Quantitative Dissection of the Allosteric and Sequence-Dependent Regulatory Genome in *E. coli*.

Thesis by
Nathan Maurice Belliveau

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Bioengineering

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended December 14, 2017

ACKNOWLEDGEMENTS

It is no question that I owe my wonderful time at Caltech to Professor Rob Phillips. His breadth of interest and never-ending enthusiasm about the science have been a constant source of inspiration for me. I also believe his constant questioning of assumptions and skepticism throughout my journey here have led to more interesting results and made me a better researcher. I hope more students continue to stumble upon the group as I did.

I would especially like to extend my gratitude to the members of the group. The socialist experiment that is Chapter 2 was truly a team effort and I enjoyed working with Manuel, Tal, Griffin, and Stephanie throughout the project. The positive atmosphere within the group has made my many hours stuck in the lab more manageable. Manuel, Griffin, and Soichi in particular were constant sounding boards to my many random thoughts and questions, and I appreciate their willingness to listen and provide feedback. In addition, Justin Kinney, who is at Cold Spring Harbor, has provided expert guidance in the Sort-Seq efforts of Chapter 3 and 4, and I have appreciated the counter perspectives that he was able to offer throughout that work.

I also had the opportunity to join Rob and other members in the group on many of his teaching engagements around the world and I have been happy to tag along whenever I could. I'm deeply indebted to Heun Jin Lee, who taught me everything I know about optics. While I'm sure I've only managed to absorb but a small fraction of his wealth of knowledge. I have done my best to pass that knowledge on accurately in the courses that I've been involved in.

Lastly, and easily most important, I acknowledge my family for their support, in particular, my wife, Christie, and mother, Marie, who were the constant ears of my hopes and frustrations over these last five or so years.

ABSTRACT

Transcriptional regulation of gene expression is one of the most ubiquitous processes in biology. But while the catalog of bacterial genomes continues to expand rapidly, we remain ignorant about how almost all of the genes in these genomes are regulated. Given a gene, we would like to know the transcription factors that regulate them, how strongly they bind to the DNA, and how they interact with RNA polymerase and other external signals to control gene expression.

The theoretical framework of statistical thermodynamics provides us with a useful way to quantitatively describe the different mechanisms of regulation. One of the important ways genes are regulated is through external signals. To that end, we begin by presenting a general theory of allosteric transcriptional regulation using a statistical mechanical formulation of the Monod-Wyman-Changeux model. Allostery is central to many biological processes, and in the context of gene regulation, it describes a transcription factor's conformational changes that modulate activity in response to external signals. We rigorously test this model using the ubiquitous simple repression motif with the transcription factor LacI in *Escherichia coli*. Our model not only accurately captures the allosteric response of these strains but also enables us to derive analytic expressions for key phenotypic properties such as the available dynamic range in gene expression.

We then move to consider the consequences for gene expression of the regulatory sequences themselves. Understanding how regulatory sequence maps to function remains a difficult problem in biology. Here we apply a massively parallel reporter assay, Sort-Seq, to build models that describe the sequence-dependent binding energies of transcription factors and RNA polymerase to DNA. By coupling such models to our thermodynamic models of regulation, we construct a genotype to phenotype map that predicts gene expression as a function of regulatory sequence. Here we demonstrate this approach by designing roughly 30 mutant LacI binding site sequences, and accurately predict expected levels of gene expression as a function of these sequences. We also show how such regulatory sequences can be designed to optimize the inducible response of LacI in the context of the allosteric simple repression motif considered above.

Given any particular promoter across a bacterial genome, we would like to be able to build genotype to phenotype mappings that predict gene expression more broadly. However, much of the quantitative insight available on transcriptional regulation relies on careful and extensive work of only a few model regulatory systems such as LacI that was considered above. Here we develop an approach, through a combination of massively parallel reporter assays, mass spectrometry, and information-theoretic modeling that can be used to dissect bacterial promoters in a systematic and scalable way. We demonstrate this method on both well-studied and previously uncharacterized promoters in *E. coli*. In all cases we recover nucleotide-resolution models of promoter mechanism and open up the possibility of exhaustively dissecting the mechanisms of promoter function in *E. coli* and a wide range of other bacteria.

PUBLISHED CONTENT AND CONTRIBUTIONS

Belliveau, N. M., Barnes, S. L., Ireland, W. T., Jones, D. L., Sweredoski, M. J., Moradian, A., Hess, S., Kinney, J. B., and Phillips, R. (2017). A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *bioRxiv*. DOI: https://doi.org/10.1101/239335.
N. M. B. helped conceive of the project, performed experiments, analyzed results, prepared data, and co-wrote the manuscript.

Razo-Mejia, M., Barnes, S. L., Belliveau, N. M., Chure, G., Einav, T., Lewis, M., and Phillips, R. (2017). Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. *bioRxiv*. DOI: https://doi.org/10.1101/111013.
M.R.M., S.L.B., N.M.B., G.C., T.E. contributed equally to the work.
N. M. B. helped conceive of the project, co-performed experiments and data analysis, and co-wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

Since the technological developments that led us to the first draft of the human genome in 2001 (Lander et al., 2001), our catalog of sequenced genomes has expanded at an incredible pace (Loman and Pallen, 2015; Land et al., 2015). While this global effort took 13 years and almost 3 billion dollars, today we could accomplish it for about 1,000 dollars and in roughly a day (Levy and Myers, 2016). As further illustration of this pace, between 2010 and 2014 a bird species from each major avian clades was sequenced (Jarvis et al., 2014), and larger scale intiatives such as the G10K project to sequence species in every vertebrate genus are well under way (Koepfli et al., 2015).

The advent of so-called next-generation sequencing technologies (and now third-generation technologies that offer much longer read lengths) has revolutionized how we perform research in biology (Goodwin et al., 2016). It has led to the development of over one hundred sequencing-based methods (Pachter, 2013) to study almost any aspect of biology, from chromosomal structure (e.g., Hi-C-Seq (Belton et al., 2012)), transcriptional regulation (e.g., Sort-Seq (Kinney et al., 2010) and ChIP-Seq (Park, 2009)) and RNA structure (e.g. SHAPE-Seq (Loughrey et al., 2014)) to name but only a few of these techniques. Many of these approaches have become important tools to biophysicists due to their ability to provide a quantitative measure of biological function. It is also having a pervasive impact on society that is only beginning to be realized; in the form of personalized medicines that may significantly improve the effectiveness of current treatment options for many diseases (Rabbani et al., 2016).

With these advances, however, the accumulation of sequence information easily outpaces our understanding of all the information contained within a genome (Galperin and Koonin, 2010). While it is quite easy, for example, to identify the open reading frames that code for different genes, we are still unable to discern the many intricate regulatory sequences that are necessary for proper biological functioning of a cell (Münch et al., 2003; Cipriano et al., 2013; Kılıç et al., 2013). With regard to gene regulation, we owe much of our initial insights from the pioneering work of Jacob, Monod, and Lwoff (Jacob, 2011; Jacob and Monod, 1961), which has spurred a substantial amount of research in biology toward understanding the variety of regulatory mechanisms involved.

At the core of this dissertation is an aim to develop quantitative descriptions of the mechanisms of transcriptional regulation that will allow us to describe and predict the input-output responses in gene expression across a bacterial genome. In **Chapter 1** we begin by first reviewing the processes associated with gene expression (i.e., the central dogma of molecular biology) and regulation. We then describe how ideas from statistical mechanics can be used to quantitatively describe the regulatory mechanisms of transcription. We end with a discussion of the current state of regulatory knowledge in *E. coli* to highlight our still limited understanding of the regulatory code which motivate the efforts of Chapter 4.

In **Chapter 2** we extend previous theoretical and experimental efforts of regulation to account for the phenomenon of allostery. Allosteric proteins respond to changes in their environment by binding to ligands or other effector molecules and are central to most metabolic and signal-transduction pathways (Fenton, 2008; Motlagh et al., 2014). While it is common to apply phenomenological models such as Hill functions to describe this response, such models consist of lumped parameters that are a conceptual dead-end (Kuhlman et al., 2007). Instead, we develop a general quantitative framework with which to describe allosteric transcriptional regulation with the Monod-Wyman-Changeux (MWC) model of allostery (Monod et al., 1965), that instead more accurately recognizes that proteins as existing in two of more structural states that drive this allosteric response. We use the model to produce a set of predictions that we then test experimentally in the context of Lac repressor (LacI) and the simple repression motif (Bintu et al., 2005a).

We then move to consider the consequence of the regulatory sequences themselves on gene expression. In order to connect regulatory sequence to biophysical mechanisms of regulation across the *E. coli* genome, we propose applying a massively parallel reporter assay, Sort-Seq (Kinney et al., 2010), to characterize the regulatory DNA. This approach will allow us to identify transcription factor binding sites and enable the quantitative dissection of individual promoters with base pair resolution. In **Chapter 3** we begin to test this strategy by performing extensive characterization of the sequence specificity of LacI, and use Sort-Seq to generate energy matrix models that describe the *in vivo* interaction energy between LacI and its DNA binding site sequence. Here we are able to show the validity of our models and demonstrate that we can use the regulatory sequence as a tunable parameter to modify gene expression, which we further show has direct consequences on gene expression characteristics such as dynamic range, saturation, and leakiness (Martins and Swain, 2011).

Finally, with the preceding chapters demonstrating the type of quantitative rigor we intend to apply to regulation more broadly across a bacterial genome, in **Chapter 4** we then develop a systematic approach to quantitatively decipher the regulation of any promoter more generally. Here we first apply Sort-Seq across different bacterial promoters to uncover the functional binding sites where transcription factors bind to regulate gene expression. Using DNA affinity chromatography and mass spectrometry we then identify the transcription factors that bind these sites, and apply information-theoretic modeling to infer energy matrix models of binding by each transcription factor. We demonstrate the validity of the approach by first applying it to the well-characterized promoters of *lacZYA*, *relBE*, and *marRAB*. We then apply it to uncover the regulatory architectures for the promoters of *purT*, *xylE*, and *dgoRKADT*, whose regulation was previously unknown.

1.1 Transcription and transcriptional regulation

The central dogma of molecular biology describes the order in which genetic information flows from genomic DNA to produce proteins that will then perform biological functions (Figure 1.1) (Crick, 1970). The process begins when RNA polymerase *transcribes* a gene's coding sequence on the DNA into a single-stranded mRNA template. A ribosome then *translates* the gene coded on the mRNA into a protein that consists of a polypeptide of amino acids. While the nucleotides on the DNA map directly to those on mRNA (A→A, C→C, G→G, T→U), each amino acid that makes up a protein will map to triplet sets of nucleotides on the mRNA known as codons.

To begin transcription, RNA polymerase must first bind the DNA in the upstream region that precedes a gene, known as the gene's promoter. Figure 1.2 shows a general schematic of the transcription process. During growth in nutrient-rich conditions, the core RNA polymerase enzyme (which consists of five subunits, $\beta\beta'\alpha_2\omega$) recognizes particular DNA binding sites by forming a complex with the primary sigma factor, RpoD (also written as σ_{70} or σ_D), which prefers the consensus -35 and -10 sequence TTGACA$(N)_n$TATAAT (where $(N)_n$ is a spacer sequence, with n optimally 17 bp) (Feklístov et al., 2014). The core RNA polymerase may also bind to other sigma factors that recognize different promoter sequences and allow the cell to respond to changes in the environment or exposure to adverse conditions (Gruber and Gross, 2003; Maeda, 2000). Binding of RNA polymerase to the DNA can be

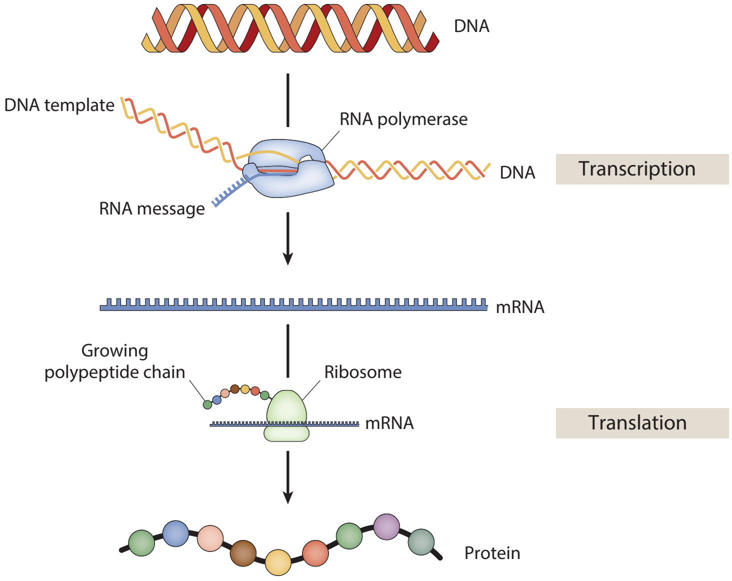


Figure 1.1: **Central dogma of molecular biology.** The pathway from DNA to protein. When a gene is on, RNA polymerase translocates along the DNA and transcribes the coding sequence into an mRNA template. This mRNA template is then read by the ribosome, which produces a polypeptide chain by stringing together amino acids according to triplet sets of nucleotides known as codons specified by the mRNA. Obtained from reference (Phillips, 2015).

enhanced by the presence of A/T rich sequences called ‘UP elements,’ that are found upstream of the -35 site (see Figure 1.2B). This is due to recognition by the flexibly tethered α-subunit C-terminal domain (αCTD) of the core RNA polymerase enzyme (Browning and Busby, 2016; Murakami and Darst, 2003). Upon binding to the DNA, RNA polymerase will then proceed through several well-characterized steps to separate the two strands of DNA and begin transcription. Here, the polymerase first transitions from a closed complex with the DNA into a stable open complex where transcription can then begin to generate the mRNA template (Browning and Busby, 2016; Murakami and Darst, 2003).

At each step of the central dogma, mechanisms exist that will regulate expression. For example, other small RNA molecules can interfere with translation by binding mRNA through direct complementarity of the RNA sequence, or in complex with RNA chaperone proteins such as Hfq (Van Assche et al., 2015). The mRNA molecule can itself form thermodynamically stable secondary structures that influence mRNA degradation or accessibility by the ribosomes that perform translation (Salis et al., 2009). Even after the protein is synthesized, it may contain certain amino acid sequences that are targeted by cellular proteases that will help degrade the protein

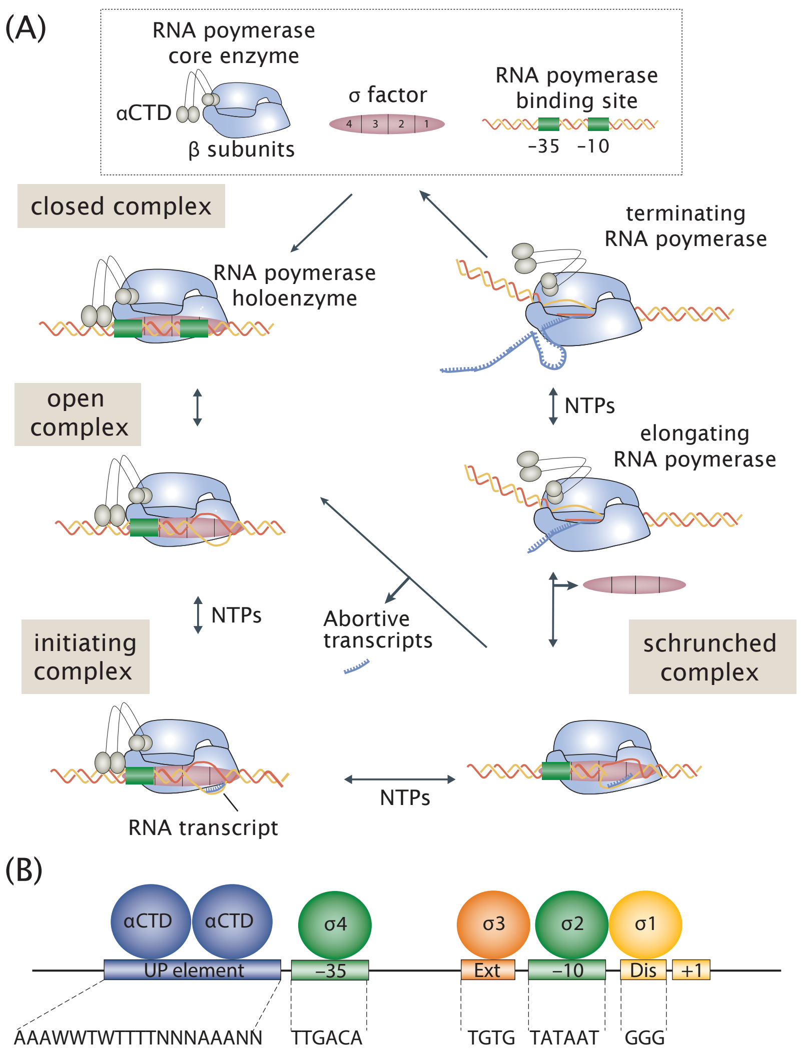


Figure 1.2: **Transcription of bacterial genes by RNA polymerase.** (A) The bacterial transcription cycle. The RNA polymerase holoenzyme comprises the RNA polymerase core enzyme, $\beta\beta'\alpha_2\omega$, and a sigma factor, which interact with a binding site on the DNA to form the closed complex. A transition is made to an open complex by unwinding the DNA in the region of the transcription start site. Addition of nucleoside triphosphates (NTPs) then allows transition to the initiating complex to begin synthesis of the RNA transcript (initially through a process termed scrunching). Abortive initiation cycles may result in small RNA fragments, but otherwise the RNA polymerase will enter the elongation phase. At this stage, the sigma factor is generally lost, and elongation of the RNA proceeds through addition of NTPs and until reaching a transcription terminator on the DNA where the RNA is released and RNA polymerase dissociates from the DNA. (B) The consensus DNA sequence recognized by RNA polymerase. Although natural promoters are quite variable, many bacterial promoters contain an UP element, a -35 site, an extended -10, a -10 site, and then a discriminator sequence. The figure shows the regions of the sigma factor and the carboxy-terminal domain of the α-subunit of RNA polymerase (αCTD) that contact the DNA. This figure was adapted from an excellent review on transcription initiation in bacteria (Browning and Busby, 2016).

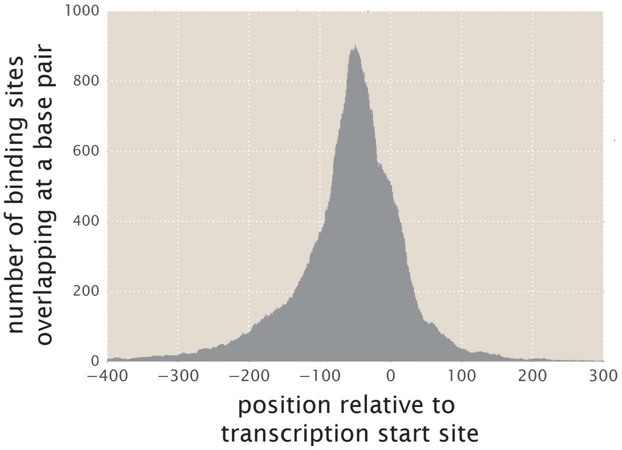


Figure 1.3: **Summary of known transcription factor binding sites of the *E. coli* genome.** The locations of all annotated transcription factor binding sites on RegulonDB were used to generate a histogram of their locations on the genome. Each binding site is listed relative to a transcription start sites for the promoter where it binds. Figure was adapted from reference (Rydenfelt et al., 2014).

under certain physiological conditions (Gur et al., 2011). Despite this variety of control mechanisms that exist, regulation at the level of transcription is arguably the among the dominant ways in which cells across all domains of life regulate their expression. Here, cells decide when a gene is ‘switched’ on or off in large part by proteins called transcription factors that bind the DNA and modulate the activity of RNA polymerase at each promoter (Browning and Busby, 2016).

If we take a survey of the location of binding sites where transcription factors bind the genomic DNA, we find them distributed near the transcription start site where RNA polymerase begins transcription (see Figure 1.3). The transcription factors that bind these sites can be categorized as either repressors (preventing transcription) or activators (enhancing transcription) (Seshasayee et al., 2011). It is interesting to note however that even LacI, the canonical example of a repressor, can also be converted into an activator (Labow et al., 1990), so this categorization is somewhat fluid and may depend on context.

Repressors prevent transcription, where they either prevent binding by RNA polymerase through through steric hindrance or by modulating the activity of activators (Browning and Busby, 2016). Repressor binding sites may also lie several hundred base pairs away from the transription start site, and binding by repressors at these sites can further modulate expression through the formation of DNA loops (Cournac and Plumbridge, 2013; Garcia et al., 2007). Examples are found in the promoters of *lac* (Boedicker et al., 2013b; Boedicker et al., 2013a), *gal* (Mandal et al., 1990), and *araC* (Martin et al., 1986; Schleif, 2010) in *E. coli*.

Activators generally bind upstream of RNA polymerase, where they enhance transcription through interaction with the αCTD domain of RNA polymerase (class I activation), or directly with the sigma factor (class II activation) (Lee et al., 2012). It is interesting to note that if the activator binding site is slid along the DNA just upstream of the RNA polymerase binding site, a periodic pattern is observed in the extent of activation (experimentally shown using synthetic constructs (Ushida and Aiba, 1990; Gaston et al., 1990)). By noting that a full turn of the DNA helix requires about 10.5 base pairs, this is explained by the need for both activator and RNA polymerase to share the same face of the DNA and serves to highlight the physical mechanism underlying this process.

It is common for promoters to have binding sites for both activators and repressors that modulate transcription and we usually refer to this as the promoter's regulatory architecture. Figure 1.4 shows the architecture of the *lac* operon, to which we owe much of our early insights into transcriptional regulation (along with the regulation of the lytic and lysogenic cycles of phage λ (Lewis, 2011)). The promoter contains three *lac* repressor (LacI) binding sites, two of which are shown, and a cyclic AMP receptor (CRP) activator binding site. This combination of repressor and activators sites causes the promoter to exhibit a now classic catabolic switch-like behavior that results in diauxie when *E. coli* is grown in the presence of glucose and lactose sugars (Loomis and Magasanik, 1967; Oehler et al., 1990; Busby and Ebright, 1999). The third LacI binding site is found within the *lacZ* gene and is involved in the formation of DNA loops that were noted earlier.

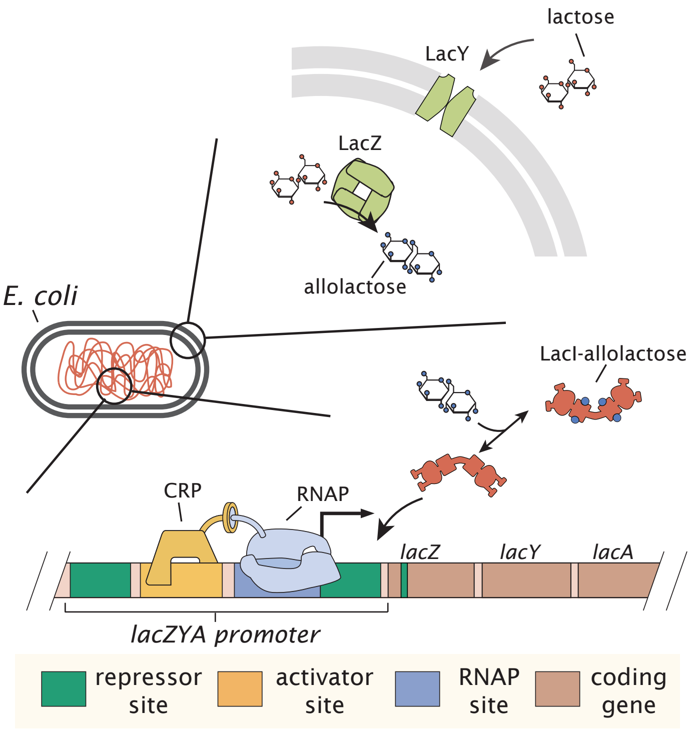


Figure 1.4: **The *lac* operon.** The promoter for the *lac* operon drives expression of *lacZ*, *lacY*, and *lacA*. Transcription by RNA polymerase is regulated: a) repression by LacI which binds at three binding sites (O1 and O2 shown; O3 is within the *lacZ* gene), and b) activation by CRP, which binds upstream of RNA polymerase. After translation, the LacZ protein forms a homotetramer that catalyzes cleavage of lactose to glucose and galactose (and lactose into allolactose). The LacY protein is a membrane protein that allows intake of lactose from the cell's environment. The functional role of LacA is not well known. In the absence of allolactose, the LacI tetramer strongly represses transcription. In the presence of allolactose, LacI is allosterically induced and no longer binds strongly to the LacI binding sites, and transcription can be enhanced by CRP. Note that binding sizes and coding regions are not shown to scale.

