each of the four possible subtracted Lrp WT vs. *lrp::kanR* replicates was calculated, as described above in Section 2.6.6, for each of 1000 bootstrap replicates. To test for technically reproducible enrichment, we considered a null hypothesis that the RSE is normally distributed centered at 0. A Z-score for each location n was then determined as follows:

$$Z_{(n)} = \frac{RSE_{0(n)}}{\text{median}(\left|RSE_{B(n)(m)} - \text{median}(RSE_{B(n)})\right|) \times 1.4826} \tag{2.4}$$

Where RSE_0 is the unsampled dataset and RSE_B represents the bootstrap replicates for which $m = 1 : 1000$. The resulting Z-score was converted to a p-value using a one-sided Z test through the scipy.stats normal cumulative distribution function [187]. These p-values were FDR corrected using the procedure described by Benjamini and Hochberg [188]. A region was considered to be technically reproducible if its q-value was less than 0.001.

2.6.9 Assessment of Lrp-specific enrichment

To assess enrichment of ChIP signal above the input background and to differentiate from off-target antibody enrichments seen in pulldowns using the *lrp::kanR* strain, an RZ score (see eq. 2.3 above) was calculated for each of the four possible combinations of WT-*lrp::kanR* replicates, yielding positive signal only when the WT pulldown value was substantially above that of the *lrp::kanR*