1.2 Thermodynamic models

In this dissertation we will rely heavily on a class of models called thermodynamic or statistical mechanical models of gene regulation (Bintu et al., 2005a; Bintu et al., 2005b). These models provide a way to formalize the qualitative descriptions noted above into falsifiable quantitative predictions about how expression will change as biophysical details such as the concentration of each regulatory protein change within the cell. At their most basic assumption, gene expression from a promoter is taken to be proportional to the equilibrium probability that the promoter is occupied by RNA polymerase (Kuhlman et al., 2007; Buchler et al., 2003; Vilar and Leibler, 2003; Bintu et al., 2005a; Garcia and Phillips, 2011; Brewster et al., 2014; Ackers et al., 1982). This might seem a little absurd since cells are definitely out of equilibrium. However, these models have been quite successful in making quantitative predictions about gene regulation. Indeed, due to a separation of times scales for different biological processes, a quasi- equilibrium treatment of regulation is generally valid. In particular, the relevant interactions, such as binding by transcription factors to the DNA, occurs with fast on/off rates relative to the rate of transcription and translation (Moran et al., 2010).

We can derive such models of gene expression by first enumerating all possible states of a promoter and their corresponding statistical weights (Bintu et al., 2005a). Here we briefly consider the simple repression architecture, which we will be used extensively in Chapters 2 and 3. As shown in Figure 1.5, the promoter can be empty, occupied by RNA polymerase, or occupied by a repressor. In addition to the specific binding sites at the promoter, we have assumed that there are N_{NS} non-specific binding sites elsewhere (i.e., on parts of the genome outside the simple repression architecture) where the RNA polymerase or the repressor can bind. Our model explicitly ignores the complexity of the distribution of non-specific binding affinities across the genome, and makes the assumption that a single parameter can capture the energy difference between our binding site of interest and the average non-specific site in the genome background. Thus, $\Delta\varepsilon_P$ represents the energy difference between the specific and non-specific binding for RNA polymerase to the DNA. Likewise, $\Delta\varepsilon_R$ represents the difference in specific and non-specific binding energies for the repressor.

We can now calculate the probability that RNA polymerase is bound to the promoter p_{bound}, which is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}{1 + \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_R} + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}, \tag{1.1}$$

with $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant and T is the temperature of the system. P represents the RNA polymerase copy number per cell, while R represents the copy number of repressor.

Measuring p_{bound} directly is experimentally difficult and we instead measure the fold-change in gene expression, which we define as the ratio of expression in the presence of repressor relative to expression in the absence of repressor (i.e., constitutive expression) (Garcia and Phillips, 2011; Weinert et al., 2014). We will explore this further in Chapters 2 and 3.

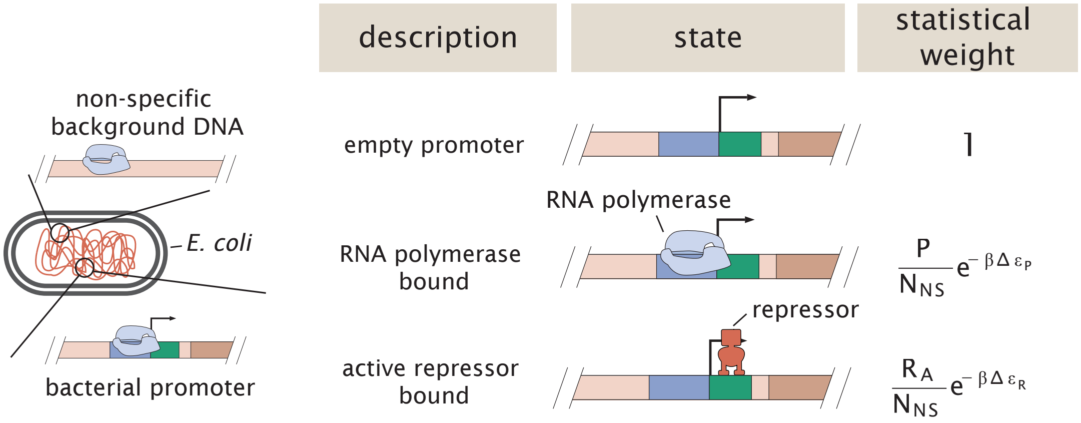


Figure 1.5: **States and weights for the simple repression motif.** There are P RNA polymerase (blue) and a R repressors (red) per cell that compete for binding to a promoter of interest. The difference in energy between a repressor bound to the promoter of interest versus another non-specific positions elsewhere on the DNA equals $\Delta\varepsilon_R$ in the active state and $\Delta\varepsilon_{RI}$ in the inactive state; the P RNA polymerase have a corresponding energy difference $\Delta\varepsilon_P$ relative to non-specific binding on the DNA. N_{NS} represents the number of non-specific binding sites for both RNA polymerase and repressor.

1.3 Allostery

Allosteric control is a molecular mechanism in which a conformational change occurs at one site on a protein in response to binding by a ligand or effector molecule at another distinct site of that same protein (Fenton, 2008; Motlagh et al., 2014). Allostery is central to most metabolic and signal-transduction pathways and examples can be found in a wide variety of cellular processes that include ligand-gated ion channels (Auerbach, 2012), enzymatic reactions (Einav et al., 2016; Velyvis et al., 2007), chemotaxis (Keymer et al., 2006), quorum sensing (Swem et al., 2008), and G-protein coupled receptors (Canals et al., 2012). While first described more than 50 years ago to account for the feedback inhibition that was apparent in the activity of certain enzymes of metabolic pathways (Changeux, 1961; Gerhart and Pardee, 1961; Gerhart and Pardee, 1962; Monod and Jacob, 1961), it continues to be an important area of study (Fenton, 2008). In particular, quantitative models describing the molecular mechanisms that provide this action at a distance remain elusive for most known cases of allostery.

One of the most common and well known examples of allostery comes from transcriptional regulation, where binding of a ligand to a transcription factor changes the conformation of a transcription factor, and reduce (or enhances) its ability to bind

DNA. The elucidation of the *lac* operon was only just completed when the ideas of allostery were begining to take a more concrete form, and those ideas brought better insight into the mechanism of repression by LacI (and the λ repressor in phage λ) (Monod et al., 1963; Changeux, 2013). Here we provide several examples of allosteric gene regulation, and then in Chapter 2 develop an analytical framework for allosteric transcriptional regulation in bacteria through the Monod-Wyman-Changeux (MWC) model of allostery (Monod et al., 1965). While not the focus of Chapters 3 and 4, allostery appears to underlie how several newly identified regulatory architectures respond and modify their regulatory state.

Proteins are capable of being allosteric due to a probability of being in multiple structural conformations. When viewed through the MWC model, binding of an effector is seen to shift the protein's allosteric equilibrium toward another state or conformation available to the protein. In the example of LacI, the repressor binds the DNA and represses expression from the *lac* operon when cells are grown in the absence of lactose. However, in the presence of lactose, cells produce the metabolite allolactose from lactose (shown in Figure 1.4) that binds to LacI and 'induce' the repressor such that it no longer favors binding to its DNA binding site (Lewis, 2011) (see Figure 1.6A). The activator of the *lac* operon (and many other genes across the *E. coli* genome), CRP, is also allosteric. Here the cellular concentration of the nucleotide cyclic AMP influences the conformational state of CRP that enables it to bind DNA and regulate transcription (Schultz et al., 1991; Sharma et al., 2009)

As another example of allosteric regulation, more recently it has also been proposed that intrinsically disordered domains can provide a so-called 'entropic barrier' that may contribute to the conditional cooperativity that is commonly observed in regulation of type II toxin-antitoxin systems (Motlagh et al., 2014; Garcia-Pino et al., 2010; Garcia-Pino et al., 2016) (see Figure 1.6B). The promoters that drive expression of toxin and antitoxin genes commonly contain tandem binding sites where the antitoxin protein can repress its own promoter. In the case of the toxin-antitoxin system Doc-PhD, the repressor (the antitoxin, PhD) contains an intrinsically disordered domain that prevents it from occupying both binding sites (Garcia-Pino et al., 2016). However, in the presence of its cognate binding partner (the toxin, Doc) there is shift in allosteric equilibrium, and the disordered domain gains a structured domain that allows the two tandem repressor sites to be fully occupied and provide repression of the promoter. When the toxin Doc is too abundant, however, each antitoxin PhD will bind two toxin proteins, which again prevents full occupancy of

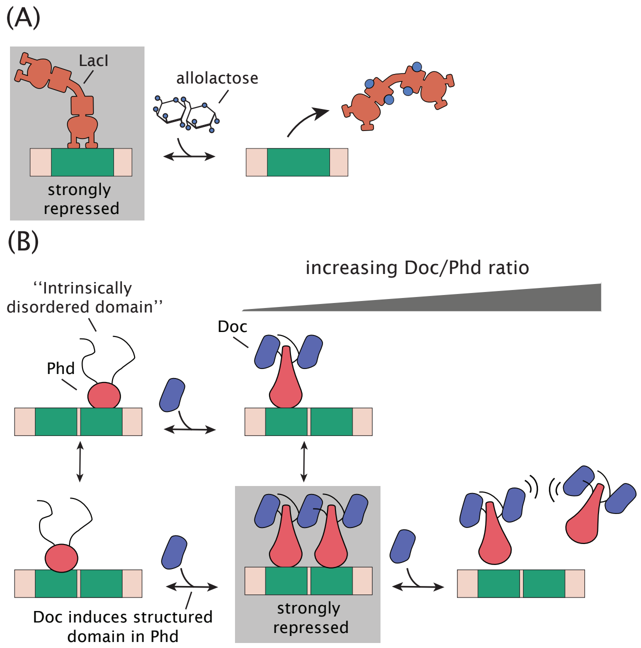


Figure 1.6: **Examples of allosteric regulation in *E. coli*.** (A) Allosteric induction of LacI occurs when allolactose, or the synthetic compound, isopropyl β-D-1-thiogalactopyranoside (IPTG) is present and binds to ligand binding sites on the LacI protein. Each dimer has two binding sites. (B) Doc/Phd toxin–antitoxin system and regulation through conditional coopertivity. In the absence of antitoxin PhD protein, the toxin Doc causes cell arrest due to disruption of translation by ribosomes. As the ratio of antitoxin to toxin increases, the antitoxin binds the toxin and disables its toxicity. The antitoxin also acts as a repressor it provides autoregulation of its promoter. As a toxin-antitoxin complex, it binds more strongly to the DNA. Due to steric hindrance from an intrinsically disordered domain on the antitoxin, strong repression is only expected when the toxin/antitoxin is not too high or low, and two antitoxin dimers are able to bind the DNA. PhD is shown as a homodimer (red) and Doc is depicted as the blue ligand.

the tandem repressor binding sites.

Just as in our example of transcriptional regulation of the simple repression architecture in the last section, we can appeal to statistical mechanics to calculate the probability of the regulatory protein being in any particular conformation. While we will develop the MWC model of allostery more fully in Chapter 2, in Figure 1.7 we show the possible states of such an allosteric regulatory protein under the MWC model, that is applicable to LacI. In contrast to our thermodynamic model example of simple repression, here we assume the repressor exists in two states, an active and inactive state. The number of inducer binding sites will depend on the protein, which in the case of LacI is a repressor dimer with two inducer binding sites (Lewis, 2011).

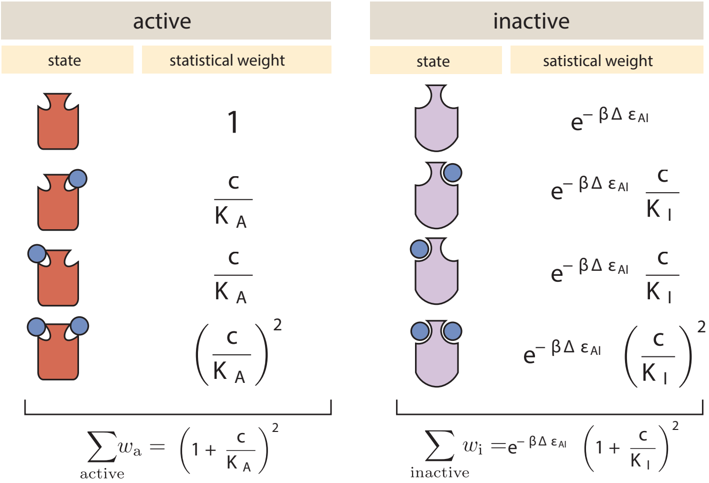

active
state
statistical weight
inactive
state
satistical weight
1
$\frac{c}{K_A}$
$\frac{c}{K_A}$
$\left(\frac{c}{K_A}\right)^2$
$e^{-\beta\Delta\varepsilon_{AI}}$
$e^{-\beta\Delta\varepsilon_{AI}}\frac{c}{K_I}$
$e^{-\beta\Delta\varepsilon_{AI}}\frac{c}{K_I}$
$e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{c}{K_I}\right)^2$
$\sum_{\text{active}} w_a = \left(1+\frac{c}{K_A}\right)^2$
$\sum_{\text{inactive}} w_i = e^{-\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^2$

Figure 1.7: **States and weights for the simple repression motif.** A repressor has an active conformation (red, left column) and an inactive conformation (purple, right column), with the energy difference between these two states given by $\Delta\varepsilon_{AI}$. The inducer (blue circle) at concentration c is capable of binding to the repressor with dissociation constants K_A in the active state and K_I in the inactive state. The eight states for a dimer with two inducer binding sites are shown along with the sums of the active and inactive states.

As another example, the repressor MarR, which represses the *marRAB* promoter, has been found to have four inducer binding sites (Wilkinson and Grove, 2006).

1.4 Status of regulatory knowledge in *E. coli*

Much of the insight we have on transcriptional regulation relies on careful and extensive work of a few model regulatory systems (Daber et al., 2011; Kuhlman et al., 2007; Buchler et al., 2003; Vilar and Leibler, 2003; Ackers et al., 1982). The Phillips group has relied on much of these efforts and used components of the *lac* operon to develop and test models of gene regulation (Garcia and Phillips, 2011; Garcia et al., 2011; Brewster et al., 2012; Boedicker et al., 2013b; Boedicker et al., 2013a; Brewster et al., 2014). While impressive advances in molecular biology have made it possible to map thousands of gene interactions and create genetic networks for a variety of organisms, they still leave us with a regulatory landscape that is qualitative in description. Here we take stock of what is known about regulation in *E. coli*. As we will find, we still remain ignorant to how most genes across the genome are regulated, and this prevents any attempt to begin to write down the types of quantitative models considered so far. This inability motivates much of the work of Chapter 4.

We can begin to take stock of what is known about how genes in *E. coli* are regulated

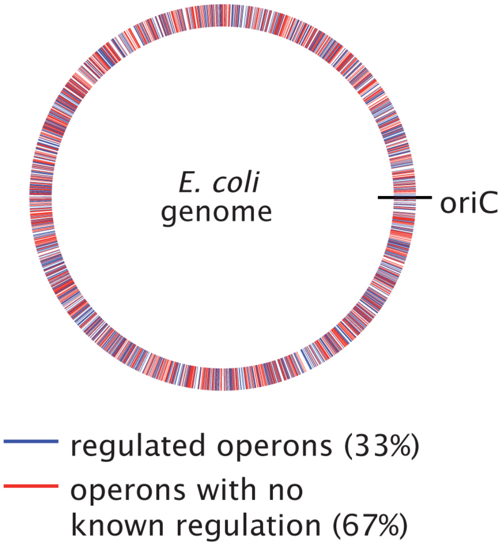


Figure 1.8: **Identification of operons in *E. coli* with and without regulatory annotation.** The plot identifies the genomic location of different operons with annotated TF binding sites (blue), and those lacking regulatory descriptions (red). The identification of regulated operons was performed using data from RegulonDB (Gama-Castro et al., 2016), which are based on manually curated experimental and computational data. All operons listed in the database were considered, where an operon was assumed to be regulated if it had at least one transcription factor binding site associated with it (lists of operons and transcription factor binding sites are available on the RegulonDB 'Download' page, http://regulondb.ccg.unam.mx).

from the database of RegulonDB (Gama-Castro et al., 2016), which lists all the known regulatory features in this organism. Using this database, Figure 1.8 identifies the positions of each operon on the *E. coli* genome and whether it contains annotated transcription factor binding sites (blue) or not (red).

It is striking that over half of the operons lack any listed transcription factor binding sites. One explanation might be that these operons are constitutively expressing (i.e., no transcription factors regulate these operons). Alternatively, transcription might be controlled through changes in sigma factor concentrations, which would provide an alternative mechanism of regulation. For example, in stationary phase there is an increase in the cellular concentration of stationary phase sigma factor, RpoS (σ_{38}), and anti-sigma factors that decrease the level of functional sigma factor RpoD (σ_{70}) and alter the genome-wide transcription output (Jishage and Ishihama, 1995; Jishage et al., 1996). We can begin to consider whether these unannotated operons might be regulated by looking at the results of a recent proteome-wide census that was taken in *E. coli* across 22 growth conditions (Schmidt et al., 2016). In this work Schmidt

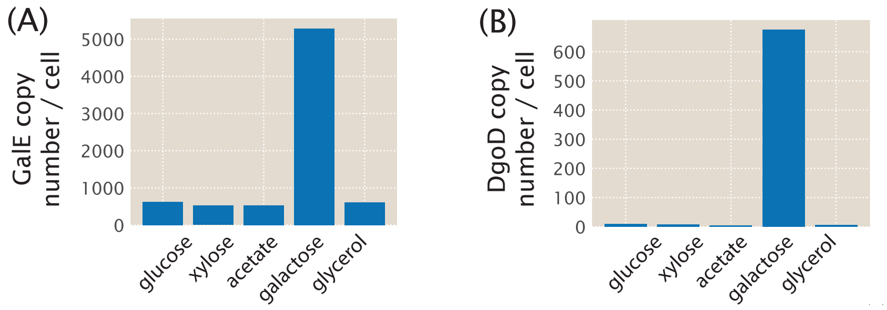


Figure 1.9: **Analysis of Schmidt *et al.* census study in *E. coli*.** (A) Here we show the protein copy numbers per cell for GalE across several carbon sources. Expression was sensitive to the presence of galactose which is consistent with its known regulation (with about 5000 copies per cell, versus about 500 for most other growth conditions). (B) DgoD was also found to be sensitive to the presence of galactose as the carbon source. The copy number was measured to be 675 copies per cell when cells were grown in galactose, and 15 copies per cell or less in all other conditions considered. For both (A) and (B), values are shown for growth in M9 minimal media, with glucose, xylose, acetate, galactose, and glycerol as carbon sources and obtained from Schmidt *et al.*, 2016.

et al. measured the copy number per cell of more than 2,300 proteins (about 55% of the *E. coli* proteome) across conditions that included different carbon sources, temperature, pH, growth phase, media, and growth in chemostats.

As a confirmation that the data of Schmidt *et al.* could identify regulated operons, we find that the GalE protein shows significantly higher expression when cells were grown in galactose (Figure 1.9A). GalE is involved in galactose catabolism, and its expression is known to increase due to loss of repression of the *galE* promoter when cells are grown in galactose (Irani et al., 1983; Semsey et al., 2007). Among promoters without any known regulation, we show the expression of DgoD in Figure 1.9B in several carbon sources. Cells grown in galactose showed much higher expression, with about 675 copies per cell, compared to at most 15 copies per cell across the other growth conditions. This is only one of many examples where a protein showed a large differential expression level across growth conditions and suggests that RegulonDB is incomplete.

In addition, we find that the expression variability for unannotated genes appears almost as variable as those with known regulation, further suggesting that many of the unannotated operons are under regulation. This is shown by the coefficient of variation (the ratio of the standard deviation to the mean protein copy number) for each protein across the 22 growth conditions (see Figure 1.10). Regulated proteins

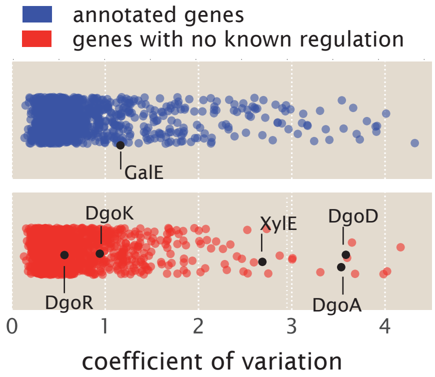
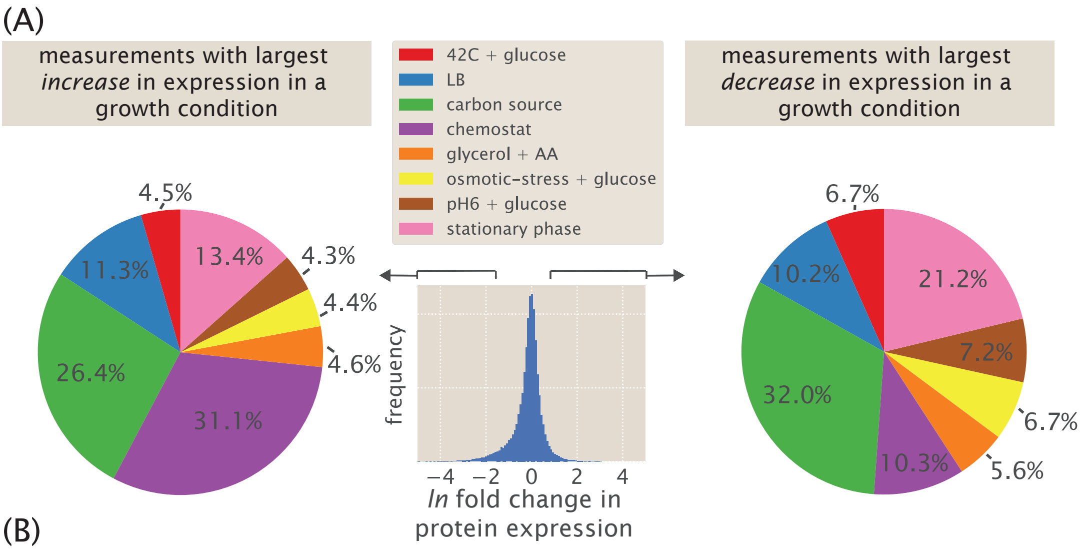


Figure 1.10: **Analysis of expression variability in Schmidt *et al.* census study across 22 growth conditions.** Coefficient of variation is calculated (standard deviation divided by mean copy number) across the 22 growth conditions for each protein measured in Schmidt *et al.*, 2016. Proteins are identified as either having regulatory annotation (blue) or not (red) using the annotations in RegulonDB (Gama-Castro et al., 2016). GalE is noted among the annotated genes and provides a reference as a gene that is known to be regulated and be perturbed in this study, as shown in Figure 1.9(A). Among the unannotated genes, those assocaited with the promoters of *purT*, *xylE*, and *dgoRKADT* are noted and are investigated in Chapter 4.

should be among those that exhibit a large change in copy number in one or a few growth conditions.

Lastly, this data represents a valuable resource to identify specific candidate genes for further regulatory investigation. In addition to calculating the coefficient of variation above, we can directly identify genes that are likely subject to transcriptional regulation and the growth condition necessary to perturb them. Here we do so by calculating the fold-change in expression for each protein relative to its average expression across all 22 growth conditions and summarize the analysis in Figure 1.11. Interestingly, among the highest fold-change values calculated, a substantial fraction are due to growth in a chemostat and in different carbon sources (Figure 1.11A, left plot). This contrasts with the lowest fold-changes values found, which are dominated by growth in different carbon sources or growth in stationary phase (Figure 1.11A, right plot). The highest and lowest fold-changes that were calculated from the data are summarized in Figure 1.11B and in Supplemental Section 1.5. Among these candidates, we consider the promoters of *purT*, *xylE*, and *dgoRKADT* in Chapter 4, where we use Sort-Seq to demonstrate that they are indeed under regulation at the transcriptional level.

Ultimately each promoter across the genome is not in isolation, and transcription factors may regulate expression from as few as a single gene, to several hundreds



(B)

unannotated genes with largest *increase* in expression		
gene	fold expression over average	growth condition
yqeG	22.0	stationary phase
dgoA	22.0	galactose
fdoI	22.0	LB
dusC	22.0	glycerol + AA
yjiM	22.0	fructose
amiC	19.3	glycerol + AA
fdoH	18.4	LB
dgoD	18.3	galactose
ygiQ	18.0	LB
fdoG	17.5	LB
xylE	17.1	xylose
ymdF	16.3	stationary phase
sdaB	16.0	LB
ykgE	14.9	LB

unannotated genes with largest *decrease* in expression		
gene	fold expression over average	growth condition
sgcB	0.005	chemostat
yeeD	0.007	stationary phase
ybhA	0.007	stationary phase
yqcA	0.012	stationary phase
yebO	0.012	stationary phase
ylaC	0.012	stationary phase
ymdF	0.014	stationary phase
yqcC	0.017	stationary phase
sgcB	0.019	pyruvate
ybeY	0.019	glycerol
ydiZ	0.019	stationary phase
syd	0.021	42C + glucose
sgcB	0.022	Succinate
ymdF	0.023	pyruvate

Figure 1.11: **Identification of unannotated genes that are sensitive to particular growth conditions in the Schmidt *et al.* census study.** The fold-change in protein copy number was calculated for all measured proteins (for each growth condition) relative to the average expression across all growth conditions. Note that the protein copy numbers were first normalized to total protein content within each growth condition before calculating fold-change. (A) The histogram shows the distribution of all calculated fold-change values associated with each protein across the 22 growth conditions. For the largest (left) and smallest (right) calculated fold-change values, pie charts summarize the fraction of each growth condition that led to the perturbation in protein copy number. Five percent of the measured fold-change values at each tail of the distribution were used to generate each pie chart. (B) The largest (left) and smallest (right) calculated fold-change values and the associated growth condition are summarized in the two tables. See Supplemental Section 1.5 for all candidates that showed a fold-change larger than 10, or less than 1/10.

of genes (Gama-Castro et al., 2016). At the level of the genome-wide regulon, we can also consider how sets of genes are controlled (i.e., gene network maps). An example of this is shown in Figure 1.12, which shows an inferred genetic network for *E. coli*. This represents the average mapping from a variety of inference approaches, using about 800 microarray datasets across about 500 growth conditions (Marbach et al., 2012). The map is quite enlightening; for example, they were able to identify clusters of physiologically associated genes, some of which had no prior known function. However, the authors note the accuracy of their network map at about 50%, which they checked in two ways. The first is by comparing their network map to the manually curated RegulonDB database that is the gold-standard for regulatory knowledge in *E. coli*. Here they find that the model appears to be consistent with RegulonDB about 50% of the time. To their credit, RegulonDB is likely incomplete and perhaps they have identified new regulatory interactions. However, the authors also tested 53 target gene predictions across five transcription factors, and only about half the time they found experimental support for the predicted regulatory connection.

This serves to highlight that at both the promoter level of detail and at a higher level of connectivity, our understanding is far from complete. By developing more complete mechanistic models of regulation at each promoter in the genome, it may be possible to develop more accurate descriptions of global genome regulation. In Chapter 4 we make the first steps toward this goal by developing an approach to systematically decipher the regulation of each promoter on the genome and demonstrate it in several well-characterized and unannotated promoters.

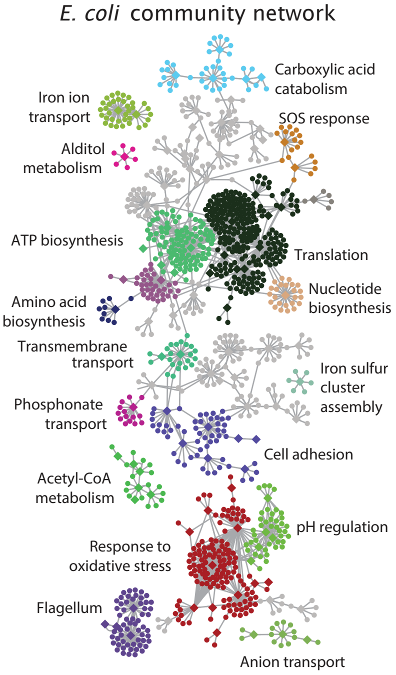


Figure 1.12: ***E. coli* network map.** The map represents an averaged inference from multiple approaches that use genome-wide datasets as input into their models. The map shows 1,505 genes (roughly one-third), including 204 transcription factors that are represented with diamonds. Edges represent predicted network connections. Network clusters were identified by testing for Gene Ontology-term enrichment. The figure was adapted from Marbach *et al.*, 2012.

1.5 Supplemental Information: Candidate genes with growth-dependent differential expression

Gene	Condition	Fold change in protein expression
yqeG	Stationary phase 3 days	22.000
dgoA	Galactose	22.000
fdoI	LB	22.000
dusC	Glycerol + AA	22.000
yjiM	Fructose	22.000
amiC	Glycerol + AA	19.286
fdoH	LB	18.434
dgoD	Galactose	18.303
ygiQ	LB	18.000
fdoG	LB	17.548
xylE	Xylose	17.083
ymdF	Stationary phase 3 days	16.259
sdaB	LB	15.987
ykgE	LB	14.941
sgcX	Chemostat mu=1.12	14.241
yedJ	Chemostat mu=1.35	14.143
ykgG	LB	13.082
yicI	Xylose	12.517
ygjP	Fumarate	12.312
sgcQ	Chemostat mu=1.12	11.844
rspA	Galactose	11.586
yqhC	Glucose	11.313
mutM	LB	11.273
htrB	LB	11.273
thiM	LB	11.057
yhcN	Stationary phase 1 day	10.970
ybgA	Stationary phase 3 days	10.486

Table 1.1: **Candidate unannotated genes with increased expression.** Fold change in protein expression was calculated relative to the average protein expression in the data of Schmidt *et al.* (2016). Proteins and growth conditions associated with a fold change of 10 or higher are summarized.

Gene	Condition	Fold change in protein expression
sdaB	Stationary phase 1 day	0.100
yidE	Glucose	0.100
maa	pH6 glucose	0.100
sgcB	pH6 glucose	0.099
rssA	Stationary phase 1 day	0.099
yffB	Galactose	0.099
djlA	Stationary phase 1 day	0.099
rihC	pH6 glucose	0.098
yedW	pH6 glucose	0.098
dgoK	Fumarate	0.098
yjiX	pH6 glucose	0.098
yagE	Stationary phase 3 days	0.098
rssA	Chemostat mu=1.21	0.097
rph	pH6 glucose	0.097
yddM	42C glucose	0.097
sgcX	pH6 glucose	0.097
ycfS	Galactose	0.096
yidE	Osmotic-stress glucose	0.096
yqaB	Stationary phase 3 days	0.096
cspB	Glucose	0.095
psuG	pH6 glucose	0.095
yjiX	Acetate	0.094
dgoD	Stationary phase 3 days	0.094
ybeY	Acetate	0.094
maa	Osmotic-stress glucose	0.093
hspQ	pH6 glucose	0.093
ybgK	Stationary phase 3 days	0.093
eutM	Stationary phase 3 days	0.093
ybaQ	Glycerol + AA	0.093
ugd	pH6 glucose	0.093
cueR	pH6 glucose	0.092
yagI	Stationary phase 3 days	0.092
yphF	LB	0.092
yidE	pH6 glucose	0.092
yehX	Stationary phase 1 day	0.092
yfhA	Stationary phase 1 day	0.092
fdoH	Pyruvate	0.092
cspB	42C glucose	0.092
rsmF	Stationary phase 3 days	0.091
glsA	Glycerol	0.091
glsA	Glycerol	0.091

yfgJ	Stationary phase 1 day	0.091
yigA	Glycerol + AA	0.090
acpH	Stationary phase 1 day	0.090
wecG	Stationary phase 3 days	0.090
ygdR	Glycerol + AA	0.090
sapF	Pyruvate	0.089
ytfL	Stationary phase 1 day	0.089
eutM	Fructose	0.088
hda	Chemostat mu=1.5	0.088
yagI	Chemostat mu=1.21	0.088
chpS	LB	0.088
djlA	Acetate	0.088
yfeC	Stationary phase 3 days	0.088
ymgD	pH6 glucose	0.088
ygfJ	Mannose	0.088
cspB	LB	0.087
tatE	Stationary phase 1 day	0.087
ynfL	42C glucose	0.087
ybiJ	Chemostat mu=1.12	0.086
fxsA	Glucose	0.086
ybeD	Stationary phase 1 day	0.086
fdoG	Stationary phase 1 day	0.085
yfeC	Stationary phase 1 day	0.085
fdoH	Chemostat mu=1.5	0.085
ymdF	Chemostat mu=1.21	0.084
ybeY	Glycerol + AA	0.084
ygdR	Glucose	0.083
cutC	Stationary phase 1 day	0.083
yjhH	Fumarate	0.082
psuG	42C glucose	0.082
fdoH	Glucosamine	0.082
fxsA	42C glucose	0.082
kefF	Stationary phase 3 days	0.082
fdoH	Glycerol + AA	0.081
rph	Stationary phase 3 days	0.082
ade	42C glucose	0.081
ytfQ	pH6 glucose	0.081
ybiJ	LB	0.081
cobU	Stationary phase 1 day	0.081
dosC	42C glucose	0.081
ygfJ	Xylose	0.080
ygdR	42C glucose	0.080
ytfL	Galactose	0.080

ybiJ	Chemostat mu=1.21	0.078
fdoH	Xylose	0.078
glsA	Pyruvate	0.078
glsA	Pyruvate	0.078
yhcN	Glucosamine	0.078
yfeY	42C glucose	0.078
sgcB	Osmotic-stress glucose	0.077
ydcH	Glycerol + AA	0.077
yheV	Stationary phase 3 days	0.077
ygfJ	Fructose	0.077
mltC	Glycerol + AA	0.076
truA	Chemostat mu=1.12	0.076
ybeY	LB	0.075
ybcL	LB	0.075
uppS	Stationary phase 3 days	0.075
ygfJ	Osmotic-stress glucose	0.074
yajI	Chemostat mu=1.21	0.074
cspB	Stationary phase 3 days	0.074
ycfS	Xylose	0.074
dgoD	Pyruvate	0.074
yfiC	Stationary phase 1 day	0.073
ybeY	Pyruvate	0.073
yfiC	Chemostat mu=1.21	0.072
ybdF	Chemostat mu=1.21	0.072
ycaP	LB	0.071
ygfJ	pH6 glucose	0.072
yhhK	pH6 glucose	0.071
psuG	Xylose	0.071
cspB	Chemostat mu=1.21	0.071
mltC	Acetate	0.070
pptA	Stationary phase 3 days	0.070
yciU	Stationary phase 1 day	0.070
hslR	Glycerol + AA	0.069
tusD	42C glucose	0.069
ilvG	pH6 glucose	0.069
ycfS	42C glucose	0.069
yjhH	42C glucose	0.068
yodC	Osmotic-stress glucose	0.068
yjiX	42C glucose	0.068
ybeD	Stationary phase 3 days	0.068
maa	Stationary phase 3 days	0.067
ygiD	Chemostat mu=1.21	0.067
mdfA	Chemostat mu=1.35	0.066

yjhH	pH6 glucose	0.066
yedW	LB	0.065
ygdR	Xylose	0.064
mltC	Pyruvate	0.064
greB	LB	0.063
pyrB	Glycerol + AA	0.063
ygdR	Stationary phase 1 day	0.062
ymdF	Glycerol	0.062
ymdF	Chemostat mu=1.12	0.062
yfjG	Chemostat mu=1.12	0.062
eutM	Xylose	0.062
ybeY	Fumarate	0.061
ymdF	pH6 glucose	0.061
mltC	Fumarate	0.061
priC	Glycerol + AA	0.061
dicA	Fructose	0.061
yidE	LB	0.061
ybcL	Osmotic-stress glucose	0.059
nudI	Chemostat mu=1.12	0.059
ydiZ	LB	0.059
fdoH	Stationary phase 3 days	0.059
yfeD	Chemostat mu=1.12	0.059
yjdI	Galactose	0.058
dgoD	Osmotic-stress glucose	0.058
ydcH	Glucosamine	0.058
yjiX	Chemostat mu=1.12	0.058
yceA	Stationary phase 1 day	0.058
mltC	Glycerol	0.058
yebO	Stationary phase 1 day	0.058
mltC	Glucosamine	0.057
ilvG	Stationary phase 1 day	0.056
ygfJ	Glycerol + AA	0.056
tdk	Stationary phase 3 days	0.055
ybaV	Pyruvate	0.054
rpmJ	Glucose	0.054
pyrB	LB	0.054
ydcH	Glucose	0.054
ymdF	Chemostat mu=1.5	0.053
ymgD	Chemostat mu=1.12	0.052
yjjG	Stationary phase 3 days	0.052
holD	Xylose	0.051
yacG	Fructose	0.051
sgcX	Osmotic-stress glucose	0.050

ynfL	Chemostat mu=1.5	0.050
ybaQ	Stationary phase 3 days	0.050
cobU	pH6 glucose	0.050
ybeY	Glucose	0.050
yeeD	42C glucose	0.050
ydcH	pH6 glucose	0.049
ybaQ	Stationary phase 1 day	0.048
ugd	Osmotic-stress glucose	0.048
mdfA	pH6 glucose	0.048
yjiX	Galactose	0.048
djlA	Glycerol + AA	0.047
rnhB	Stationary phase 3 days	0.047
cspB	Osmotic-stress glucose	0.046
ydcI	Stationary phase 1 day	0.046
poxA	Chemostat mu=1.12	0.045
dedD	Stationary phase 3 days	0.044
nudI	Pyruvate	0.043
ycaP	Glycerol + AA	0.042
yfeY	Stationary phase 3 days	0.042
rfaC	Stationary phase 1 day	0.041
ybhA	Stationary phase 1 day	0.041
ypfH	Succinate	0.041
yfeY	Stationary phase 1 day	0.041
yjiX	Succinate	0.041
yffB	Glucosamine	0.040
eutM	Galactose	0.040
yjiX	Chemostat mu=1.5	0.040
hicB	Glycerol + AA	0.039
ydcI	Stationary phase 3 days	0.039
yddM	Stationary phase 3 days	0.039
hicB	LB	0.039
yjiX	Glycerol	0.039
ybiJ	Glycerol + AA	0.038
ybcL	pH6 glucose	0.038
fdoH	Fructose	0.037
cspG	Chemostat mu=1.35	0.037
yigP	Stationary phase 1 day	0.037
ymdF	Osmotic-stress glucose	0.036
fdoH	Osmotic-stress glucose	0.036
djlA	Glucosamine	0.036
ybiJ	Galactose	0.036
ymgD	Chemostat mu=1.5	0.035
fdoH	pH6 glucose	0.035

nudI	Osmotic-stress glucose	0.035
ymdF	LB	0.035
ybiJ	Chemostat mu=1.35	0.034
djlA	Glucose	0.033
yqcC	Stationary phase 1 day	0.033
ymdF	Succinate	0.033
maa	Stationary phase 1 day	0.032
mltC	LB	0.032
rnhB	pH6 glucose	0.028
ymdF	42C glucose	0.027
yacG	Stationary phase 3 days	0.027
ymdF	Galactose	0.025
ylaC	Stationary phase 1 day	0.024
yddM	pH6 glucose	0.023
ymdF	Pyruvate	0.023
sgcB	Succinate	0.022
syd	42C glucose	0.021
ydiZ	Stationary phase 1 day	0.019
ybeY	Glycerol	0.019
sgcB	Pyruvate	0.019
yqcC	Stationary phase 3 days	0.017
ymdF	Stationary phase 1 day	0.014
ylaC	Stationary phase 3 days	0.012
yebO	Stationary phase 3 days	0.012
yqcA	Stationary phase 3 days	0.012
ybhA	Stationary phase 3 days	0.007
yeeD	Stationary phase 1 day	0.007
yeeD	Stationary phase 3 days	0.007
sgcB	Chemostat mu=1.35	0.005

Table 1.2: **Candidate unannotated genes with decreased expression.** Fold change in protein expression was calculated relative to the average protein expression in the data of Schmidt *et al.* (2016). Proteins and growth conditions associated with a fold change of 1/10 or lower are summarized.

References

Ackers, G. K., Johnson, A. D., and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences* 79.4, pp. 1129–33.

Auerbach, A. (2012). Thinking in Cycles: MWC is a Good Model for Acetylcholine Receptor-Channels. *The Journal of Physiology* 590.1, pp. 93–8.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* 58.3, pp. 268–276.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005a). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics and Development* 15.2, pp. 116–124.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005b). Transcriptional regulation by the numbers: applications. *Current Opinion in Genetics and Development* 15.2, pp. 125–135.

Boedicker, J. Q., Garcia, H. G., Johnson, S., and Phillips, R. (2013a). DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Physical Biology* 10.6, p. 066005.

Boedicker, J. Q., Garcia, H. G., and Phillips, R. (2013b). Theoretical and Experimental Dissection of DNA Loop-Mediated Repression. *Physical Review Letters* 110.1, p. 018101.

Brewster, R. C., Jones, D. L., and Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology* 8.12.

Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. *Cell* 156.6, pp. 1312–1323.

Browning, D. F. and Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology* 14.10, pp. 638–650.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–41.

Busby, S. and Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology* 293.2, pp. 973–979.

Canals, M., Lane, J. R., Wen, A., Scammells, P. J., Sexton, P. M., and Christopoulos, A. (2012). A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation. *Journal of Biological Chemistry* 287.1, pp. 650–659.

Changeux, J.-P. (2013). The Origins of Allostery: From Personal Memories to Material for the Future. *Journal of Molecular Biology* 425.9, pp. 1396–1406.

Changeux, J.-P. (1961). The Feedback Control Mechanism of Biosynthetic L-Threonine Deaminase by L-Isoleucine. *Cold Spring Harbor Symposia on Quantitative Biology* 26.1, pp. 313–318.

Cipriano, M. J., Novichkov, P. N., Kazakov, A. E., Rodionov, D. A., Arkin, A. P., Gelfand, M. S., and Dubchak, I. (2013). RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* 14.1, pp. 213–221.

Cournac, A. and Plumbridge, J. (2013). DNA Looping in Prokaryotes: Experimental and Theoretical Approaches. *Journal of Bacteriology* 195.6, pp. 1109–1119.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature* 227.5258, pp. 561–563.

Daber, R., Sochor, M. A., and Lewis, M. (2011). Thermodynamic analysis of mutant lac repressors. *Journal of Molecular Biology* 409.1, pp. 76–87.

Einav, T., Mazutis, L., and Phillips, R. (2016). Statistical Mechanics of Allosteric Enzymes. EN. *The Journal of Physical Chemistry B* 121 (15).

Feklístov, A., Sharon, B. D., Darst, S. A., and Gross, C. A. (2014). Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annual Review of Microbiology* 68.1, pp. 357–376.

Fenton, A. W. (2008). Allostery: an illustrated definition for the 'second secret of life'. *Trends in Biochemical Sciences* 33.9, pp. 420–425.

Galperin, M. Y. and Koonin, E. V. (2010). From complete genome sequence to "complete" understanding? *Trends in Biotechnology* 28.8, pp. 398–406.

Gama-Castro, S. et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* 44.D1, pp. D133–D143.

Garcia, H. G. and Phillips, R. (2011). Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–8.

Garcia, H. G., Grayson, P., Han, L., Inamdar, M., Kondev, J., Nelson, P. C., Phillips, R., Widom, J., and Wiggins, P. A. (2007). Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. *Biopolymers* 85.2, pp. 115–130.

Garcia, H. G., Lee, H. J., Boedicker, J. Q., and Phillips, R. (2011). Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. *Biophysical Journal* 101.3, 535–544.

Garcia-Pino, A., Balasubramanian, S., Wyns, L., Gazit, E., De Greve, H., Magnuson, R. D., Charlier, D., Nuland, N. A. J. van, and Loris, R. (2010). Allostery and Intrinsic Disorder Mediate Transcription Regulation by Conditional Cooperativity. *Cell* 142.1, pp. 101–111.

Garcia-Pino, A., De Gieter, S., Talavera, A., De Greve, H., Efremov, R. G., and Loris, R. (2016). An intrinsically disordered entropic switch determines allostery in Phd–Doc regulation. *Nature Chemical Biology* 12.7, pp. 490–496.

Gaston, K., Bell, A., Kolb, A., Buc, H., and Busby, S. (1990). Stringent spacing requirements for transcription activation by CRP. *Cell* 62.4, pp. 733–743.

Gerhart, J. C. and Pardee, A. B. (1962). The Enzymology of Control by Feedback Inhibition. *Journal of Biological Chemistry* 237.3, pp. 891–896.

Gerhart, J. C. and Pardee, A. B. (1961). Separation of feedback inhibition from activity of aspartate transcarbamylase (ATCase). In: *Fed. Proc.* Vol. 20, p. 224.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17.6, pp. 333–351.

Gruber, T. M. and Gross, C. A. (2003). Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. *Annual Review of Microbiology* 57.1, pp. 441–466.

Gur, E., Biran, D., and Ron, E. Z. (2011). Regulated proteolysis in Gram-negative bacteria - how and when? *Nature Reviews Microbiology* 9.12, pp. 839–848.

Irani, M. H., Orosz, L., and Adhya, S. (1983). A control element within a structural gene: The *gal* operon of *Escherichia coli*. *Cell* 32.3, pp. 783–788.

Jacob, F. (2011). The Birth of the Operon. *Science* 332.6031, pp. 767–767.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* 3.3, pp. 318–356.

Jarvis, E. D. et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346.6215, pp. 1320–1331.

Jishage, M. and Ishihama, A. (1995). Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. *Journal of Bacteriology* 177.23, pp. 6832–6835.

Jishage, M., Iwata, A., Ueda, S., and Ishihama, A. (1996). Regulation of RNA polymerase sigma subunit synthesis in*Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions. *Journal of Bacteriology* 178.18, pp. 5447–5451.

Keymer, J. E., Endres, R. G., Skoge, M., Meir, Y., and Wingreen, N. S. (2006). Chemosensing in *Escherichia coli*: two regimes of two-state receptors. *Proceedings of the National Academy of Sciences* 103.6, pp. 1786–91.

Kılıç, S., White, E. R., Sagitova, D. M., Cornish, J. P., and Erill, I. (2013). CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Research* 42.D1, pp. D156–D160.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* 107.20, 9158–9163.

Koepfli, K.-P., Paten, B., O'Brien, S. J., and Scientists, t. G.K.C. o. (2015). The Genome 10K Project: A Way Forward. *Annual Review of Animal Biosciences* 3.1, pp. 57–111.

Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences* 104.14, pp. 6043–6048.

Labow, M. A., Baim, S. B., Shenk, T., and Levine, A. J. (1990). Conversion of the lac repressor into an allosterically regulated transcriptional activator for mammalian cells. *Molecular and Cellular Biology* 10.7, pp. 3343–3356.

Land, M. et al. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* 15.2, pp. 141–161.

Lander, E. S. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409.6822, pp. 860–921.

Lee, D. J., Minchin, S. D., and Busby, S. J. W. (2012). Activating Transcription in Bacteria. *Annual Review of Microbiology* 66.1, pp. 125–152.

Levy, S. E. and Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics* 17.1, pp. 95–115.

Lewis, M. (2011). A tale of two repressors – a historical perspective. *Journal of Molecular Biology* 409.1, pp. 14–27.

Loman, N. J. and Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* 13.12, pp. 787–794.

Loomis, W. F. and Magasanik, B. (1967). The catabolite repression gene of the Lac operon in *Escherichia coli*. *Journal of Molecular Biology* 23.3, pp. 487–494.

Loughrey, D., Watters, K. E., Settle, A. H., and Lucks, J. B. (2014). SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, pp. 1–10.

Maeda, H. (2000). Competition among seven *Escherichia coli* sigma subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Research* 28.18, pp. 3497–3503.

Mandal, N., Su, W., Haber, R., Adhya, S., and Echols, H. (1990). DNA looping in cellular repression of transcription of the galactose operon. *Genes & Development* 4.3, pp. 410–418.

Marbach, D. et al. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods* 9.8, pp. 796–804.

Martin, K., Huo, L., and Schleif, R. F. (1986). The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. *Proceedings of the National Academy of Sciences* 83.11, pp. 3654–3658.

Martins, B. M. C. and Swain, P. S. (2011). Trade-Offs and constraints in allosteric sensing. *PLoS Computational Biology* 7.11, pp. 1–13.

Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* 12.1, pp. 88–118.

Monod, J. and Jacob, F. (1961). General Conclusions - Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. *Cold Spring Harbor Symposia on Quantitative Biology* 26, pp. 389–401.

Monod, J., Changeux, J.-P., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology* 6.4, pp. 306–329.

Moran, U., Phillips, R., and Milo, R. (2010). SnapShot: Key Numbers in Biology. *Cell* 141.7, 1262–1262.e1.

Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. (2014). The ensemble nature of allostery. *Nature* 508.7496, pp. 331–339.

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Research* 31.1, pp. 266–269.

Murakami, K. S. and Darst, S. A. (2003). Bacterial RNA polymerases: the wholo story. *Current Opinion in Structural Biology* 13.1, pp. 31–39.

Oehler, S., Eismann, E. R., Krämer, H., and Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal* 9.4, pp. 973–979.

Pachter, L. (2013). *Seq. https://liorpachter.wordpress.com/seq/. Blog.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10.10, pp. 669–680.