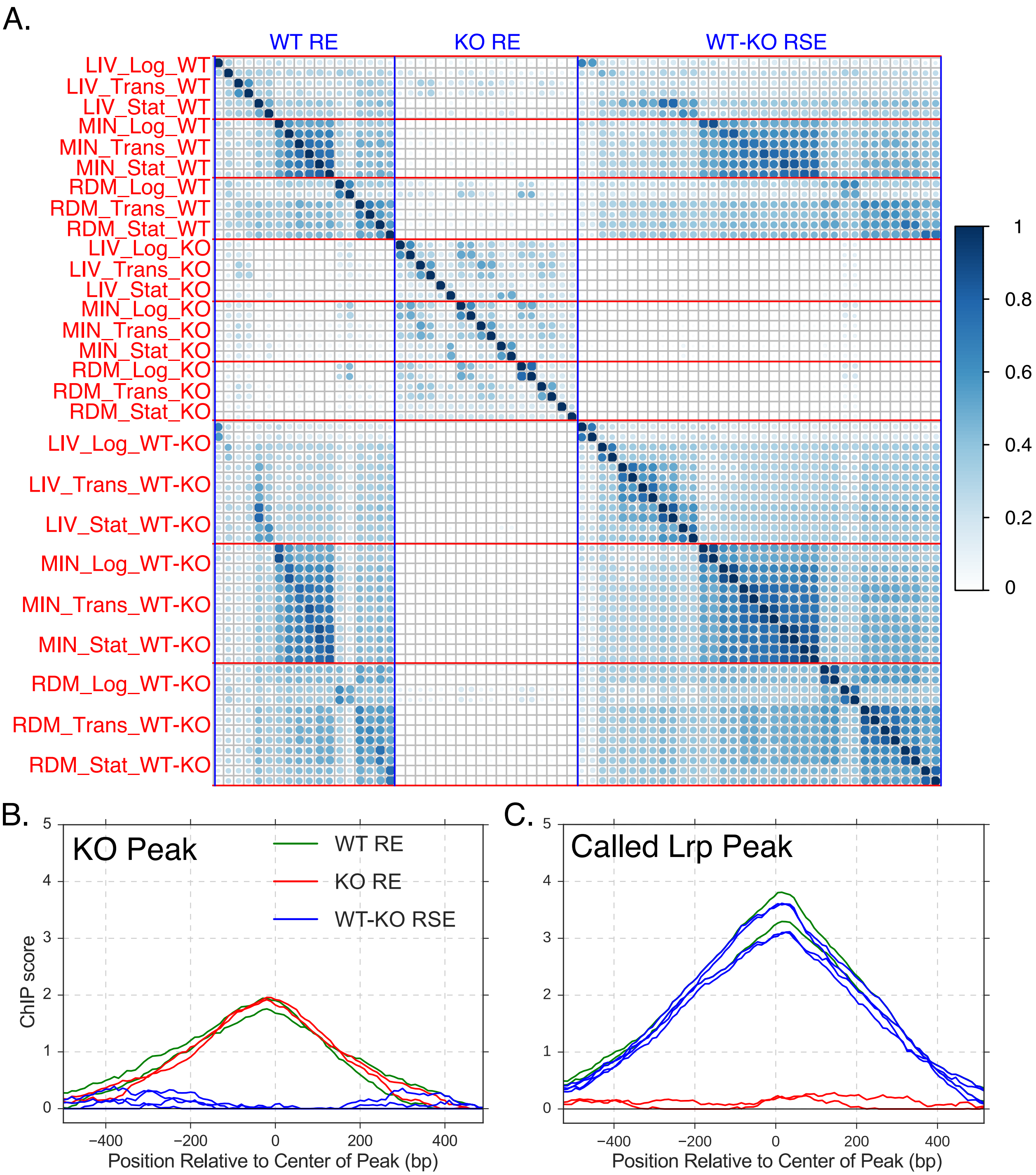

Figure 2.18: Lrp ChIP-Seq data is highly reproducible. (A) Heatmap displaying the similarity between replicates based on shared locations in the highest 2% of signal in each replicate as quantified by the Jaccard statistic $(\frac{A\cap B}{A\cup B})$. Replicates for each WT Raw Enrichment(WT RE), *lrp::kanR* 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 [186]. (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.

signal. We then tested for enrichment of the RZ score above the median signal for that track through the use of a one-sided Z-test using scipy.stats normal cumulative distribution function and FDR correction of the resulting p-value to a q-value. To be considered enriched above background, a region was required to have an enrichment q-value less than 0.001.

2.6.10 Assessment of biological reproducibility

To assess the biological reproducibility of each region n, the irreducible discovery rate [189] was calculated for each data point between the RSE signals for each possible combination of the four Lrp WT - lrp::kanR combinations for each condition and time point (i.e. WT1 - KO1 and WT2-KO2; WT1-KO2 and WT2-KO1). Starting parameters for the IDR calculation for each condition included $\mu = 0.0$, $\sigma = 1.4826$, $\rho = 0.1$ and an associated weight based on the estimated number of bound Lrp octamers for each nutrient condition x as determined in [146]:

$$W = \frac{250\text{bp} \times L_x}{10\text{bp}} \times 463968 \tag{2.5}$$

Where $L_{min} = 684$, $L_{LIV} = 616$, $L_{RDM} = 188$. A region was considered to be biologically reproducible if the FDR-corrected IDR q-value for regions passing the previous 2 filters was less than 0.01 for both combinations of RSE replicates.

2.6.11 Combining enrichment and reproducibility into final peaks

Final peaks were determined if a region n passed the biological reproducibility filter (using each possible pair of RSE signals), and at least one of the four subtracted replicate combinations passed both the technical and enrichment filters (using each possible RSE signal). Adjacent passing regions were consolidated into one region if they were within 30 base pairs. The applied cutoffs and other thresholds were confirmed to be reasonable through manual inspection of called peaks and candidate peaks that narrowly missed one or more cutoffs. An example peak in comparison to a non-Lrp-specific peak can be seen in Figure 2.18B and C.

2.6.12 Preprocessing of RNA-Seq data

Similar to the ChIP-Seq reads, sequencing adapters were removed from all sequences using CutAdapt version 1.8.1 [177] with parameters –quality-base=33 -a AGATCGGAAGAGC -A AGATCGGAAGAGC -n 3 -m 20 –mask-adapter –match-read-wildcards. Low quality reads were trimmed with Trimmomatic version 0.32 [178] using the parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20. The quality of the raw and preprocessed fastq files was assessed using FastQC version 0.10.1 [179] and MultiQC version 1.2 [180]. The number of raw and surviving reads for each sample are described in Supplemental File 3.