Phillips, R. (2015). Napoleon Is in Equilibrium. *dx.doi.org* 6.1, pp. 85–111.

Rabbani, B., Nakaoka, H., Akhondzadeh, S., Tekin, M., and Mahdieh, N. (2016). Next generation sequencing: implications in personalized medicine and pharmacogenomics. *Mol. BioSyst.* 12.6, pp. 1818–1830.

Rydenfelt, M., Garcia, H. G., Cox III, R. S., and Phillips, R. (2014). The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS ONE* 9.12, e114347.

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27.10, pp. 946–950.

Schleif, R. (2010). AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiology Reviews* 34.5, pp. 779–796.

Schmidt, A. et al. (2016). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology* 34 (1), pp. 104–111.

Schultz, S, Shields, G, and Steitz, T (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* 253.5023, pp. 1001–1007.

Semsey, S., Krishna, S., Sneppen, K., and Adhya, S. (2007). Signal integration in the galactose network of *Escherichia coli*. *Molecular Microbiology* 65.2, pp. 465–476.

Seshasayee, A. S. N., Sivaraman, K., and Luscombe, N. M. (2011). An Overview of Prokaryotic Transcription Factors. In: *A Handbook of Transcription Factors*. Dordrecht: Springer, Dordrecht, pp. 7–23.

Sharma, H., Yu, S., Kong, J., Wang, J., and Steitz, T. A. (2009). Structure of apo-CAP reveals that large conformational changes are necessary for DNA binding. *Proceedings of the National Academy of Sciences* 106.39, pp. 16604–16609.

Swem, L. R., Swem, D. L., Wingreen, N. S., and Bassler, B. L. (2008). Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio harveyi. *Cell* 134.3, pp. 461–473.

Ushida, C. and Aiba, H. (1990). Helical phase dependent action of CRP: effect of the distance between the CRP site and the -35 region on promoter activity. *Nucleic Acids Research* 18.21, pp. 6325–6330.

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., and Steenackers, H. P. (2015). RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in Microbiology* 6.300, p. 134.

Velyvis, A., Yang, Y. R., Schachman, H. K., and Kay, L. E. (2007). A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences* 104.21, pp. 8815–20.

Vilar, J. M. G. and Leibler, S. (2003). DNA Looping and Physical Constraints on Transcription Regulation. *Journal of Molecular Biology* 331.5, pp. 981–989.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., and Kegel, W. K. (2014). Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters* 113.25, pp. 1–5.

Wilkinson, S. P. and Grove, A. (2006). Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. *Current Issues in Molecular Biology* 8.1, pp. 51–62.

Chapter 2

TUNING TRANSCRIPTIONAL REGULATION THROUGH SIGNALING: A PREDICTIVE THEORY OF ALLOSTERIC INDUCTION

A version of this chapter originally appeared as M. Razo-Mejia, S. L. Barnes, N. M. Belliveau, G. Chure, T. Einav, M. Lewis, R. Phillips (2017). Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. bioRxiv, 111013. http://doi.org/10.1101/111013. It is also in preparation for publication in a peer-reviewed journal. M.R.M., S.L.B., N.M.B., G.C., T.E. contributed equally to this work.

2.1 Introduction

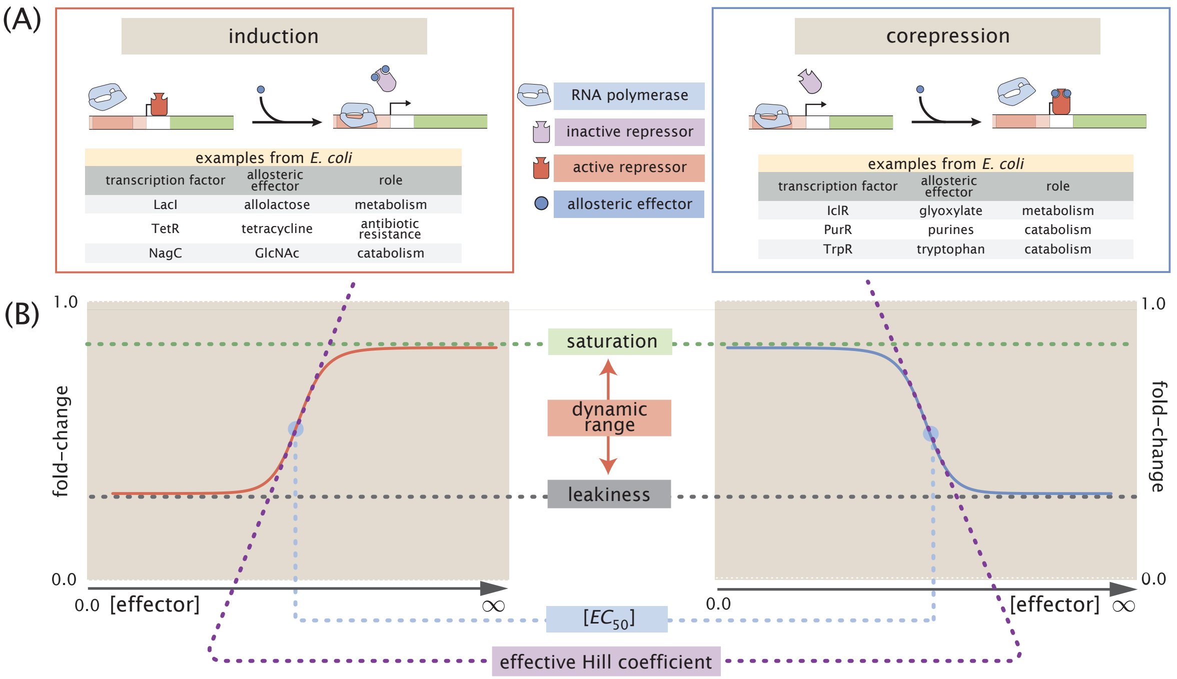
Understanding how organisms sense and respond to changes in their environment has long been a central theme of biological inquiry. At the cellular level, this interaction is mediated by a diverse collection of molecular signaling pathways. A pervasive mechanism of signaling in these pathways is allosteric regulation, in which the binding of a ligand induces a conformational change in some target molecule, triggering a signaling cascade (Lindsley and Rutter, 2006). One of the most important examples of such signaling is offered by transcriptional regulation, where a transcription factor's propensity to bind to DNA will be altered upon binding to an allosteric effector.

Despite the overarching importance of this mode of signaling, a quantitative understanding of the molecular interactions between extracellular inputs and gene expression remains poorly explored. Attempts to reconcile theoretical models and experiments have often been focused on fitting data retrospectively after experiments have been conducted (Kuhlman et al., 2007; Daber et al., 2009). Further, many treatments of induction are strictly phenomenological, electing to treat induction curves individually either using Hill functions or as ratios of polynomials without acknowledging that allosteric proteins have distinct conformational states depending upon whether an effector molecule is bound to them or not (Setty et al., 2003; Poelwijk et al., 2011; Vilar and Saiz, 2013; Rogers et al., 2015; Rohlhill et al., 2017). These fits are made in experimental conditions in which there is great uncertainty about the copy number of both the transcription factor and the regulated locus, meaning that the underlying minimal set of parameters cannot be pinned down unequivocally. Without minimal models involving clear, specific parameters, such fits are primarily descriptive and can do little to predict the system response as parameters are varied. Furthermore, phenomenological fits with unclear parameters provide little prospect for predicting or understanding what molecular properties determine key phenotypic parameters such as leakiness, dynamic range, $[EC_{50}]$, and the effective Hill coefficient as discussed in Martins and Swain (2011) and Marzen et al. (2013) and illustrated in Fig. 2.1. In response to these concerns, we formulate a minimal Monod-Wyman-Changeux (MWC) model of transcription factor induction in conjunction with a corresponding thermodynamic model of repression. While some treatments of induction have used MWC models to predict transcriptional outputs (Daber et al., 2009; Daber et al., 2011; Sochor, 2014), there has been no systematic experimental test of how well such a model can predict the induction process over broad swathes of regulatory parameter space. To that end, we use the MWC model to make parameter-free predictions about how the induction response will be altered when transcription factor copy number and operator strength are systematically varied.

We test our model in the context of the simple repression motif – a widespread bacterial genetic regulatory architecture in which binding of a transcription factor occludes binding of an RNA polymerase thereby inhibiting transcription initiation. A recent survey of different regulatory architectures within the *E. coli* genome revealed that more than 100 genes are characterized by the simple repression motif, making it a common and physiologically relevant architecture (Rydenfelt et al., 2014). Building upon previous work (Garcia and Phillips, 2011; Brewster et al., 2014; Weinert et al., 2014), we present a statistical mechanical rendering of allostery in the context of induction and corepression, shown schematically in Fig. 2.1(A), and use this model as the basis of parameter-free predictions which we then probe experimentally. Specifically, we model the allosteric response of transcriptional repressors using the MWC model, which stipulates that an allosteric protein fluctuates between two distinct conformations – an active and inactive state – in thermodynamic equilibrium (Monod et al., 1965). In the context of induction, effector binding increases the probability that a repressor will be in the inactive state, weakening its ability to bind to the promoter and resulting in increased expression. The framework presented here provides considerable insight beyond that of simply fitting a sigmoidal curve to inducer titration data. We aim to explain and predict the relevant biologically important parameters of an induction profile, such as characterizing the midpoint and steepness of its response as well as the limits of minimum and maximum expression as shown in Fig. 2.1(B). By combining this MWC treatment of induction with a thermodynamic model of transcriptional regulation (Fig. 2.2), we create a general quantitative model of allosteric transcriptional regulation that is applicable to a wide range of regulatory architectures such as activation, corepression, and various combinations thereof, extending our quantitative understanding of these schemes (Bintu et al., 2005) to include signaling.

To demonstrate the predictive power of our theoretical formulation across a wide range of both operator strengths and repressor copy numbers, we design an *E. coli* genetic construct in which the binding probability of a repressor regulates gene expression of a fluorescent reporter. Our ultimate goal was to determine if one self-consistent set of parameters can describe experimental data across a broad suite of strains, tuning variables, and experimental methods (Garcia and Phillips, 2011; Garcia et al., 2011; Brewster et al., 2012; Boedicker et al., 2013a; Boedicker et al., 2013b; Brewster et al., 2014). Using components from the well-characterized *lac* system in *E. coli*, we first quantify the three parameters associated with the induction of the repressor, namely, the binding affinity of the active and inactive repressor to the inducer and the free energy difference between the active and inactive repressor states. We fix these parameters by fitting to measurements of the fold-change in gene expression as a function of inducer concentration for a circuit with known repressor copy number and repressor-operator binding energy. Using these newly estimated allosteric parameters, we make accurate, parameter-free predictions of the induction response for many other combinations of repressor copy number and binding energy. This goes well beyond previous treatments of the induction phenomenon and shows that one extremely compact set of parameters can be applied self-consistently and predictively to vastly different regulatory situations including simple repression on chromosome, cases in which decoy binding sites for repressor are put on plasmids, cases in which multiple genes compete for the same regulatory machinery, cases involving multiple binding sites for repressor leading to DNA looping, and the induction experiments described here. The broad reach of this minimal parameter set is highlighted in Fig. 2.3.

Rather than viewing the behavior of each circuit as giving rise to its own unique input-output response, the formulation of the MWC model presented here provides a means to characterize these seemingly diverse behaviors using a single unified framework governed by a small set of parameters, applicable even to mutant repressors in much the same way that earlier work showed how mutants in quorum sensing and chemotaxis receptors could be understood within a minimal MWC-based model (Keymer et al., 2006; Swem et al., 2008). Another insight that emerges from our theoretical treatment is how a subset in the parameter space of repressor copy number, operator binding site strength, and inducer concentration can all yield the same level of gene expression. Our application of the MWC model allows us to understand these degeneracies in parameter space through an expression for the free energy of repressor binding, a nonlinear combination of physical parameters which determines the system's mean response and is the fundamental quantity that dictates the phenotypic cellular response to a signal.



examples from *E. coli*

transcription factor	allosteric effector	role
LacI	allolactose	metabolism
TetR	tetracycline	antibiotic resistance
NagC	GlcNAc	catabolism

examples from *E. coli*

transcription factor	allosteric effector	role
IclR	glyoxylate	metabolism
PurR	purines	catabolism
TrpR	tryptophan	catabolism

Figure 2.1: **Transcription regulation architectures involving an allosteric repressor.** (A) We consider a promoter regulated solely by an allosteric repressor. When bound, the repressor prevents RNAP from binding and initiating transcription. Induction is characterized by the addition of an effector which binds to the repressor and stabilizes the inactive state (defined as the state which has a low affinity for DNA), thereby increasing gene expression. In corepression, the effector stabilizes the repressor's active state and thus further reduces gene expression. We list several characterized examples of induction and corepression that support different physiological roles in *E. coli* (Huang et al., 2011; Li et al., 2014). (B) A schematic regulatory response of the two architectures shown in Panel (A) plotting the fold-change in gene expression as a function of effector concentration, where fold-change is defined as the ratio of gene expression in the presence versus the absence of repressor. We consider the following key phenotypic properties that describe each response curve: the minimum response (leakiness), the maximum response (saturation), the difference between the maximum and minimum response (dynamic range), the concentration of ligand which generates a fold-change halfway between the minimal and maximal response ($[EC_{50}]$), and the log-log slope at the midpoint of the response (effective Hill coefficient).

2.2 Results

Characterizing Transcription Factor Induction using the Monod-Wyman-Changeux (MWC) Model

We begin by considering the induction of a simple repression genetic architecture, in which the binding of a transcriptional repressor occludes the binding of RNA polymerase (RNAP) to the DNA (Ackers et al., 1982; Buchler et al., 2003). When an effector (hereafter referred to as an 'inducer' for the case of induction) binds to the repressor, it shifts the repressor's allosteric equilibrium towards the inactive state as specified by the MWC model (Monod et al., 1965). This causes the repressor to bind more weakly to the operator, which increases gene expression. Simple repression motifs in the absence of inducer have been previously characterized by an equilibrium model where the probability of each state of repressor and RNAP promoter occupancy is dictated by the Boltzmann distribution (Ackers et al., 1982; Buchler et al., 2003; Vilar and Leibler, 2003; Bintu et al., 2005a; Garcia and Phillips, 2011; Brewster et al., 2014) (we note that non-equilibrium models of simple repression have been shown to have the same functional form that we derive below (Phillips, 2015)). We extend these models to consider the role of allostery by accounting for the equilibrium state of the repressor through the MWC model as follows.

Consider a cell with copy number P of RNAP and R repressors. Our model assumes that the repressor can exist in two conformational states. R_A repressors will be in the active state (the favored state when the repressor is not bound to an inducer; in this state the repressor binds tightly to the DNA) and the remaining R_I repressors will be in the inactive state (the predominant state when repressor is bound to an inducer; in this state the repressor binds weakly to the DNA) such that $R_A + R_I = R$. Repressors fluctuate between these two conformations in thermodynamic equilibrium (Monod et al., 1965).

Thermodynamic models of gene expression begin by enumerating all possible states of the promoter and their corresponding statistical weights. As shown in Fig. 2.2(A), the promoter can either be empty, occupied by RNAP, or occupied by either an active or inactive repressor. We assign the repressor a different DNA binding affinity in the active and inactive state. In addition to the specific binding sites at the promoter, we assume that there are N_{NS} non-specific binding sites elsewhere (i.e. on parts of the genome outside the simple repression architecture) where the RNAP or the repressor can bind. All specific binding energies are measured relative to the average non-specific binding energy. Our model explicitly ignores the complexity of the distribution of non-specific binding affinities in the genome and makes the assumption that a single parameter can capture the energy difference between our binding site of interest and the average site in the reservoir. Thus, $\Delta\varepsilon_P$ represents the energy difference between the specific and non-specific binding for RNAP to the DNA. Likewise, $\Delta\varepsilon_{RA}$ and $\Delta\varepsilon_{RI}$ represent the difference in specific and non-specific binding energies for repressor in the active or inactive state, respectively.

Thermodynamic models of transcription (Ackers et al., 1982; Buchler et al., 2003; Vilar and Leibler, 2003; Bintu et al., 2005; Bintu et al., 2005a; Kuhlman et al., 2007; Daber et al., 2011; Garcia and Phillips, 2011; Brewster et al., 2014; Weinert et al., 2014) posit that gene expression is proportional to the probability that the RNAP is bound to the promoter p_{bound}, which is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} + \frac{R_I}{N_{NS}} e^{-\beta\Delta\varepsilon_{RI}} + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P}}, \tag{2.1}$$

with $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant and T is the temperature of the system. As $k_B T$ is the natural unit of energy at the molecular length scale, we treat the products $\beta\Delta\varepsilon_j$ as single parameters within our model. Measuring p_{bound} directly is fraught with experimental difficulties, as determining the exact proportionality between expression and p_{bound} is not straightforward. Instead, we measure the fold-change in gene expression due to the presence of the repressor. We define fold-change as the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor (i.e. constitutive expression), namely,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \tag{2.2}$$

We can simplify this expression using two well-justified approximations: (1) $\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} \ll 1$ implying that the RNAP binds weakly to the promoter ($N_{NS} = 4.6 \times 10^6$, $P \approx 10^3$ (Klumpp and Hwa, 2008), $\Delta\varepsilon_P \approx -2$ to $-5\ k_B T$ (Brewster et al., 2012), so that $\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_P} \approx 0.01$) and (2) $\frac{R_I}{N_{NS}} e^{-\beta\Delta\varepsilon_{RI}} \ll 1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}$ which reflects our assumption that the inactive repressor binds weakly to the promoter of interest. Using these approximations, the fold-change reduces to the form

$$\text{fold-change} \approx \left(1 + \frac{R_A}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} \equiv \left(1 + p_A(c) \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.3}$$

where in the last step we have introduced the fraction $p_A(c)$ of repressors in the active state given a concentration c of inducer, which is defined as $R_A(c) = p_A(c)R$. Since inducer binding shifts the repressors from the active to the inactive state, $p_A(c)$ is a decreasing function of c (Marzen et al., 2013).

We compute the probability $p_A(c)$ that a repressor with n inducer binding sites will be active using the MWC model. After first enumerating all possible configurations of a repressor bound to inducer (see Fig. 2.2(B)), $p_A(c)$ is given by the sum of the

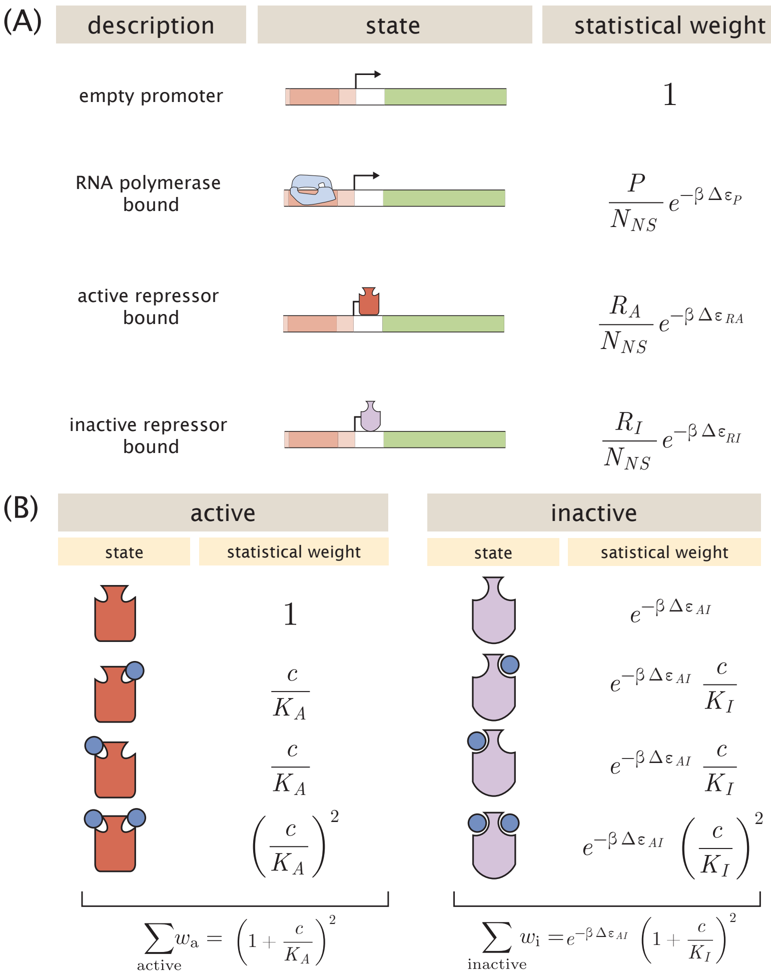

(A)
description
state
statistical weight
empty promoter
1
RNA polymerase bound
$\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}$
active repressor bound
$\frac{R_A}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}$
inactive repressor bound
$\frac{R_I}{N_{NS}} e^{-\beta \Delta \varepsilon_{RI}}$
(B)
active
inactive
state
statistical weight
state
satistical weight
1
$e^{-\beta \Delta \varepsilon_{AI}}$
$\frac{c}{K_A}$
$e^{-\beta \Delta \varepsilon_{AI}} \frac{c}{K_I}$
$\frac{c}{K_A}$
$e^{-\beta \Delta \varepsilon_{AI}} \frac{c}{K_I}$
$\left(\frac{c}{K_A}\right)^2$
$e^{-\beta \Delta \varepsilon_{AI}} \left(\frac{c}{K_I}\right)^2$
$\sum_{\text{active}} w_{\text{a}} = \left(1 + \frac{c}{K_A}\right)^2$
$\sum_{\text{inactive}} w_{\text{i}} = e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^2$

Figure 2.2: **States and weights for the simple repression motif.** (A) RNAP (light blue) and a repressor compete for binding to a promoter of interest. There are R_A repressors in the active state (red) and R_I repressors in the inactive state (purple). The difference in energy between a repressor bound to the promoter of interest versus another non-specific site elsewhere on the DNA equals $\Delta\varepsilon_{RA}$ in the active state and $\Delta\varepsilon_{RI}$ in the inactive state; the P RNAP have a corresponding energy difference $\Delta\varepsilon_P$ relative to non-specific binding on the DNA. N_{NS} represents the number of non-specific binding sites for both RNAP and repressor. (B) A repressor has an active conformation (red, left column) and an inactive conformation (purple, right column), with the energy difference between these two states given by $\Delta\varepsilon_{AI}$. The inducer (blue circle) at concentration c is capable of binding to the repressor with dissociation constants K_A in the active state and K_I in the inactive state. The eight states for a dimer with $n = 2$ inducer binding sites are shown along with the sums of the active and inactive states.

weights of the active states divided by the sum of the weights of all possible states, namely,

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}, \tag{2.4}$$

where K_A and K_I represent the dissociation constant between the inducer and repressor in the active and inactive states, respectively, and $\Delta\varepsilon_{AI} = \varepsilon_I - \varepsilon_A$ stands for the free energy difference between a repressor in the inactive and active state (the quantity $e^{-\Delta\varepsilon_{AI}}$ is sometimes denoted by L (Monod et al., 1965; Marzen et al., 2013) or $K_{\mathrm{RR}*}$ (Daber et al., 2011)). A repressor which favors the active state in the absence of inducer ($\Delta\varepsilon_{AI} > 0$) will be driven towards the inactive state upon inducer binding when $K_I < K_A$. The specific case of a repressor dimer with $n = 2$ inducer binding sites is shown in Fig. 2.2(B).

Substituting $p_A(c)$ from Eq. (2.4) into Eq. (2.3) yields the general formula for induction of a simple repression regulatory architecture (Phillips, 2015), namely,

$$\text{fold-change} = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_{RA}}\right)^{-1}. \tag{2.5}$$

While we have used the specific case of simple repression with induction to craft this model, we reiterate that the exact same mathematics describe the case of corepression in which binding of an allosteric effector stabilizes the active state of the repressor and decreases gene expression (see Fig. 2.1(B)). Interestingly, we shift from induction (governed by $K_I < K_A$) to corepression ($K_I > K_A$) as the ligand transitions from preferentially binding to the inactive repressor state to stabilizing the active state. Furthermore, this general approach can be used to describe a variety of other motifs such as activation, multiple repressor binding sites, and combinations of activator and repressor binding sites (Bintu et al., 2005; Brewster et al., 2014; Weinert et al., 2014).

This key formula presented in Eq. (2.5) enables us to make precise quantitative statements about induction profiles. Motivated by the broad range of predictions implied by this equation and by the belief that no previous work has made a thorough test of the predictive power of the MWC framework in the context of induction, we designed a series of experiments using the *lac* system in *E. coli* to tune the control parameters for a simple repression genetic circuit. As discussed in Fig. 2.3, previous studies from our lab have provided us with well-characterized values for many of the parameters in our experimental system, leaving only the values of the the MWC parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$) to be determined. We note that while previous studies have obtained values for K_A, K_I, and $L = e^{-\beta\Delta\varepsilon_{AI}}$ (O'Gorman et al., 1980; Daber et al., 2011), they were either based upon clever biochemical experiments or *in vivo* conditions involving poorly characterized transcription factor copy numbers and gene copy numbers. These differences relative to our experimental conditions and fitting techniques led us to believe that it was important to perform our own analysis of these parameters. Indeed, after inferring these three MWC parameters (see Supplemental Section 2.5 for details regarding the inference of $\Delta\varepsilon_{AI}$, which was fitted separately from K_A and K_I), we were able to predict the input/output response of the system under a broad range of experimental conditions. For example, this framework can predict the response of the system at different repressor copy numbers R, repressor-operator affinities $\Delta\varepsilon_{RA}$, inducer concentrations c, and gene copy numbers (see Supplemental Section 2.6).

Experimental Design

To test this model of allostery, we build off of a collection of work that has developed both a quantitative understanding of and experimental control over the simple repression motif. While other studies have investigated and parameterized induction curves of simple repression motifs, they have often relied on expression systems where the copy numbers of the various components are expressed from plasmids, resulting in highly variable and unconstrained copy numbers (Murphy et al., 2007; Daber et al., 2009; Murphy et al., 2010; Daber et al., 2011; Sochor, 2014). To adequately test the predictions of our model, we generated our own data sets in which the precise copy number of the repressor protein and the target gene copy number were known. Such a quantitative approach relies strongly on a solid foundation of previous work as depicted in Fig. 2.3. Earlier work from our laboratory used *E. coli* constructs based on components of the *lac* system to demonstrate how the Lac repressor (LacI) copy number R and operator binding energy $\Delta\varepsilon_{RA}$ affect gene expression the theory used in that work to the case of multiple promoters competing for a given transcription factor, which was demonstrated experimentally by Brewster et al. (2014), who modified this system to consider expression from multiple-copy plasmids as well as the presence of competing repressor binding sites.

Although the current work focuses on systems with a single site of repression, in Supplemental Section 2.5 we utilize data from Brewster et al. (2014), in which multiple sites of repression are explored to characterize the allosteric free energy difference $\Delta\varepsilon_{AI}$ between the repressor's active and inactive states. As explained in that section, this additional data set is critical because multiple degenerate sets of parameters can characterize an induction curve equally well, with the $\Delta\varepsilon_{AI}$ parameter compensated by the inducer dissociation constants K_A and K_I (see Fig. 2.9). After fixing $\Delta\varepsilon_{AI}$ as described in the supplemental section, the present work considers the effects of an inducer on gene expression, adding yet another means for tuning the behavior of the system. A remarkable feature of our approach is how accurately our simple model quantitatively describes the mean response of the promoter in a wide variety of regulatory contexts. We extend this body of work by introducing three additional biophysical parameters – $\Delta\varepsilon_{AI}$, K_A, and K_I – which capture the allosteric nature of the transcription factor and complement the results shown by Garcia and Phillips (2011) and Brewster et al. (2014).

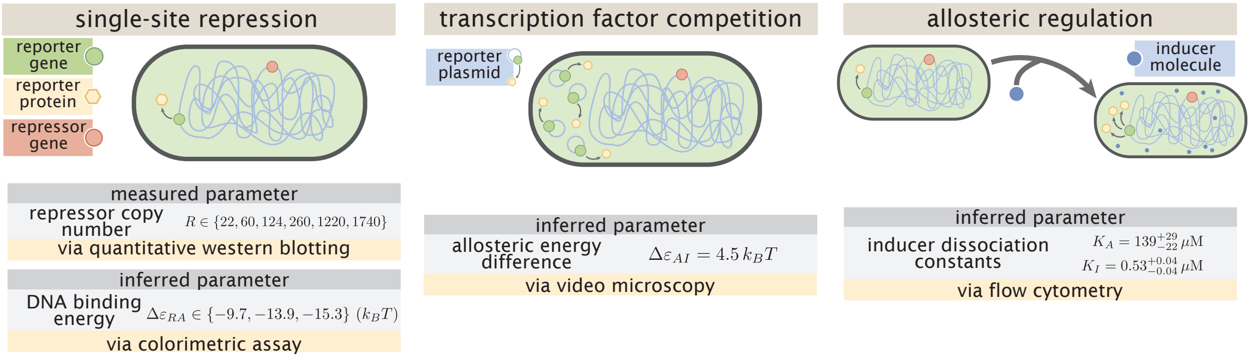

single-site repression
reporter gene
reporter protein
repressor gene
measured parameter
repressor copy number
$R \in \{22, 60, 124, 260, 1220, 1740\}$
via quantitative western blotting
inferred parameter
DNA binding energy
$\Delta\varepsilon_{RA} \in \{-9.7, -13.9, -15.3\}\ (k_BT)$
via colorimetric assay
transcription factor competition
reporter plasmid
inferred parameter
allosteric energy difference
$\Delta\varepsilon_{AI} = 4.5\, k_BT$
via video microscopy
allosteric regulation
inducer molecule
inferred parameter
inducer dissociation constants
$K_A = 139^{+29}_{-22}\, \mu\mathrm{M}$
$K_I = 0.53^{+0.04}_{-0.04}\, \mu\mathrm{M}$
via flow cytometry

Figure 2.3: **Understanding the modular components of induction.** Over the past decade, we have refined both our experimental control over and theoretical understanding of the simple repression architectures. A first round of experiments used colorimetric assays and quantitative Western blots to investigate how single-site repression is modified by the repressor copy number and repressor-DNA binding energy (Garcia and Phillips, 2011). A second round of experiments used video microscopy to probe how the copy number of the promoter and presence of competing repressor binding sites affect gene expression, and we use this data set to determine the free energy difference between the repressor's inactive and active conformations (Brewster et al., 2014) (see Supplemental Section 2.5). Both of the previous experiments characterized the system in the absence of an inducer, and in the present work we consider this additional important feature of the simple repression architecture. We used flow cytometry to determine the inducer-repressor dissociation constants and demonstrate that with these parameters we can predict *a priori* the behavior of the system for any repressor copy number, DNA binding energy, gene copy number, and inducer concentration.

To test this extension to the theory of transcriptional regulation by simple repression, we predicted the induction profiles for an array of strains that could be made using the previously characterized repressor copy number and DNA binding energies. More specifically, we used modified *lacI* ribosomal binding sites from Garcia and Phillips (2011) to generate strains with mean repressor copy number per cell of $R = 22 \pm 4$, 60 ± 20, 124 ± 30, 260 ± 40, 1220 ± 160, and 1740 ± 340, where the error denotes standard deviation of at least three replicates as measured by Garcia and Phillips (2011). We note that repressor copy number R refers to the number of repressor dimers in the cell, which is twice the number of repressor tetramers reported by Garcia and Phillips (2011); since both heads of the repressor are assumed to always be either specifically or non-specifically bound to the genome, the two repressor dimers in each LacI tetramer can be considered independently. Beyond LacI, it is suspected that most transcription factors are either bound specifically or non-specifically to DNA (Kao-Huang et al., 1977). We further assume that the rates of production and degradation for both mRNA and protein are the same across all constructs. Gene expression was measured using a Yellow Fluorescent Protein (YFP) gene, driven by a *lacUV5* promoter. Each of the six repressor copy number variants were paired with the native O1, O2, or O3 LacI operator (Oehler et al., 1994) placed at the YFP transcription start site, thereby generating eighteen unique strains. The repressor-operator binding energies (O1 $\Delta\varepsilon_{RA} = -15.3 \pm 0.2\ k_BT$, O2 $\Delta\varepsilon_{RA} = -13.9\ k_BT \pm 0.2$, and O3 $\Delta\varepsilon_{RA} = -9.7 \pm 0.1\ k_BT$) were previously inferred by measuring the fold-change of the *lac* system at different repressor copy numbers, where the error arises from model fitting (Garcia and Phillips, 2011). Additionally, we were able to obtain the value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ by fitting to previous data as discussed in Supplemental Section 2.5. We measure fold-change over a range of known IPTG concentrations c, using $n = 2$ inducer binding sites per LacI dimer and approximating the number of non-specific binding sites as the length in base-pairs of the *E. coli* genome, $N_{NS} = 4.6 \times 10^6$. We proceed by first inferring the values of the repressor-inducer dissociation constants K_A and K_I using Bayesian inferential methods as discussed below (Sivia and Skilling, 2006). When combined with the previously measured parameters within Eq. (2.5), this enables us to predict gene expression for any concentration of inducer, repressor copy number, and DNA binding energy.

Our experimental pipeline for determining fold-change using flow cytometry is shown in Fig. 2.4. Briefly, cells were grown to exponential phase, in which gene expression reaches steady state (Scott et al., 2010), under concentrations of the inducer IPTG ranging between 0 and 5 mM. We measure YFP fluorescence using flow cytometry and automatically gate the data to include only single-cell measurements (see Supplemental Section 2.7). To validate the use of flow cytometry, we also measured the fold-change of a subset of strains using the established method of single-cell microscopy (see Supplemental Section 2.8). We found that the fold-change measurements obtained from microscopy were indistinguishable from that of flow-cytometry and yielded values for the inducer binding constants K_A and K_I that were within error.

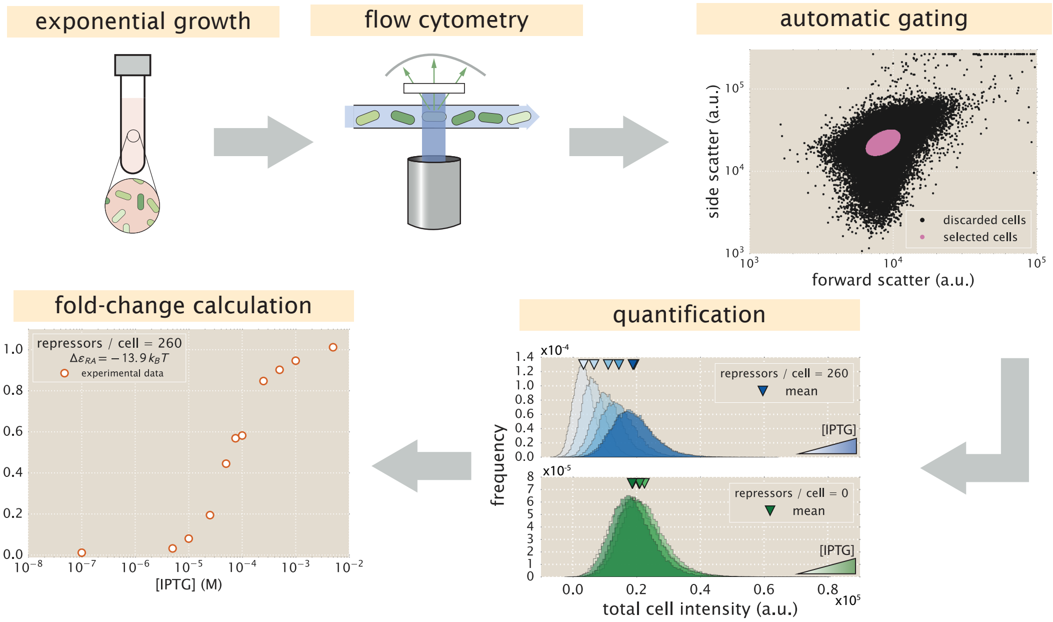


Figure 2.4: **An experimental pipeline for high-throughput fold-change measurements.** Cells are grown to exponential steady state and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single cells (see Methods). Mean expression is then quantified at different IPTG concentrations (top, blue histograms) and for a strain without repressor (bottom, green histograms), which shows no response to IPTG as expected. Fold-change is computed by dividing the mean fluorescence in the presence of repressor by the mean fluorescence in the absence of repressor.

Determination of the *in vivo* MWC Parameters

The three parameters that we tune experimentally are shown in Fig. 2.5(A), leaving the three allosteric parameters ($\Delta\varepsilon_{AI}$, K_A, and K_I) to be determined by fitting. Using previous LacI fold-change data (Brewster et al., 2014), we infer that $\Delta\varepsilon_{AI} = 4.5\ k_BT$ (see Supplemental Section 2.5). Rather than fitting K_A and K_I to our entire data set of eighteen unique constructs, we performed a Bayesian parameter estimation on the data from a single strain with $R = 260$ and an O2 operator ($\Delta\varepsilon_{RA} = -13.9\ k_BT$ Garcia and Phillips, 2011) shown in Fig. 2.5(D) (white circles). Using Markov Chain Monte

Carlo, we determine the most likely parameter values to be $K_A = 139^{+29}_{-22} \times 10^{-6}$ M and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M, which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95th percentile of the parameter value distributions as depicted in Fig. 2.5(B). Unfortunately, we are not able to make a meaningful value-for-value comparison of our parameters to those of earlier studies (Daber et al., 2009; Daber et al., 2011) because of the effects induced by uncertainties in both gene copy number and transcription factor copy numbers, the importance of which is illustrated by the plots in Supplemental Section 2.6. To demonstrate the strength of our parameter-free model, we then predicted the fold-change for the remaining seventeen strains with no further fitting (see Fig. 2.5(C)-(E)) together with the specific phenotypic properties described in Fig. 2.1 and discussed in detail below (see Fig. 2.5(F)-(J)). The shaded regions in Fig. 2.5(C)-(J) denote the 95% credible regions. An interesting aspect of our predictions of fold-change is that the width of the credible regions increases with repressor copy number and inducer concentration but decreases with the repressor-operator binding strength. Note that the fold-change Eq. (2.5) depends on the product of $\frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}$ with the MWC parameters K_A, K_I, and $\Delta \varepsilon_{AI}$. As a result, strains with small repressor copy numbers, as well as strains with weak operators such as O3, will necessarily suppress variation in the MWC parameters (see Supplemental Section 2.9).

We stress that the entire suite of predictions in Fig. 2.5 is based upon the induction profile of a single strain. Our ability to make such a broad range of predictions stems from the fact that our parameters of interest —such as the repressor copy number and DNA binding energy —appear as distinct physical parameters within our model. While the single data set in Fig. 2.5(D) could also be fit using a Hill function, such an analysis would be unable to predict any of the other curves in the figure (see Supplemental Section 2.10). Phenomenological expressions such as the Hill function can describe data, but lack predictive power and are thus unable to build our intuition, help us design *de novo* input-output functions, or guide future experiments (Kuhlman et al., 2007; Murphy et al., 2007).

Comparison of Experimental Measurements with Theoretical Predictions

We tested the predictions shown in Fig. 2.5 by measuring the fold-change induction profiles using strains that span this broad range in repressor copy numbers and repressor binding energies as characterized in Garcia and Phillips (2011), and inducer concentrations spanning several orders of magnitude. With a few provocative

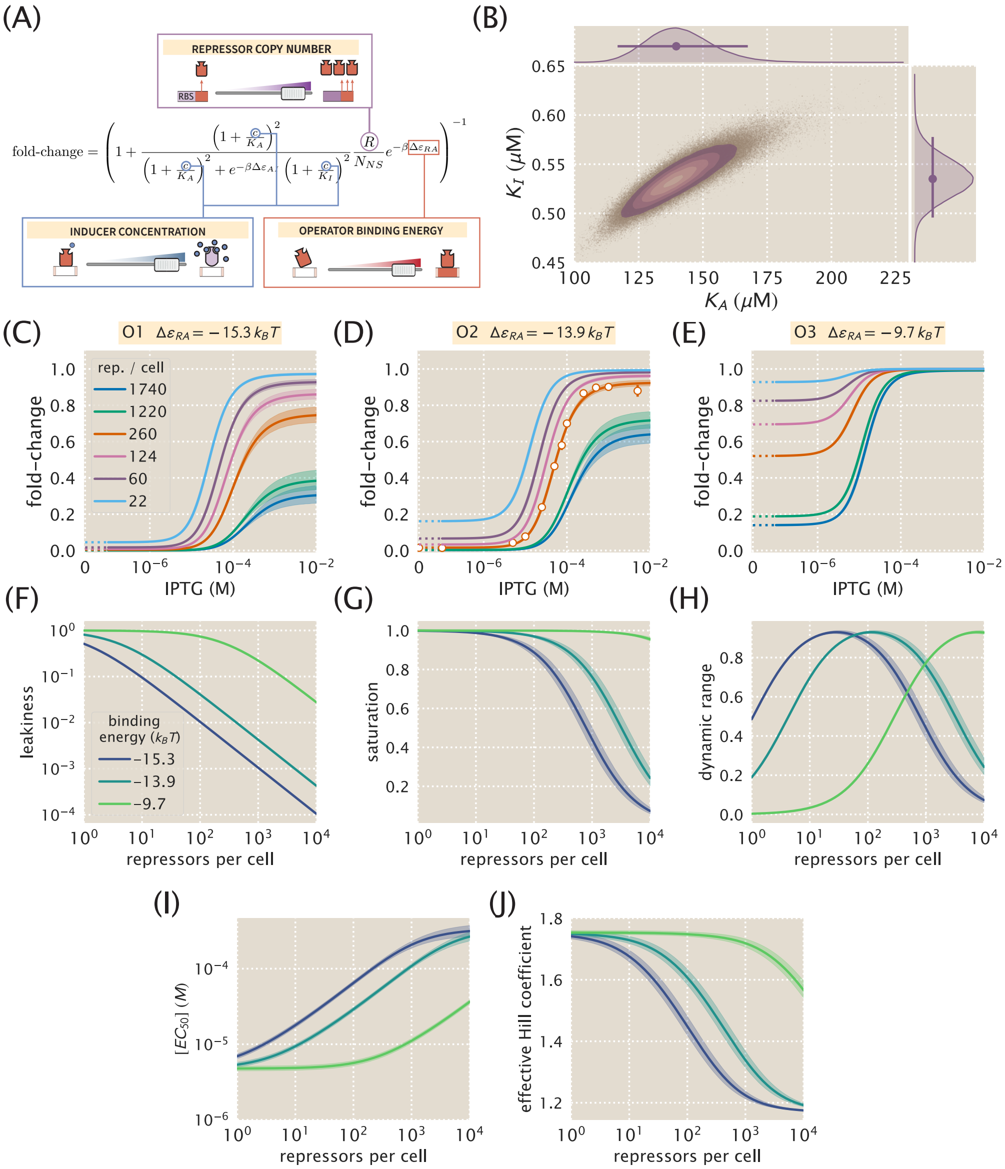


Figure 2.5: **Predicting induction profiles for different biological control parameters.** (A) We can quantitatively tune R via ribosomal binding site (RBS) modifications, $\Delta\varepsilon_{RA}$ by mutating the operator sequence, and c by adding different amounts of IPTG to the growth medium. (B) Previous experiments have characterized the R, N_{NS}, $\Delta\varepsilon_{RA}$, and $\Delta\varepsilon_{AI}$ parameters (see Fig. 2.3), leaving only the unknown dissociation constants K_A and K_I between the inducer and the repressor in the active and inactive states, respectively. These two parameters can be inferred using Bayesian parameter estimation from a single induction curve. (C-E) Predicted IPTG titration curves for different repressor copy numbers and operator strengths. Titration data for the O2 strain (white circles in Panel (D)) with $R = 260$, $\Delta\varepsilon_{RA} = -13.9\ k_BT$, $n = 2$, and $\Delta\varepsilon_{AI} = 4.5\ k_BT$ can be used to determine the thermodynamic parameters $K_A = 139^{+29}_{-22} \times 10^{-6}$ M and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M (orange line). *(Caption continued on next page)*

Figure 2.5: *(continued from previous page)* The remaining solid lines predict the fold-change Eq. (2.5) for all other combinations of repressor copy numbers (shown in the legend) and repressor-DNA binding energies corresponding to the O1 operator ($-15.3\ k_BT$), O2 operator ($-13.9\ k_BT$), and O3 operator ($-9.7\ k_BT$). Error bars of experimental data show the standard error of the mean (eight or more replicates) when this error is not smaller than the diameter of the data point. The shaded regions denote the 95% credible region, although the credible region is obscured when it is thinner than the curve itself. To display the measured fold-change in the absence of inducer, we alter the scaling of the x-axis between 0 and 10^{-7} M to linear rather than logarithmic, as indicated by a dashed line. Additionally, our model allows us to investigate key phenotypic properties of the induction profiles (see Fig. 2.1(B)). Specifically, we show predictions for the (F) leakiness, (G) saturation, (H) dynamic range, (I) $[EC_{50}]$, and (J) effective Hill coefficient of the induction profiles.

exceptions, the results shown in Fig. 2.6 demonstrate a consistent agreement between theory and experiment. We note that there was an apparently systematic shift in the O3 $\Delta\varepsilon_{RA} = -9.7\ k_BT$ strains (Fig. 2.6(C)) and all of the $R = 1220$ and $R = 1740$ strains. This may be partially due to imprecise previous determinations of their $\Delta\varepsilon_{RA}$ and R values. By performing a global fit where we infer all parameters including the repressor copy number R and the binding energy $\Delta\varepsilon_{RA}$, we found better agreement for these particular strains, although a discrepancy in the steepness of the response for all O3 strains remains (see Supplemental Section 2.11). We considered a number of plausible hypotheses for the origins of these discrepancies such as including other states (e.g. non-negligible binding of the inactive repressor), relaxing the weak promoter approximation, and accounting for variations in gene and repressor copy number throughout the cell cycle, but none explained the observed differences. As an additional test of our model, we also considered strains using the synthetic Oid operator which exhibits stronger repression, $\Delta\varepsilon_{RA} = -17\ k_BT$ (Garcia and Phillips, 2011), than the O1, O2, and O3 operators. The global fit agrees well with the Oid microscopy data, though it asserts a stronger Oid binding energy of $\Delta\varepsilon_{RA} = -17.7\ k_BT$ (see Supplemental Section 2.12 for more details).

To ensure that the agreement between our predictions and data is not an accident of the strain we chose to perform our fitting, we explored the effects of using each of our other strains to estimate K_A and K_I. As shown in Supplemental Section 2.13 and Fig. 2.6(D), the inferred values of K_A and K_I depend very minimally upon which strain is chosen, demonstrating that these parameter values are highly robust. As previously mentioned, we performed a global fit using the data from all eighteen strains for the following parameters: the inducer dissociation constants K_A and K_I,

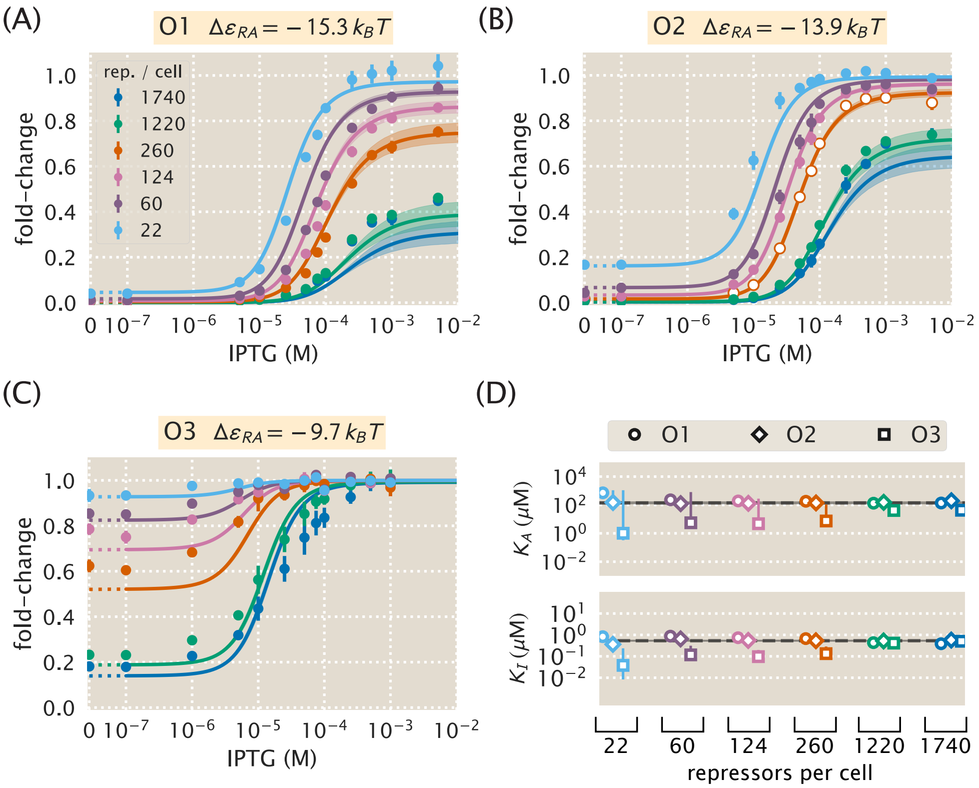


Figure 2.6: **Comparison of predictions against measured and inferred data.** Flow cytometry measurements of fold-change over a range of IPTG concentrations for (A) O1, (B) O2, and (C) O3 strains at varying repressor copy numbers, overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Fig. 2.5, all of the predicted induction curves were generated prior to measurement by inferring the MWC parameters using a single data set (O2 $R = 260$, shown by white circles in Panel (B)). The predictions may therefore depend upon which strain is used to infer the parameters. (D) The inferred parameter values of the dissociation constants K_A and K_I using any of the eighteen strains instead of the O2 $R = 260$ strain. Nearly identical parameter values are inferred from each strain, demonstrating that the same set of induction profiles would have been predicted regardless of which strain was chosen. The points show the mode, and the error bars denote the 95% credible region of the parameter value distribution. Error bars not visible are smaller than the size of the marker.

the repressor copy numbers R, and the repressor DNA binding energy $\Delta\varepsilon_{RA}$ (see Supplemental Section 2.11), and the resulting parameter values were nearly identical to those measured from a single strain. For the remainder of the text we proceed using our analysis on the strain with $R = 260$ repressors and an O2 operator.