2.6.13 Filtering highly abundant RNAs from analysis

In some, but not all, of our samples as much as 70% of our RNA-seq reads were ribosomal reads or the highly abundant RNA products from *ssrA* and *ssrS* (Supplemental File 3). To filter highly abundant RNA reads and thus avoid having variations in ribosome depletion efficiency interfere with proper normalization, all RNA-seq reads were aligned using bowtie2 version 2.1.0 [182] to the same ATCC 47076-modified version of the U00096.2 genome used for the ChIP-Seq data. The following parameters were used for bowtie2: -q –end-to-end –very-sensitive -p 5 –phred33 –dovetail. The subsequent alignments were parsed for reads that overlapped with ribosomal reads in a strand specific manner using custom python scripts. New fastq files were written that only included reads that did not overlap ribosomal reads, and these files were used for downstream gene expression analyses. In all replicates at least two million reads survived this final filter with the smallest size replicate containing 2.6 million reads after filtering (Supplemental File 3).

2.6.14 Gene-centric quantification of RNA-Seq data

Gene-centric quantification of RNA expression for all samples was performed using kallisto version 0.43.0 [190] with the arguments: quant -t 4 -b 100 –rf-stranded. The appropriate transcriptome file needed for alignment was created through converting the GeneProductSet dataset from RegulonDB version 9.4 [48] to the appropriate ATCC 47076 coordinates and input file format for kallisto using custom python scripts.

2.6.15 Determination of Lrp-dependent changes in transcription

To determine Lrp-dependent changes in transcription, we used kallisto's companion post-processing data analysis software, sleuth [191] to model the transcript abundance for each condition and time point. We tested for differential expression between the WT and *lrp::kanR* strains separately for each condition and time point by using a Wald test on the genotype term of the simple model: transcript abundance $\sim$ genotype; here the *lrp::kanR* is the baseline condition. Additionally, we used the bootstrapped read counts from Kallisto to calculate the average WT to *lrp::kanR* expression. We first normalized the count k for gene i using a scaling factor for each replicate j as adapted from equation 5 in [192] and shown below:

$$s_j = e^{\text{median}(k_i - \frac{1}{N}\sum_{i=1}^{N} \log k_i)} \tag{2.6}$$

Expression for gene i in replicate j is thus:

$$\text{expr}_i = \frac{k_i}{s_j} \tag{2.7}$$

We then calculated the $\log_2$ expression ratio between WT and *lrp::kanR* as below:

$$\log_2(\text{expr ratio}) = \text{mean}(\log_2(WT_1), \log_2(WT_2)) - \text{mean}(\log_2(KO_1), \log_2(KO_2)) \tag{2.8}$$

Transcripts that passed both an FDR corrected p-value of less than 0.05 and a $\log_2$ expression ratio magnitude of greater than $\log_2(1.5)$ were considered to have a significant Lrp-dependent RNA expression change under that condition.

To obtain a maximally conservative credible interval on the $\log_2$ expression ratio we calculated the 95% credible interval on the $\log_2$ ratio of each of the four possible WT replicate to *lrp::kanR* replicate pairs across all 100 bootstrap replicates performed by kallisto. We then chose the minimum of the minimum credible intervals of all possible pairs and the maximum of the maximum credible intervals of all possible pairs to report in each of our RNA-seq plots. This credible interval is on average two times larger than a credible interval obtained from bootstrap replicates of the average expression ratio and best represents the true uncertainty of each ratio.

2.6.16 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 2.19A). In addition, we tested the antibody for use in Western blotting (Figure 2.19B). We also confirmed that the antibody did not bind the oligomerization interface by observing bands corresponding to Lrp octamers and hexadecamers in native Western blots (data not shown).