Predicting the Phenotypic Traits of the Induction Response

Rather than measuring the full induction response of a system, a subset of the properties shown in Fig. 2.1, namely, the leakiness, saturation, dynamic range, $[EC_{50}]$, and effective Hill coefficient, may be of greater interest. For example, synthetic biology is often focused on generating large responses (i.e. a large dynamic range) or finding a strong binding partner (i.e. a small $[EC_{50}]$) (Brophy and Voigt, 2014; Shis et al., 2014). While these properties are all individually informative, when taken together they capture the essential features of the induction response. We reiterate that a Hill function approach cannot predict these features *a priori* and furthermore requires fitting each curve individually. The MWC model, on the other hand, enables us to quantify how each trait depends upon a single set of physical parameters as shown by Fig. 2.5(F)-(J).

We define these five phenotypic traits using expressions derived from the model, Eq. (2.5). These results build upon extensive work by Martins and Swain (2011), who computed many such properties for ligand-receptor binding within the MWC model. We begin by analyzing the leakiness, which is the minimum fold-change observed in the absence of ligand, given by

$$\begin{aligned} \text{leakiness} &= \text{fold-change}(c = 0) \\ &= \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}, \end{aligned} \tag{2.6}$$

and the saturation, which is the maximum fold change observed in the presence of saturating ligand,

$$\begin{aligned} \text{saturation} &= \text{fold-change}(c \to \infty) \\ &= \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}. \end{aligned} \tag{2.7}$$

Systems that minimize leakiness repress strongly in the absence of effector while systems that maximize saturation have high expression in the presence of effector. Together, these two properties determine the dynamic range of a system's response, which is given by the difference

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \tag{2.8}$$

These three properties are shown in Fig. 2.5(F)-(H). We discuss these properties in greater detail in Supplemental Section 2.14. For example, we compute the number of repressors R necessary to evoke the maximum dynamic range and demonstrate that the magnitude of this maximum is independent of the repressor-operator binding energy $\Delta\varepsilon_{RA}$. Fig. 2.7(A)-(C) shows that the measurements of these three properties, derived from the fold-change data in the absence of IPTG and the presence of saturating IPTG, closely match the predictions for all three operators.

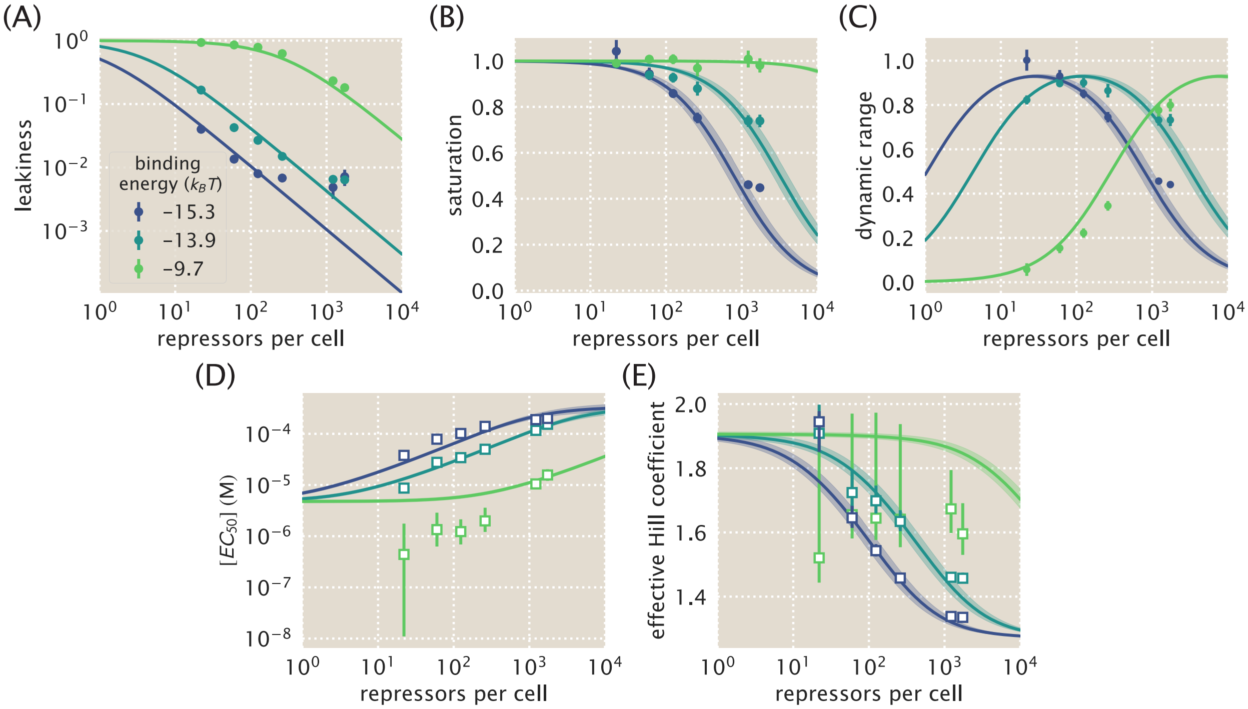


Figure 2.7: **Predictions and experimental measurements of key properties of induction profiles.** Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 2.6 in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-operator binding energies in the legend correspond to the O1 operator ($-15.3\ k_BT$), O2 operator ($-13.9\ k_BT$), and O3 operator ($-9.7\ k_BT$). Both the (D) $[EC_{50}]$ and (E) effective Hill coefficient are inferred by individually fitting each operator-repressor pairing in Fig. 2.6(A)-(C) separately to Eq. (2.5) in order to smoothly interpolate between the data points. Error bars for (A)-(C) represent the standard error of the mean for eight or more replicates; error bars for (D)-(E) represent the 95% credible region for the parameter found by propagating the credible region of our estimates of K_A and K_I into Eqs. (2.9) and (2.10).

Two additional properties of induction profiles are the $[EC_{50}]$ and effective Hill coefficient, which determine the range of inducer concentration in which the system's output goes from its minimum to maximum value. The $[EC_{50}]$ denotes the inducer concentration required to generate a system response Eq. (2.5) halfway between its minimum and maximum value,

$$\text{fold-change}(c = [EC_{50}]) = \frac{\text{leakiness} + \text{saturation}}{2}. \tag{2.9}$$

The effective Hill coefficient h, which quantifies the steepness of the curve at the $[EC_{50}]$ (Marzen et al., 2013), is given by

$$h = \left(2 \frac{d}{d \log c} \left[\log \left(\frac{\text{fold-change}(c) - \text{leakiness}}{\text{dynamic range}} \right) \right] \right)_{c=[EC_{50}]}. \tag{2.10}$$

Fig. 2.5(I)-(J) shows how the $[EC_{50}]$ and effective Hill coefficient depend on the repressor copy number. In Supplemental Section 2.14, we discuss the analytic forms of these two properties as well as their dependence on the repressor-DNA binding energy.

Fig. 2.7(D)-(E) shows the estimated values of the $[EC_{50}]$ and the effective Hill coefficient overlaid on the theoretical predictions. Both properties were obtained by fitting Eq. (2.5) to each individual titration curve and computing the $[EC_{50}]$ and effective Hill coefficient using Eq. (2.9) and Eq. (2.10), respectively. We find that the predictions made with the single strain fit closely match those made for each of the strains with O1 and O2 operators, but the predictions for the O3 operator are markedly off. In Supplemental Section 2.10, we show that the large, asymmetric error bars for the O3 $R = 22$ strain arise from its nearly flat response, where the lack of dynamic range makes it impossible to determine the value of the inducer dissociation constants K_A and K_I, as can be seen in the uncertainty of both the $[EC_{50}]$ and effective Hill coefficient. Discrepancies between theory and data for O3 are improved, but not fully resolved, by performing a global fit or fitting the MWC model individually to each curve (see Appendices 2.11 and 2.13). It remains an open question how to account for discrepancies in O3, in particular regarding the significant mismatch between the predicted and fitted effective Hill coefficients.

Data Collapse of Induction Profiles

Our primary interest heretofore was to determine the system response at a specific inducer concentration, repressor copy number, and repressor-DNA binding energy. However, the cell does not necessarily “care about” the precise number of repressors in the system or the binding energy of an individual operator. The relevant quantity for cellular function is the fold-change enacted by the regulatory system. This raises the question: given a specific value of the fold-change, what combination of parameters will give rise to this desired response? In other words, what trade-offs between the parameters of the system will give rise to the same mean cellular output? These are key questions both for understanding how the system is governed and for engineering specific responses in a synthetic biology context. To address these questions, we follow the data collapse strategy used in a number of previous studies (Sourjik and Berg, 2002; Keymer et al., 2006; Swem et al., 2008), and rewrite Eq. (2.5) as a Fermi function,

$$\text{fold-change} = \frac{1}{1 + e^{-F(c)}}, \tag{2.11}$$

where $F(c)$ is the free energy of the repressor binding to the operator of interest relative to the unbound operator state in k_BT units (Keymer et al., 2006; Swem et al., 2008; Phillips, 2015), which is given by

$$F(c) = \frac{\Delta\varepsilon_{RA}}{k_BT} - \log \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} - \log \frac{R}{N_{NS}}. \tag{2.12}$$

The first term in $F(c)$ denotes the repressor-operator binding energy, the second the contribution from the inducer concentration, and the last the effect of the repressor copy number. We note that elsewhere, this free energy has been dubbed the Bohr parameter since such families of curves are analogous to the shifts in hemoglobin binding curves at different pHs known as the Bohr effect (Mirny, 2010; Phillips, 2015; Einav et al., 2016).

Instead of analyzing each induction curve individually, the free energy provides a natural means to simultaneously characterize the diversity in our eighteen induction profiles. Fig. 2.8(A) demonstrates how the various induction curves from Fig. 2.5(C)-(E) all collapse onto a single master curve, where points from every induction profile that yield the same fold-change are mapped onto the same free energy. Fig. 2.8(B) shows this data collapse for the 216 data points in Fig. 2.6(A)-(C), demonstrating the close match between the theoretical predictions and experimental measurements across all eighteen strains.

There are many different combinations of parameter values that can result in the same free energy as defined in Eq. (2.12). For example, suppose a system originally has a fold-change of 0.2 at a specific inducer concentration, and then operator mutations increase the $\Delta\varepsilon_{RA}$ binding energy (Garcia et al., 2012). While this serves to initially increase both the free energy and the fold-change, a subsequent increase in the repressor copy number could bring the cell back to the original fold-change level. Such trade-offs hint that there need not be a single set of parameters that evoke a specific cellular response, but rather that the cell explores a large but degenerate space of parameters with multiple, equally valid paths.

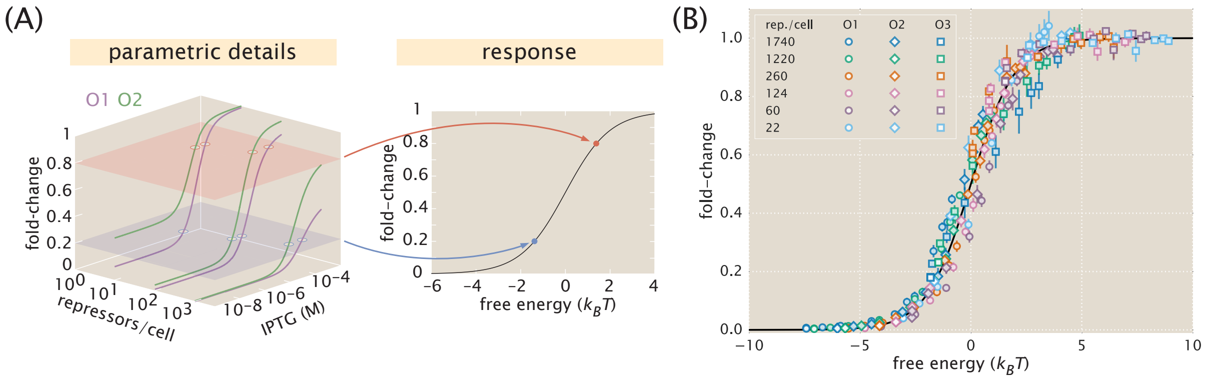


Figure 2.8: **Fold-change data from a broad collection of different strains collapse onto a single master curve.** (A) Any combination of parameters can be mapped to a single physiological response (i.e. fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from Fig. 2.6 collapse onto a single master curve as a function of the free energy Eq. (2.12). The free energy for each strain was calculated from Eq. (2.12) using $n = 2$, $\Delta\varepsilon_{AI} = 4.5\ k_BT$, $K_A = 139 \times 10^{-6}$ M, $K_I = 0.53 \times 10^{-6}$ M, and the strain-specific R and $\Delta\varepsilon_{RA}$. All data points represent the mean, and error bars are the standard error of the mean for eight or more replicates.

2.3 Discussion

Since the early work by Monod, Wyman, and Changeux (Monod et al., 1963; Monod et al., 1965), a broad list of different biological phenomena have been tied to the existence of macromolecules that switch between inactive and active states. Examples can be found in a wide variety of cellular processes that include ligand-gated ion channels (Auerbach, 2012), enzymatic reactions (Velyvis et al., 2007; Einav et al., 2016), chemotaxis (Keymer et al., 2006), quorum sensing (Swem et al., 2008), G-protein coupled receptors (Canals et al., 2012), physiologically important proteins (Milo et al., 2007; Levantino et al., 2012), and beyond. One of the most ubiquitous examples of allostery is in the context of gene expression, where an array of molecular players bind to transcription factors to either aid or deter their ability to regulate gene activity (Huang et al., 2011; Li et al., 2014). Nevertheless, no definitive study has been made of the applicability of the MWC model to transcription factor function, despite the clear presence of different conformational states in their structures in the presence and absence of signaling molecules (Lewis et al., 1996). The purpose of this work was to derive an analytical framework for allosteric transcriptional regulation that is sufficiently general to quantitatively predict gene expression across a variety of regulatory architectures. We then rigorously tested this framework and its predictions upon the specific case of simple repression using the *lac* system.

Others have developed quantitative models describing different aspects of allosteric regulatory systems. Martins and Swain (2011) and Marzen et al. (2013) analytically derived fundamental properties of the MWC model, including the leakiness and dynamic range described in this work, noting the inherent trade-offs in these properties when tuning the microscopic parameters of the model. Work in the Church and Voigt labs, among others, has expanded on the availability of allosteric circuits for synthetic biology (Lutz and Bujard, 1997; Moon et al., 2012; Rogers et al., 2015; Rohlhill et al., 2017). Recently, Daber *et al.* theoretically explored the induction of simple repression within the MWC model (Daber et al., 2009) and experimentally measured how mutations alter the induction profiles of transcription factors (Daber et al., 2011). Vilar and Saiz considered the broad range of interactions in inducible *lac*-based systems including the effects of oligomerization and DNA folding on transcription factor induction (Saiz and Vilar, 2008; Vilar and Saiz, 2013). Other work has attempted to use the *lac* system to reconcile *in vitro* and *in vivo* measurements (Tungtur et al., 2011; Sochor, 2014). Although this body of work has done much to improve our understanding of allosteric transcription factors, there has remained a disconnect between model and experiment. In order to rigorously test a model's applicability to natural systems, the model's predictions must be weighed against data from precise experiments specifically designed to test those predictions.

Here, we expand upon this body of work by generating a predictive model of allosteric transcriptional regulation and then testing the model against a thorough set of experiments using well-characterized regulatory components. Specifically, we used the MWC model to build upon and refine a well-established thermodynamic model of transcriptional regulation (Bintu et al., 2005; Garcia and Phillips, 2011), allowing us to compose the model from a minimal set of biologically meaningful parameters. This minimal model captures the key players of transcriptional regulation – namely the repressor copy number, the DNA binding energy, and the concentration of inducer – and enables us to predict how the system will behave when we change each of these parameters. We tested these predictions on a range of strains whose repressor copy number spanned two orders of magnitude and whose DNA binding affinity spanned 6 k_BT. We argue that one would not be able to generate such a wide array of predictions by using a Hill function, which abstracts away the biophysical meaning of the parameters into phenomenological parameters (Forsén and Linse, 1995).

Specifically, we tested our model in the context of a *lac*-based simple repression system by first determining the allosteric dissociation constants K_A and K_I from a single induction data set (O2 operator with binding energy $\Delta\varepsilon_{RA} = -13.9\ k_BT$ and repressor copy number $R = 260$) and then using these values to make parameter-free predictions of the induction profiles for seventeen other strains where $\Delta\varepsilon_{RA}$ and R were varied significantly (see Fig. 2.5). We next measured the induction profiles of these seventeen strains using flow cytometry and found that our predictions consistently and accurately captured the primary features for each induction data set, as shown in Fig. 2.6(A)-(C). Surprisingly, we find that the inferences for the repressor-inducer dissociation constants that would have been derived from any other single strain (instead of the O2 operator with $R = 260$) would have resulted in nearly identical predictions (see Fig. 2.6(D) and Appendix 2.13). This suggests that a few carefully chosen measurements can lead to a deep quantitative understanding of how simple regulatory systems work without requiring an extensive sampling of strains that span the parameter space. Moreover, the fact that we could consistently achieve reliable predictions after fitting only two free parameters stands in contrast to the common practice of fitting several free parameters simultaneously, which can nearly guarantee an acceptable fit provided that the model roughly resembles the system response, regardless of whether the details of the model are tied to any underlying molecular mechanism.

Beyond observing changes in fold-change as a function of effector concentration, our application of the MWC model allows us to explicitly predict the values of the induction curves' key parameters, namely, the leakiness, saturation, dynamic range, $[EC_{50}]$, and the effective Hill coefficient (see Fig. 2.7). This allows us to quantify the unique traits of each set of strains examined here. Strains using the O1 operator consistently have a low leakiness value, a consequence of the operator's strong binding energy. The saturation values for these strains, however, vary significantly with R. This trend is reversed for strains using O3, which has the weakest binding energy of our constructs. Leakiness values for constructs using O3 vary strongly with R, but their saturation values approach 1 regardless of R. Strains with the intermediate O2 binding energy have both a leakiness and saturation that vary markedly with R. For both the O1 and O2 data sets, our model also accurately predicts the effective Hill coefficient and $[EC_{50}]$, though these predictions for O3 are noticeably less accurate. While performing a global fit for all model parameters marginally improves the prediction for O3 (see Appendix 2.11), we are still unable to accurately predict the effective Hill coefficient or the $[EC_{50}]$. We further tried including additional states (such as allowing the inactive repressor to bind to the operator), relaxing the weak promoter approximation, accounting for changes in gene and repressor copy number throughout the cell cycle (Jones et al., 2014), and refitting the original binding energies from Garcia et al. (2011), but we were still unable to account for the O3 data. It remains an open question as to how the discrepancy between the theory and measurements for O3 can be reconciled.

Because this model allows us to derive expressions for individual features of induction curves, we are able to examine how these features may be tuned by careful selection of system parameters. Fig. 2.7 shows how each of the induction curves' key features vary as a function of $\Delta\varepsilon_{RA}$ and R, which makes it possible to select desired properties from among the possible phenotypes available to the system. For instance, it is possible to obtain a high dynamic range using fewer than 100 repressors if the binding energy is strong. As an example of the constraints inherent to the system, one cannot design a strain with a leakiness of 0.1 and a saturation of 0.4 by only varying the repressor copy number and repressor-operator binding affinity, since these two properties are coupled by Eqs. (2.6) and (2.7). Achieving this particular behavior would require changing the ratio K_A/K_I of repressor-inducer dissociation constants, as may be done by mutating the repressor's inducer binding pocket.

The dynamic range, which is of considerable interest when designing or characterizing a genetic circuit, is revealed to have an interesting property: although changing the value of $\Delta\varepsilon_{RA}$ causes the dynamic range curves to shift to the right or left, each curve has the same shape and in particular the same maximum value. This means that strains with strong or weak binding energies can attain the same dynamic range when the value of R is tuned to compensate for this energy. This feature is not immediately apparent from the IPTG induction curves, which show very low dynamic ranges for several of the O1 and O3 strains. Without the benefit of models that can predict such phenotypic traits, efforts to engineer genetic circuits with allosteric transcription factors must rely on trial and error to achieve specific responses (Rogers et al., 2015; Rohlhill et al., 2017). This is a compelling example showing that our predictive modeling approach has a significant advantage over descriptive models.

To our knowledge this is the first work of its kind in which a single family of parameters is demonstrated to predict a vast range of induction curves with qualitatively different behaviors. One of the demanding criteria of our approach is that a small set of parameters must consistently describe data from a diverse collection of data sets taken using distinct methods such as Miller assays and bulk and single-cell fluorescence experiments to measure fold-change (see Appendices 2.7 and 2.12), as well as quantitative Western blots (Garcia and Phillips, 2011) and binomial partitioning methods to count repressors (Rosenfeld et al., 2005; Brewster et al., 2014) Furthermore, we build off of our previous studies that use the simple repression architecture and we demand that the parameters derived from these studies account for constructs that are integrated into the chromosome, plasmid-borne, and even for cases where there are competing binding sites to take repressors out of circulation (Garcia and Phillips, 2011; Brewster et al., 2014) (see Appendix 2.6) or where there are multiple operators to allow DNA looping (Boedicker et al., 2013b). The resulting model not only predicts the individual titration profiles as a function of IPTG, but describes key properties of the response. The general agreement with the entire body of work presented here demonstrates that our model captures the underlying mechanism governing simple repression. We are unaware of any comparable study in transcriptional regulation that demands one predictive framework cover such a broad array of regulatory situations.

Despite the diversity observed in the induction profiles of each of our strains, our data are unified by their reliance on fundamental biophysical parameters. In particular, we have shown that our model for fold-change can be rewritten in terms of the free energy Eq. (2.12), which encompasses all of the physical parameters of the system. This has proven to be an illuminating technique in a number of studies of allosteric proteins (Sourjik and Berg, 2002; Keymer et al., 2006; Swem et al., 2008). Although it is experimentally straightforward to observe system responses to changes in effector concentration c, framing the input-output function in terms of c can give the misleading impression that changes in system parameters lead to fundamentally altered system responses. Alternatively, if one can find the ‘natural variable’ that enables the output to collapse onto a single curve, it becomes clear that the system’s output is not governed by individual system parameters, but rather the contributions of multiple parameters that define the natural variable.

When our fold-change data are plotted against the respective free energies for each construct, they collapse cleanly onto a single curve (see Fig. 2.8). This enables us to analyze how parameters can compensate each other. For example, we may wish to determine which combinations of parameters result in a system that is strongly repressed (free energy $F(c) \leq -5\ k_BT$). We know from our understanding of the induction phenomenon that strong repression is most likely to occur at low values of c. However, from Eq. (2.12) we can clearly see that increases in the value of c can be compensated by an increase in the number of repressors R, a decrease in the binding energy $\Delta\varepsilon_{RA}$ (i.e. stronger binding), or some combination of both. Likewise, while the system tends to express strongly ($F(c) \geq 5\ k_BT$) when c is high, one could design a system that expresses strongly at low values of c by reducing R or increasing the value of $\Delta\varepsilon_{RA}$. As a concrete example, given a concentration $c = 10^{-5}$ M, a system using the O1 operator ($\Delta\varepsilon_{RA} = -15.3\ k_BT$) requires 745 or more repressors for $F(c) \leq -5\ k_BT$, while a system using the weaker O3 operator ($\Delta\varepsilon_{RA} = -9.7\ k_BT$) requires 2×10^5 or more repressors for $F(c) \leq -5\ k_BT$.

While our experiments validated the theoretical predictions in the case of simple repression, we expect the framework presented here to apply much more generally to different biological instances of allosteric regulation. For example, we can use this model to study more complex systems such as when transcription factors interact with multiple operators (Bintu et al., 2005). We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli* (see Appendix 2.15). This work can also serve as a springboard to characterize not just the mean but the full gene expression distribution and thus quantify the impact of noise on the system (Eldar and Elowitz, 2010). Another extension of this approach would be to theoretically predict and experimentally verify whether the repressor-inducer dissociation constants K_A and K_I or the energy difference $\Delta\varepsilon_{AI}$ between the allosteric states can be tuned by making single amino acid substitutions in the transcription factor (Daber et al., 2011; Phillips, 2015). Finally, we expect that the kind of rigorous quantitative description of the allosteric phenomenon provided here will make it possible to construct biophysical models of fitness for allosteric proteins similar to those already invoked to explore the fitness effects of transcription factor binding site strengths and protein stability (Gerland and Hwa, 2002; Berg et al., 2004; Zeldovich and Shakhnovich, 2008).

To conclude, we find that our application of the MWC model provides an accurate, predictive framework for understanding simple repression by allosteric transcription factors. To reach this conclusion, we analyzed the model in the context of a well-characterized system, in which each parameter had a clear biophysical meaning. As many of these parameters had been measured or inferred in previous studies, this gave us a minimal model with only two free parameters which we inferred from a single data set. We then accurately predicted the behavior of seventeen other data sets in which repressor copy number and repressor-DNA binding energy were systematically varied. In addition, our model allowed us to understand how key properties such as the leakiness, saturation, dynamic range, $[EC_{50}]$, and effective Hill coefficient depended upon the small set of parameters governing this system. Finally, we show that by framing inducible simple repression in terms of free energy, the data from all of our experimental strains collapse cleanly onto a single curve, illustrating the many ways in which a particular output can be targeted. In total, these results show that a thermodynamic formulation of the MWC model supersedes phenomenological fitting functions for understanding transcriptional regulation by allosteric proteins.

2.4 Methods

Bacterial Strains and DNA Constructs

All strains used in these experiments were derived from *E. coli* K12 MG1655 with the *lac* operon removed, adapted from those created and described in Garcia and Phillips (2011) and Garcia et al. (2011). Briefly, the operator variants and YFP reporter gene were cloned into a pZS25 background which contains a *lacUV5* promoter that drives expression as is shown schematically in Fig. 2.2. These constructs carried a kanamycin resistance gene and were integrated into the *galK* locus of the chromosome using λ Red recombineering (Sharan et al., 2009). The *lacI* gene was constitutively expressed via a $\text{P}_{\text{LtetO-1}}$ promoter (Lutz and Bujard, 1997), with ribosomal binding site mutations made to vary the LacI copy number as described in Salis et al. (2009) using site-directed mutagenesis (Quickchange II; Stratagene), with further details in Garcia and Phillips (2011). These *lacI* constructs carried a chloramphenicol resistance gene and were integrated into the *ybcN* locus of the chromosome. Final strain construction was achieved by performing repeated P1 transduction (Thomason et al., 2007) of the different operator and *lacI* constructs to generate each combination used in this work. Integration was confirmed by PCR amplification of the replaced chromosomal region and by sequencing. Primers and final strain genotypes are listed in Supplemental Section 2.16.

It is important to note that the rest of the *lac* operon (*lacZYA*) was never expressed. The LacY protein is a transmembrane protein which actively transports lactose as well as IPTG into the cell. As LacY was never produced in our strains, we assume that the extracellular and intracellular IPTG concentration was approximately equal due to diffusion across the membrane into the cell as is suggested by previous work (Fernández-Castané et al., 2012).

To make this theory applicable to transcription factors with any number of DNA binding domains, we used a different definition for repressor copy number than has been used previously. We define the LacI copy number as the average number of repressor dimers per cell whereas in Garcia and Phillips (2011), the copy number is defined as the average number of repressor tetramers in each cell. To motivate this decision, we consider the fact that the LacI repressor molecule exists as a tetramer in *E. coli* (Lewis et al., 1996) in which a single DNA binding domain is formed from dimerization of LacI proteins, so that wild-type LacI might be described as a dimer of dimers. Since each dimer is allosterically independent (i.e. either dimer can be allosterically active or inactive, independent of the configuration of the other dimer) (Daber et al., 2009), a single LacI tetramer can be treated as two functional repressors. Therefore, we have simply multiplied the number of repressors reported in Garcia and Phillips (2011) by a factor of two. This factor is included as a keyword argument in the numerous Python functions used to perform this analysis, as discussed in the code documentation.

A subset of strains in these experiments were measured using fluorescence microscopy for validation of the flow cytometry data and results. To aid in the high-fidelity segmentation of individual cells, the strains were modified to constitutively express an mCherry fluorophore. This reporter was cloned into a pZS4*1 backbone (Lutz and Bujard, 1997) in which mCherry is driven by the *lacUV5* promoter. All microscopy and flow cytometry experiments were performed using these strains.

Growth Conditions for Flow Cytometry Measurements

All measurements were performed with *E. coli* cells grown to mid-exponential phase in standard M9 minimal media (M9 5X Salts, Sigma-Aldrich M6030; 2 mM magnesium sulfate, Mallinckrodt Chemicals 6066-04; 100 μM calcium chloride, Fisher Chemicals C79-500) supplemented with 0.5% (w/v) glucose. Briefly, 500 μL cultures of *E. coli* were inoculated into Lysogeny Broth (LB Miller Powder, BD Medical) from a 50% glycerol frozen stock (-80°C) and were grown overnight in a 2 mL 96-deep-well plate sealed with a breathable nylon cover (Lab Pak - Nitex Nylon, Sefar America Inc. Cat. No. 241205) with rapid agitation for proper aeration. After approximately 12 to 15 hours, the cultures had reached saturation and were diluted 1000-fold into a second 2 mL 96-deep-well plate where each well contained 500 μL of M9 minimal media supplemented with 0.5% w/v glucose (anhydrous D-Glucose, Macron Chemicals) and the appropriate concentration of IPTG (Isopropyl β-D-1 thiogalactopyranoside Dioxane Free, Research Products International). These were sealed with a breathable cover and were allowed to grow for approximately eight hours. Cells were then diluted ten-fold into a round-bottom 96-well plate (Corning Cat. No. 3365) containing 90 μL of M9 minimal media supplemented with 0.5% w/v glucose along with the corresponding IPTG concentrations. For each IPTG concentration, a stock of 100-fold concentrated IPTG in double distilled water was prepared and partitioned into 100 μL aliquots. The same parent stock was used for all experiments described in this work.

Flow Cytometry

Unless explicitly mentioned, all fold-change measurements were collected on a Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer graciously provided by the Pamela Björkman lab at Caltech. Detailed information regarding the voltage settings of the photo-multiplier detectors can be found in Table 2.1. Prior to each day's experiments, the analyzer was calibrated using MACSQuant Calibration Beads (Cat. No. 130-093-607) such that day-to-day experiments would be comparable. All YFP fluorescence measurements were collected via 488 nm laser excitation coupled with a 525/50 nm emission filter. Unless otherwise specified, all measurements were taken over the course of two to three hours using automated sampling from a 96-well plate kept at approximately 4° - 10°C on a MACS Chill 96 Rack (Cat. No. 130-094-459). Cells were diluted to a final concentration of approximately 4×10^4 cells per μL which corresponded to a flow rate of 2,000-6,000 measurements per second, and acquisition for each well was halted after 100,000 events were detected. Once completed, the data were extracted and immediately processed using the following methods.

Unsupervised Gating of Flow Cytometry Data

Flow cytometry data will frequently include a number of spurious events or other undesirable data points such as cell doublets and debris. The process of restricting the collected data set to those data determined to be "real" is commonly referred to as gating. These gates are typically drawn manually (Maecker et al., 2005) and restrict the data set to those points which display a high degree of linear correlation between their forward-scatter (FSC) and side-scatter (SSC). The development of unbiased and unsupervised methods of drawing these gates is an active area of research (Lo et al., 2008; Aghaeepour et al., 2013). For our purposes, we assume that the fluorescence level of the population should be log-normally distributed about some mean value. With this assumption in place, we developed a method that allows us to restrict the data used to compute the mean fluorescence intensity of the population to the smallest two-dimensional region of the log(FSC) vs. log(SSC) space in which 40% of the data is found. This was performed by fitting a bivariate Gaussian distribution and restricting the data used for calculation to those that reside within the 40th percentile. This procedure is described in more detail in the supplementary information as well as in a Jupyter notebook located in this paper's Github repository (https://www.github.com/rpgroup-pboc/mwc_induction).

Experimental Determination of Fold-Change

For each strain and IPTG concentration, the fold-change in gene expression was calculated by taking the ratio of the population mean YFP expression in the presence of LacI repressor to that of the population mean in the absence of LacI repressor. However, the measured fluorescence intensity of each cell also includes the autofluorescence contributed by the weak excitation of the myriad protein and small molecules within the cell. To correct for this background, we computed the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{2.13}$$

where $\langle I_{R>0} \rangle$ is the average cell YFP intensity in the presence of repressor, $\langle I_{R=0} \rangle$ is the average cell YFP intensity in the absence of repressor, and $\langle I_{\text{auto}} \rangle$ is the average cell autofluorescence intensity, as measured from cells that lack the *lac*-YFP construct.

Bayesian Parameter Estimation

In this work, we determine the the most likely parameter values for the inducer dissociation constants K_A and K_I of the active and inactive state, respectively, using Bayesian methods. We compute the probability distribution of the value of each parameter given the data D, which by Bayes' theorem is given by

$$P(K_A, K_I \mid D) = \frac{P(D \mid K_A, K_I) P(K_A, K_I)}{P(D)}, \tag{2.14}$$

where D is all the data composed of independent variables (repressor copy number R, repressor-DNA binding energy $\Delta\varepsilon_{RA}$, and inducer concentration c) and one dependent variable (experimental fold-change). $P(D \mid K_A, K_I)$ is the likelihood of having observed the data given the parameter values for the dissociation constants, $P(K_A, K_I)$ contains all the prior information on these parameters, and $P(D)$ serves as a normalization constant. Eq. (2.5) assumes a deterministic relationship between the parameters and the data, so in order to construct a probabilistic relationship as required by Eq. (2.14), we assume that the experimental fold-change for the i^{th} datum given the parameters is of the form

$$\text{fold-change}^{(i)}_{\text{exp}} = \left(1 + \frac{\left(1+\frac{c^{(i)}}{K_A}\right)^2}{\left(1+\frac{c^{(i)}}{K_A}\right)^2 + e^{-\beta\Delta\varepsilon_{AI}}\left(1+\frac{c^{(i)}}{K_I}\right)^2}\frac{R^{(i)}}{N_{NS}}e^{-\beta\Delta\varepsilon^{(i)}_{RA}}\right)^{-1} + \epsilon^{(i)}, \tag{2.15}$$

where $\epsilon^{(i)}$ represents the departure from the deterministic theoretical prediction for the i^{th} data point. If we assume that these $\epsilon^{(i)}$ errors are normally distributed with mean zero and standard deviation σ, the likelihood of the data given the parameters is of the form

$$P(D|K_A, K_I, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\prod_{i=1}^{n}\exp\left[-\frac{(\text{fold-change}^{(i)}_{\text{exp}} - \text{fold-change}(K_A,K_I,R^{(i)},\Delta\varepsilon^{(i)}_{RA},c^{(i)}))^2}{2\sigma^2}\right], \tag{2.16}$$

where fold-change$^{(i)}_{\text{exp}}$ is the experimental fold-change and fold-change$(\cdots)$ is the theoretical prediction. The product $\prod_{i=1}^{n}$ captures the assumption that the n data points are independent. Note that the likelihood and prior terms now include the extra unknown parameter σ. In applying Eq. (2.16), a choice of K_A and K_I that provides better agreement between theoretical fold-change predictions and experimental measurements will result in a more probable likelihood.

Both mathematically and numerically, it is convenient to define $\tilde{k}_A = -\log\frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log\frac{K_I}{1\,\text{M}}$ and fit for these parameters on a log scale. Dissociation constants are scale invariant, so that a change from 10 μM to 1 μM leads to an equivalent increase in affinity as a change from 1 μM to 0.1 μM. With these definitions we assume for the prior $P(\tilde{k}_A, \tilde{k}_I, \sigma)$ that all three parameters are independent. In addition, we assume a uniform distribution for $\tilde{k}_A$ and $\tilde{k}_I$ and a Jeffreys prior (Sivia and Skilling, 2006) for the scale parameter σ. This yields the complete prior

$$P(\tilde{k}_A, \tilde{k}_I, \sigma) \equiv \frac{1}{(\tilde{k}_A^{\max} - \tilde{k}_A^{\min})}\frac{1}{(\tilde{k}_I^{\max} - \tilde{k}_I^{\min})}\frac{1}{\sigma}. \tag{2.17}$$

These priors are maximally uninformative meaning that they imply no prior knowledge of the parameter values. We defined the $\tilde{k}_A$ and $\tilde{k}_A$ ranges uniform on the range of -7 to 7, although we note that this particular choice does not affect the outcome provided the chosen range is sufficiently wide.

Putting all these terms together we can now sample from $P(\tilde{k}_A, \tilde{k}_I, \sigma \mid D)$ using Markov chain Monte Carlo (see GitHub repository) to compute the most likely parameter as well as the error bars (given by the 95% credible region) for K_A and K_I.

Data Curation

All of the data used in this work as well as all relevant code can be found at this dedicated website. Data were collected, stored, and preserved using the Git version control software in combination with off-site storage and hosting website GitHub. Code used to generate all figures and complete all processing step and analyses are available on the GitHub repository. Many analysis files are stored as instructive Jupyter Notebooks. The scientific community is invited to fork our repositories and open constructive issues on the GitHub repository.

Acknowledgements

The work of this chapter has been a wonderful exercise in scientific collaboration. We thank Hernan Garcia for information and advice for working with these bacterial strains, Pamela Björkman and Rachel Galimidi for access and training for use of the Miltenyi Biotec MACSQuant flow cytometer, and Colin deBakker of Milteny Biotec for useful advice and instruction in flow cytometry. The experimental front of this work began at the Physiology summer course at the Marine Biological Laboratory in Woods Hole, MA operated by the University of Chicago. We thank Simon Alamos, Nalin Ratnayeke, and Shane McInally for their work on the project during the course. We also thank Suzannah Beeler, Justin Bois, Robert Brewster, Ido Golding, Soichi Hirokawa, Jané Kondev, Tom Kuhlman, Heun Jin Lee, Muir Morrison, Nigel Orme, Alvaro Sanchez, and Julie Theriot for useful advice and discussion on this work. This work was supported by La Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA).

2.5 Supplemental Information: Inferring allosteric parameters from previous data

The fold-change profile described by Eq. (2.5) features three unknown parameters K_A, K_I, and $\Delta\varepsilon_{AI}$. In this section, we explore different conceptual approaches to determining these parameters. We first discuss how the induction titration profile of the simple repression constructs used in this paper are not sufficient to determine all three MWC parameters simultaneously, since multiple degenerate sets of parameters can produce the same fold-change response. We then utilize an additional data set from Brewster et al. (2014) to determine the parameter $\Delta\varepsilon_{AI} = 4.5\ k_BT$, after which the remaining parameters K_A and K_I can be extracted from any induction profile with no further degeneracy.

Degenerate Parameter Values

In this section, we discuss how multiple sets of parameters may yield identical fold-change profiles. More precisely, we shall show that if we try to fit the data in Fig. 2.5(C) to the fold-change Eq. (2.5) and extract the three unknown parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), then multiple degenerate parameter sets would yield equally good fits. In other words, this data set alone is insufficient to uniquely determine the actual physical parameter values of the system. This problem persists even when fitting multiple data sets simultaneously as in Supplemental Section 2.11.

In Fig. 2.9(A), we fit the $R = 260$ data by fixing $\Delta\varepsilon_{AI}$ to the value shown on the x-axis and determine the parameters K_A and K_I given this constraint. We use the fold-change function Eq. (2.5) but with $\beta\Delta\varepsilon_{RA}$ modified to the form $\beta\Delta\tilde{\varepsilon}_{RA}$ in Eq. (2.21) to account for the underlying assumptions used when fitting previous data (see below for a full explanation of why this modification is needed).

The best-fit curves for several different values of $\Delta\varepsilon_{AI}$ are shown in Fig. 2.9(B). Note that these fold-change curves are nearly overlapping, demonstrating that different sets of parameters can yield nearly equivalent responses. Without more data, the relationships between the parameter values shown in Fig. 2.9(A) represent the maximum information about the parameter values that can be extracted from the data. Additional experiments which independently measure any of these unknown parameters could resolve this degeneracy. For example, NMR measurements could be used to directly measure the fraction $(1 + e^{-\beta\Delta\varepsilon_{AI}})^{-1}$ of active repressors in the absence of IPTG (Gardino et al., 2003; Boulton and Melacini, 2016).

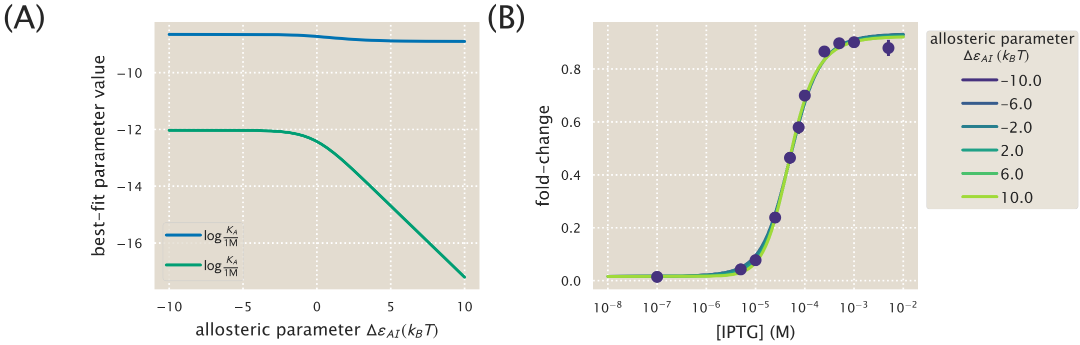


Figure 2.9: **Multiple sets of parameters yield identical fold-change responses.** (A) The data for the O2 strain ($\Delta\varepsilon_{RA} = -13.9\ k_BT$) with $R = 260$ in Fig. 2.5(C) was fit using Eq. (2.5) with $n = 2$. $\Delta\varepsilon_{AI}$ is forced to take on the value shown on the x-axis, while the K_A and K_I parameters are fit freely. (B) The resulting best-fit functions for several value of $\Delta\varepsilon_{AI}$ all yield nearly identical fold-change responses.

Computing $\Delta\varepsilon_{AI}$

As shown in the previous section, the fold-change response of a single strain is not sufficient to determine the three MWC parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), since degenerate sets of parameters yield nearly identical fold-change responses. To circumvent this degeneracy, we now turn to some previous data from the *lac* system in order to determine the value of $\Delta\varepsilon_{AI}$ in Eq. (2.5) for the induction of the Lac repressor. Specifically, we consider two previous sets of work from: (1) Garcia and Phillips (2011) and (2) Brewster et al. (2014), both of which measured fold-change with the same simple repression system in the absence of inducer ($c = 0$) but at various repressor copy numbers R. The original analysis for both data sets assumed that in the absence of inducer all of the Lac repressors were in the active state. As a result, the effective binding energies they extracted were a convolution of the DNA binding energy $\Delta\varepsilon_{RA}$ and the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. We refer to this convoluted energy value as $\Delta\tilde{\varepsilon}_{RA}$. We first disentangle the relationship between these parameters in Garcia and Phillips and then use this relationship to extract the value of $\Delta\varepsilon_{AI}$ from the Brewster et al. dataset.

Garcia and Phillips determined the total repressor copy numbers R of different strains using quantitative Western blots. Then they measured the fold-change at these repressor copy numbers for simple repression constructs carrying the O1, O2, O3, and Oid *lac* operators integrated into the chromosome. These data were then fit to the following thermodynamic model to determine the repressor-DNA binding energies $\Delta\tilde{\varepsilon}_{RA}$ for each operator,

$$\text{fold-change}(c = 0) = \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\tilde{\varepsilon}_{RA}}\right)^{-1}. \tag{2.18}$$

Note that this functional form does not exactly match our fold-change Eq. (2.5) in the limit $c = 0$,

$$\text{fold-change}(c = 0) = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.19}$$

since it is missing the factor $\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}$, which specifies what fraction of repressors are in the active state in the absence of inducer,

$$\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \tag{2.20}$$

In other words, Garcia and Phillips assumed that in the absence of inducer, all repressors were active. In terms of our notation, the convoluted energy values

$\Delta\tilde{\varepsilon}_{RA}$ extracted by Garcia and Phillips (namely, $\Delta\tilde{\varepsilon}_{RA} = -15.3\ k_BT$ for O1 and $\Delta\tilde{\varepsilon}_{RA} = -17.0\ k_BT$ for Oid) represent

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log\left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right). \tag{2.21}$$

Note that if $e^{-\beta\Delta\varepsilon_{AI}} \ll 1$, then nearly all of the repressors are active in the absence of inducer so that $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$. In simple repression systems where we definitively know the value of $\Delta\varepsilon_{RA}$ and R, we can use Eq. (2.19) to determine the value of $\Delta\varepsilon_{AI}$ by comparing with experimentally determined fold-change values. However, the binding energy values that we use from Garcia and Phillips (2011) are effective parameters $\Delta\tilde{\varepsilon}_{RA}$. In this case, we are faced with an undetermined system in which we have more variables than equations, and we are thus unable to determine the value of $\Delta\varepsilon_{AI}$. In order to obtain this parameter, we must turn to a more complex regulatory scenario which provides additional constraints that allow us to fit for $\Delta\varepsilon_{AI}$.

A variation on simple repression in which multiple copies of the promoter are available for repressor binding (for instance, when the simple repression construct is on plasmid) can be used to circumvent the problems that arise when using $\Delta\tilde{\varepsilon}_{RA}$. This is because the behavior of the system is distinctly different when the number of active repressors $p_A(0)R$ is less than or greater than the number of available promoters N. Repression data for plasmids with known copy number N allows us to perform a fit for the value of $\Delta\varepsilon_{AI}$.

To obtain an expression for a system with multiple promoters N, we follow Weinert et al. (2014), writing the fold-change in terms of the the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{2.22}$$

where $\lambda_r = e^{\beta\mu}$ is the fugacity and μ is the chemical potential of the repressor. The fugacity will enable us to easily enumerate the possible states available to the repressor.

To determine the value of λ_r, we first consider that the total number of repressors in the system, R_{tot}, is fixed and given by

$$R_{\text{tot}} = R_S + R_{NS}, \tag{2.23}$$

where R_S represents the number of repressors specifically bound to the promoter and R_{NS} represents the number of repressors nonspecifically bound throughout the genome. The value of R_S is given by

$$R_S = N \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.24}$$

where N is the number of available promoters in the cell. Note that in counting N, we do not distinguish between promoters that are on plasmid or chromosomally integrated provided that they both have the same repressor-operator binding energy (Weinert et al., 2014). The value of R_{NS} is similarly give by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{2.25}$$

where N_{NS} is the number of non-specific sites in the cell (recall that we use $N_{NS} = 4.6 \times 10^6$ for *E. coli*).

Substituting in Eqs. (2.24) and (2.25) into the modified Eq. (2.23) yields the form

$$p_A(0) R_{\text{tot}} = \frac{1}{1 + e^{-\beta \Delta\varepsilon_{AI}}} \left(N \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \tag{2.26}$$

where we recall from Eq. (2.21) that $\beta \Delta\varepsilon_{RA} = \beta \Delta\tilde{\varepsilon}_{RA} + \log \left(\frac{1}{1 + e^{-\beta \Delta\varepsilon_{AI}}} \right)$. Numerically solving for λ_r and plugging the value back into Eq. (2.22) yields a fold-change function in which the only unknown parameter is $\Delta\varepsilon_{AI}$.

With these calculations in hand, we can now determine the value of the $\Delta\varepsilon_{AI}$ parameter. Fig. 2.10(A) shows how different values of $\Delta\varepsilon_{AI}$ lead to significantly different fold-change response curves. Thus, analyzing the specific fold-change response of any strain with a known plasmid copy number N will fix $\Delta\varepsilon_{AI}$. Interestingly, the inflection point of Eq. (2.26) occurs near $p_A(0) R_{\text{tot}} = N$ (as shown by the triangles in Fig. 2.10(A)), so that merely knowing where the fold-change response transitions from concave down to concave up is sufficient to obtain a rough value for $\Delta\varepsilon_{AI}$. We note, however, that for $\Delta\varepsilon_{AI} \gtrsim 5\ k_B T$, increasing $\Delta\varepsilon_{AI}$ further does not affect the fold-change because essentially every repressors will be in the active state in this regime. Thus, if the $\Delta\varepsilon_{AI}$ is in this regime, we can only bound it from below.