2.6.17 Filtering of genes into Lrp-dependent categories

For gene target filtering, we established four categories through a two-level filtering scheme (Figure 2.6A). 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 any overlapping high confidence Lrp binding site, as de-

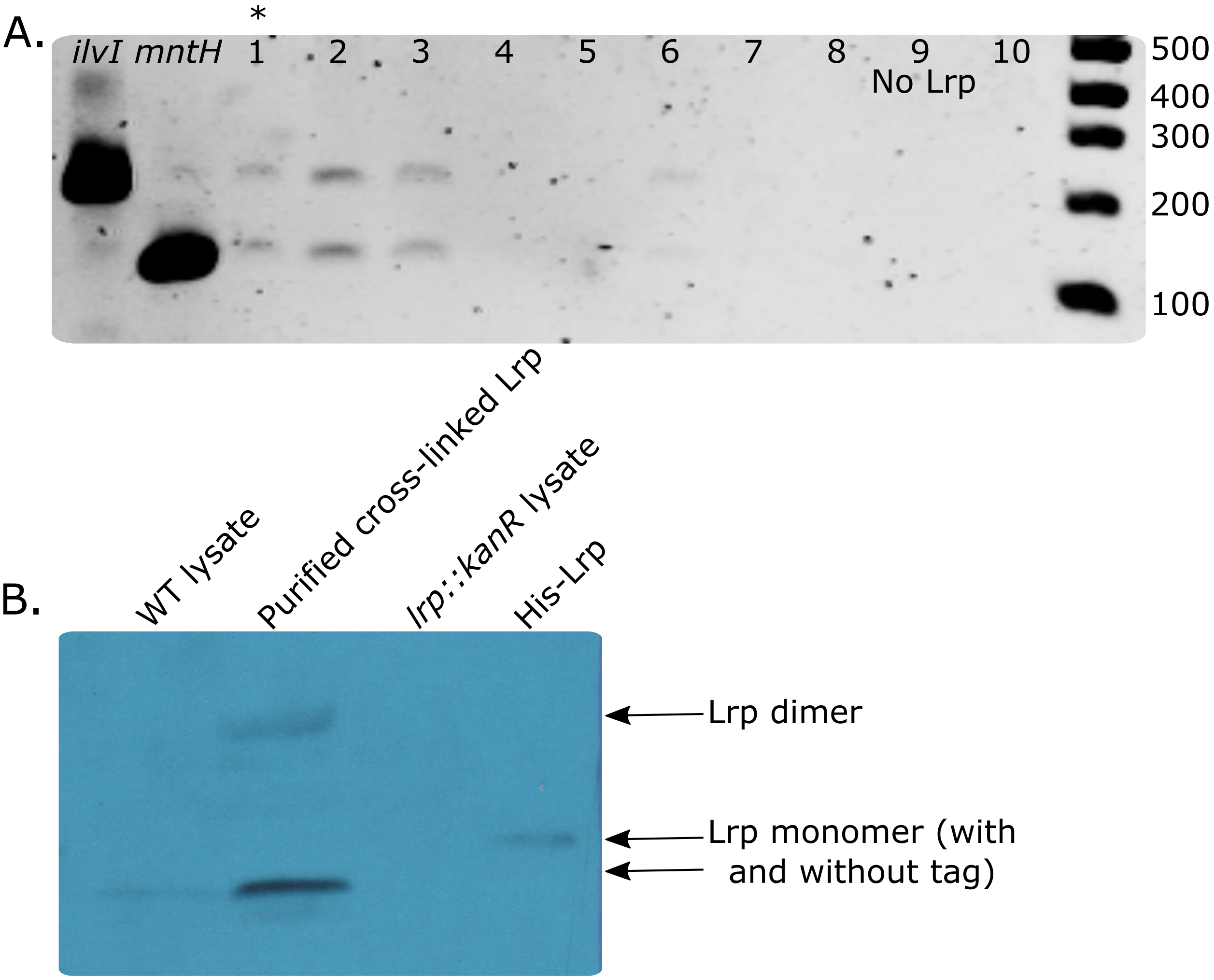

Figure 2.19: Lrp antibody does not interfere with DNA binding and is specific for Lrp. (A) Pull down assay to test ability of antibody to bind DNA-bound Lrp. The first two labeled lines show the expected size of the band for pull down of the specific (*ilvI*) and non-specific (*mntH*) DNA-fragments. Lanes 1-8 are candidate anti-Lrp antibodies. Lane 9 is a negative control with antibody but no Lrp to illustrate no off-target DNA binding to the antibody alone. 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. Lrp's ability to bind DNA non-specifically is strongly evident. (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. Note that the purified Lrp is at much higher concentration than the Lrp concentration in crude cell lysate.

fined above, within the regulatory region (defined as 250 bp upstream and downstream from the annotated transcription start site (TSS; annotations from RegulonDB [48])). If multiple TSSs were annotated for a gene, the regulatory region included 250 bp upstream of the most distal TSS and 250 bp downstream of the most proximal TSS. For unannotated TSSs present within the RegulonDB PromoterSet, we assigned the TSS to the nearest downstream gene within 500 bp based on the ORF definitions in RegulonDB's GeneProductSet and any TSS that fell outside this range was left unassigned. This automated TSS assignment is consistent with those used in other, similar applications (e.g., [193]).

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 poised target (no RNA expression change but Lrp binding), or unconnected to Lrp (neither RNA expression change or Lrp binding). For the additional classification of poised targets in Figure 2.7A, we further divided the poised targets into four subcategories poised, poised_nearby_direct, poised_uncertain_direct, and poised_unexplained. Poised targets were considered true poised targets if, in any other condition, they transitioned to a direct target. Of the genes that do not fit into the true poised classification if, within the same condition, a nearby gene (with a promoter region within 1000 bp of the poised gene) was a direct target then they were classified as a poised_nearby_direct. Failing those two classifications, genes for which the conservative credible interval on the $\log_2$ expression values (described in Section 2.6.15) spanned our $\log_2(1.5)$ ratio biological cutoff were considered poised_uncertain_direct since we do not have enough information in our data to definitively say that the RNA expression is not impacted by Lrp in that condition. Finally, any poised gene not falling in the above classifications was considered a poised_unexplained gene, representing genes where Lrp is binding at the promoter but never regulates the transcription levels of the gene under the conditions studied here.

We also subcategorized the Lrp indirect genes into three categories, indirect, indirect_operon_direct, and indirect_peak_in_cds (Figure 2.7B). Some genes classified as indirects were the result of alternative TSSs within an operon, for which Lrp binds the TSS of the first gene in the operon. Genes that fall under this category were considered indirect_operon_direct. Additionally, some genes had a called Lrp peak that overlapped within the coding region of the gene, these genes were classified as indirect_peak_in_cds. All other indirect genes were classified as truly indirect.

For comparing enrichment of Lrp targets with σ factor or transcription 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 σ factors or transcription factors, we corrected for multiple hypothesis testing with the python statsmodels.sandbox.stats.multicomp.multipletests implementation of the Benjamini-Hochberg method [188, 194]. All plots except where noted were created using ggplot2 [195] or Matplotlib [196]. All genomic features above plots were created using the DNA features viewer python library (https://github.com/Edinburgh-Genome-Foundry/DnaFeaturesViewer).

2.6.18 Accession numbers

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.

2.7 Acknowledgements

The authors wish to thank Prof. Robert Blumenthal for his advice and suggestions on this project. This research was supported by NIH R00 GM097033 (to P.L.F.), NIH R35 GM128637 (to P.L.F.), and startup funds from the University of Michigan (to P.L.F.). G.M.K. is supported by the Cellular Biotechnology Training Program (T32-GM008353) and a one-time Rackham Research Grant from the University of Michigan. M.B.W. is supported by a National Science Foundation Graduate Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.7.1 Author Contributions

G.M.K. and P.L.F. planned the experiments, G.M.K. performed the experiments, M.B.W. performed the sequencing analysis, M.B.W., G.M.K., and P.L.F. analyzed the data, M.B.W., G.M.K., and P.L.F. prepared the manuscript, and P.L.F. supervised the work.