We now analyze experimental induction data for different strains with known plasmid copy numbers to determine $\Delta\varepsilon_{AI}$. Fig. 2.10(B) shows experimental measurements of fold-change for two O1 promoters with $N = 64$ and $N = 52$ copy numbers and one Oid promoter with $N = 10$ from Brewster et al. (2014). By fitting these data to Eq. (2.22), we extracted the parameter value $\Delta\varepsilon_{AI} = 4.5\ k_B T$. Substituting this value into Eq. (2.20) shows that 99% of the repressors are in the active state in the absence of inducer and $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$, so that all of the previous energies and calculations made by Garcia and Phillips (2011) and Brewster et al. (2014) were accurate.

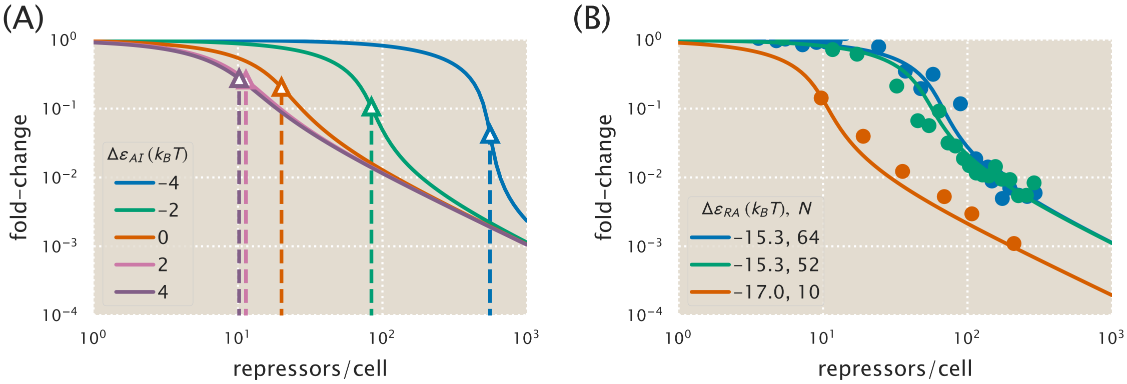


Figure 2.10: **Fold-change of multiple identical genes.** (A) In the presence of $N = 10$ identical promoters, the fold-change Eq. (2.22) depends strongly on the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. The vertical dotted lines represent the number of repressors at which $R_A = N$ for each value of $\Delta\varepsilon_{AI}$. (B) Using fold-change measurements from (Brewster et al., 2014) for the operators and gene copy numbers shown, we can determine the most likely value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ for LacI.

2.6 Supplemental Information: Induction of simple repression with multiple promoters or competitor sites

We made the choice to perform all of our experiments using strains in which a single copy of our simple repression construct had been integrated into the chromosome. This stands in contrast to the methods used by a number of other studies (Oehler et al., 1994; Setty et al., 2003; Oehler et al., 2006; Daber et al., 2009; Daber et al., 2011; Vilar and Saiz, 2013; Shis et al., 2014; Sochor, 2014), in which reporter constructs are placed on plasmid, meaning that the number of constructs in the cell is not precisely known. It is also common to express repressor on plasmid to boost its copy number, which results in an uncertain value for repressor copy number. Here we show that our treatment of the MWC model has broad predictive power beyond the single-promoter scenario we explore experimentally, and indeed can account for systems in which multiple promoters compete for the repressor of interest. Additionally, we demonstrate the importance of having precise control over these parameters, as they can have a significant effect on the induction profile.

Chemical Potential Formulation to Calculate Fold-Change

In this section, we discuss a simple repression construct which we generalize in two ways from the scenario discussed in the text. First, we will allow the repressor to bind to N_S identical specific promoters whose fold-change we are interested in measuring, with each promoter containing a single repressor binding site ($N_S = 1$ in the main text). Second, we consider N_C identical competitor sites which do not regulate the promoter of interest, but whose binding energies are substantially stronger than non-specific binding ($N_C = 0$ in the main text). As in the main text, we assume that the rest of the genome contains N_{NS} non-specific binding sites for the repressor. As in Supplemental Section 2.5, we can write the fold-change Eq. (2.2) in the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.27}$$

where λ_r is the fugacity of the repressor and $\Delta\varepsilon_{RA}$ represents the energy difference between the repressor's binding affinity to the specific operator of interest relative to the repressor's non-specific binding affinity to the rest of the genome.

We now expand our definition of the total number of repressors in the system, R_{tot}, so that it is given by

$$R_{\text{tot}} = R_S + R_{NS} + R_C, \tag{2.28}$$

where R_S, R_{NS}, and R_C represent the number of repressors bound to the specific promoter, a non-specific binding site, or to a competitor binding site, respectively. The value of R_S is given by

$$R_S = N_S \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{2.29}$$

where N_S is the number of specific binding sites in the cell. The value of R_{NS} is similarly give by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{2.30}$$

where N_{NS} is the number of non-specific sites in the cell (recall that we use $N_{NS} = 4.6 \times 10^6$ for *E. coli*), and R_C is given by

$$R_C = N_C \frac{\lambda_r e^{-\beta \Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_C}}, \tag{2.31}$$

where N_C is the number of competitor sites in the cell and $\Delta\varepsilon_C$ is the binding energy of the repressor to the competitor site relative to its non-specific binding energy to the rest of the genome.

To account for the induction of the repressor, we replace the total number of repressors R_{tot} in Eq. (2.28) by the number of active repressors in the cell, $p_A(c) R_{\text{tot}}$. Here, p_A

denotes the probability that the repressor is in the active state (Eq. (2.4)),

$$p_A(c) = \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}. \tag{2.32}$$

Substituting in Eqs. (2.29)-(2.31) into the modified Eq. (2.28) yields the form

$$p_A(c)R_{\text{tot}} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1+\lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS}\frac{\lambda_r}{1+\lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1+\lambda_r e^{-\beta\Delta\varepsilon_C}}. \tag{2.33}$$

For systems where the number of binding sites N_S, N_{NS}, and N_C are known, together with the binding affinities $\Delta\varepsilon_{RA}$ and $\Delta\varepsilon_C$, we can solve numerically for λ_r and then substitute it into Eq. (2.27) to obtain a fold-change at any concentration of inducer c. In the following sections, we will theoretically explore the induction curves given by Eq. (2.33) for a number of different combinations of simple repression binding sites, thereby predicting how the system would behave if additional specific or competitor binding sites were introduced.

Variable Repressor Copy Number (*R*) with Multiple Specific Binding Sites ($N_S > 1$)

In the the main text, we consider the induction profiles of strains with varying R but a single, specific binding site $N_S = 1$ (see Fig. 2.6). Here we predict the induction profiles for similar strains in which R is varied, but $N_S > 1$, as shown in Fig. 2.11. The top row shows induction profiles in which $N_S = 10$ and the bottom row shows profiles in which $N_S = 100$, assuming three different choices for the specific operator binding sites given by the O1, O2, and O3 operators. These values of N_S were chosen to mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid. A few features stand out in these profiles. First, as the magnitude of N_S surpasses the number of repressors R, the leakiness begins to increase significantly, since there are no longer enough repressors to regulate all copies of the promoter of interest. Second, in the cases where $\Delta\varepsilon_{RA} = -15.3\ k_BT$ for the O1 operator or $\Delta\varepsilon_{RA} = -13.9\ k_BT$ for the O2 operator, the profiles where $N_S = 100$ are notably sharper than the profiles where $N_S = 10$, and it is possible to achieve dynamic ranges approaching 1. Finally, it is interesting to note that the profiles for the O3 operator where $\Delta\varepsilon_{RA} = -9.7\ k_BT$ are nearly indifferent to the value of N_S.

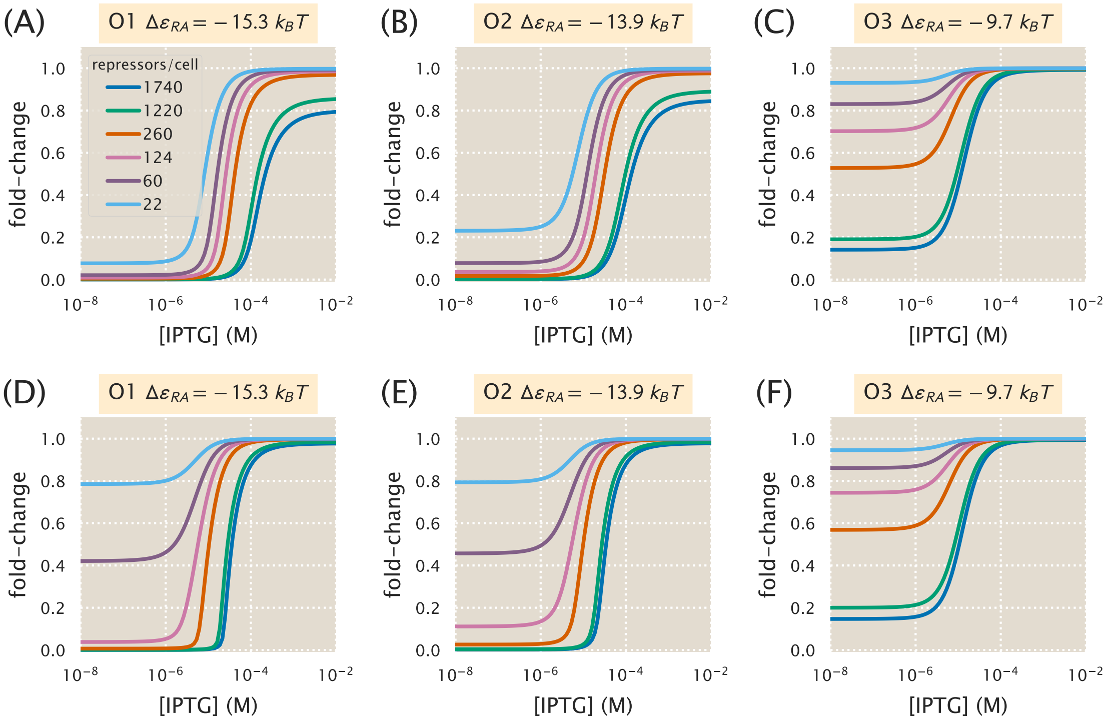


Figure 2.11: **Induction with variable *R* and multiple specific binding sites.** Induction profiles are shown for strains with variable R and $\Delta\varepsilon_{RA} = -15.3, -13.9,$ or $-9.7\ k_BT$. (A-C) The number of specific sites, N_S, is held constant at 10 as R and $\Delta\varepsilon_{RA}$ are varied. (D-F) N_S is held constant at 100 as R and $\Delta\varepsilon_{RA}$ are varied. These situations mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid.

Variable Number of Specific Binding Sites N_S with Fixed Repressor Copy Number (R)

The second set of scenarios we consider is the case in which the repressor copy number $R = 260$ is held constant while the number of specific promoters N_S is varied (see Fig. 2.12). Again we see that leakiness is increased significantly when $N_S > R$, though all profiles for $\Delta\varepsilon_{RA} = -9.7\ k_BT$ exhibit high leakiness, making the effect less dramatic for this operator. Additionally, we find again that adjusting the number of specific sites can produce induction profiles with maximal dynamic ranges. In particular, the O1 and O2 profiles with $\Delta\varepsilon_{RA} = -15.3$ and $-13.9\ k_BT$, respectively, have dynamic ranges approaching 1 for $N_S = 50$ and 100.

Competitor Binding Sites

An intriguing scenario is presented by the possibility of competitor sites elsewhere in the genome. This serves as a model for situations in which a promoter of interest is

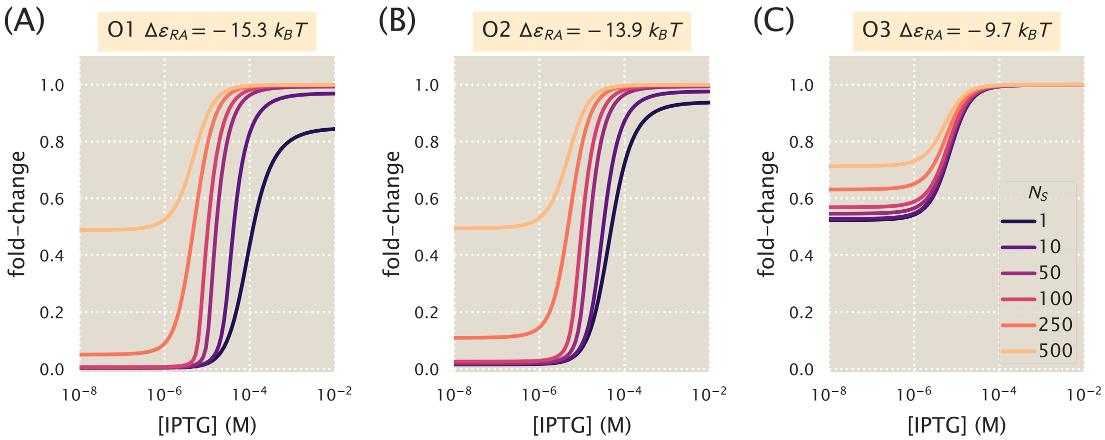


Figure 2.12: **Induction with variable specific sites and fixed *R*.** Induction profiles are shown for strains with $R = 260$ and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$. The number of specific sites N_S is varied from 1 to 500.

regulated by a transcription factor that has multiple targets. This is highly relevant, as the majority of transcription factors in *E. coli* have at least two known binding sites, with approximately 50 transcription factors having more than ten known binding sites (Rydenfelt et al., 2014; Schmidt et al., 2015). If the number of competitor sites and their average binding energy is known, however, they can be accounted for in the model. Here, we predict the induction profiles for strains in which $R = 260$ and $N_S = 1$, but there is a variable number of competitor sites N_C with a strong binding energy $\Delta\varepsilon_C = -17.0\ k_BT$. In the presence of such a strong competitor, when $N_C > R$ the leakiness is greatly increased, as many repressors are siphoned into the pool of competitor sites. This is most dramatic for the case where $\Delta\varepsilon_{RA} = -9.7\ k_BT$, in which it appears that no repression occurs at all when $N_C = 500$. Interestingly, when $N_C < R$ the effects of the competitor are not especially notable.

Properties of the Induction Response

As discussed in the main body of the paper, our treatment of the MWC model allows us to predict key properties of induction responses. Here, we consider the leakiness, saturation, and dynamic range (see Fig. 2.1) by numerically solving Eq. (2.33) in the absence of inducer, $c = 0$, and in the presence of saturating inducer $c \to \infty$. Using Eq. (2.32), the former case is given by

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}, \qquad (2.34)$$

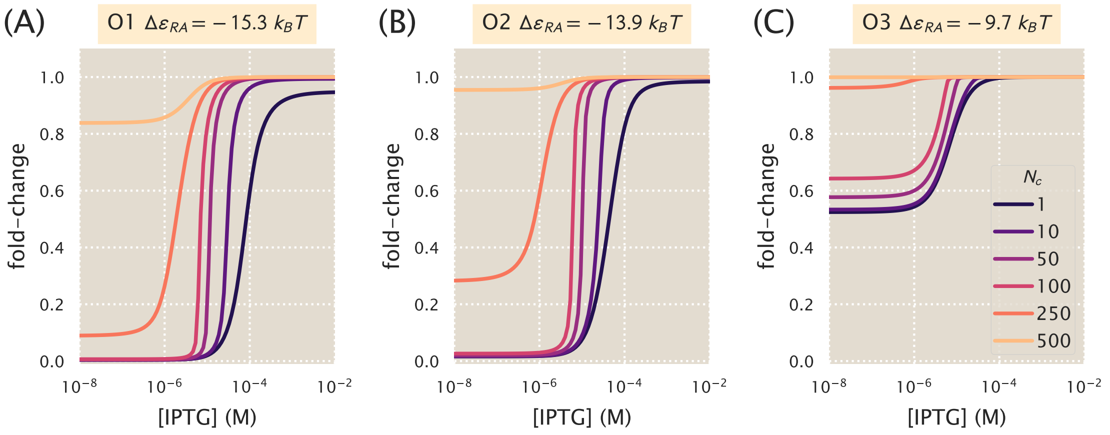


Figure 2.13: **Induction with variable competitor sites, a single specific site, and fixed *R*.** Induction profiles are shown for strains with $R = 260$, $N_s = 1$, and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$ for the O1 operator, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$ for the O2 operator, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$ for the O3 operator. The number of specific sites, N_C, is varied from 1 to 500. This mimics the common scenario in which a transcription factor has multiple binding sites in the genome.

whereupon substituting in the value of λ_r into Eq. (2.27) will yield the leakiness. Similarly, the limit of saturating inducer is found by determining λ_r from the form

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^2} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}. \quad (2.35)$$

In Fig. 2.14 we show how the leakiness, saturation, and dynamic range vary with R and $\Delta\varepsilon_{RA}$ in systems with $N_S = 10$ or $N_S = 100$. An inflection point occurs where $N_S = R$, with leakiness and dynamic range behaving differently when $R < N_S$ than when $R > N_S$. This transition is more dramatic for $N_S = 100$ than for $N_S = 10$. Interestingly, the saturation values consistently approach 1, indicating that full induction is easier to achieve when multiple specific sites are present. Moreover, dynamic range values for O1 and O2 strains with $\Delta\varepsilon_{RA} = -15.3$ and $-13.9\ k_BT$ approach 1 when $R > N_S$, although when $N_S = 10$ there is a slight downward dip owing to saturation values of less than 1 at high repressor copy numbers.

In Fig. 2.15 we similarly show how the leakiness, saturation, and dynamic range vary with R and $\Delta\varepsilon_{RA}$ in systems with $N_S = 1$ and multiple competitor sites $N_C = 10$ or $N_C = 100$. Each of the competitor sites has a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The phenotypic profiles are very similar to those for multiple specific sites shown in Fig. 2.14, with sharper transitions at $R = N_C$ due to the greater binding strength

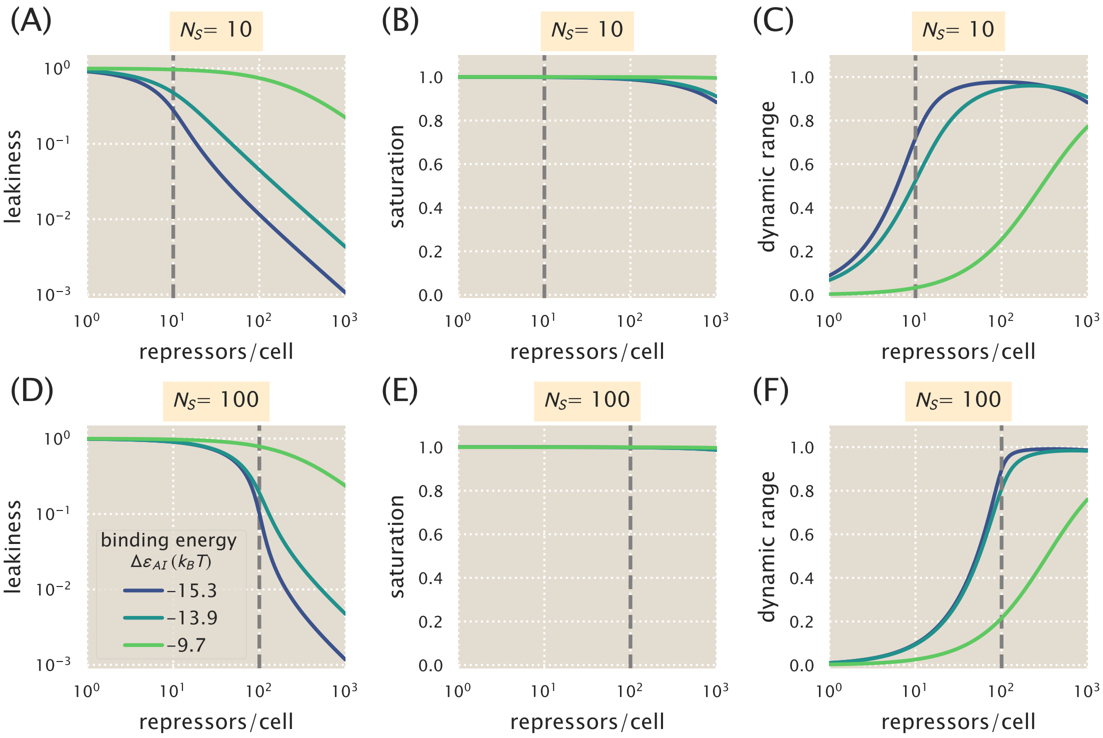


Figure 2.14: **Phenotypic properties of induction with multiple specific binding sites.** The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with number of specific binding sites $N_S = 10$ (A-C) or $N_S = 100$ (D-F). The dashed vertical line indicates the point at which $N_S = R$.

of the competitor site. This indicates that introducing competitors has much the same effect on the induction phenotypes as introducing additional specific sites, as in either case the influence of the repressors is dampened when there are insufficient repressors to interact with all of the specific binding sites.

This section of the appendix gives a quantitative analysis of the nuances imposed on induction response in the case of systems involving multiple gene copies as are found in the vast majority of studies on induction. In these cases, the intrinsic parameters of the MWC model get entangled with the parameters describing gene copy number.

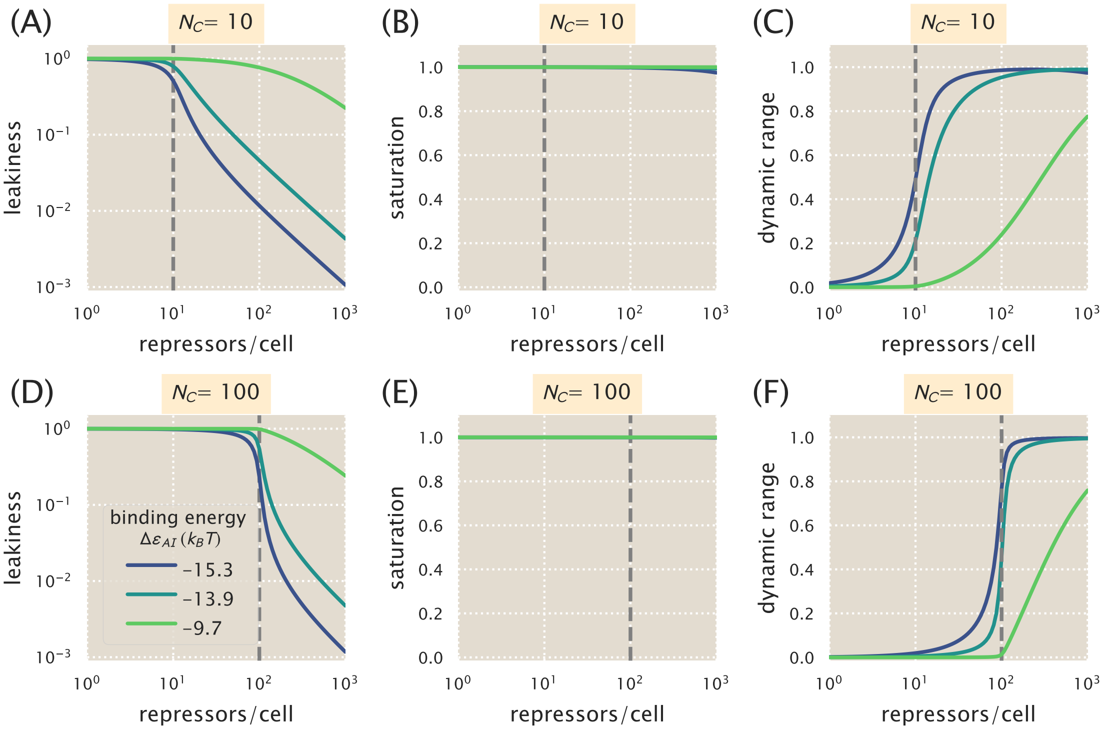


Figure 2.15: **Phenotypic properties of induction with a single specific site and multiple competitor sites.** The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with a single specific binding site $N_S = 1$ and a number of competitor sites $N_C = 10$ (A-C) or $N_C = 100$ (D-F). All competitor sites have a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The dashed vertical line indicates the point at which $N_C = R$.

2.7 Supplemental Information: Flow cytometry

In this section, we provide information regarding the equipment used to make experimental measurements of the fold-change in gene expression in the interests of transparency and reproducibility. We also provide a summary of our unsupervised method of gating the flow cytometry measurements for consistency between experimental runs.

Equipment

Due to past experience using the Miltenyi Biotec MACSQuant flow cytometer during the Physiology summer course at the Marine Biological Laboratory, we used the same flow cytometer for the formal measurements in this work. All measurements were made using an excitation wavelength of 488 nm with an emission filter set of 525/50 nm. This excitation wavelength provides approximately 40% of the maximum YFP absorbance (Chroma Technology Corporation, 2016), and this was found to be sufficient for the purposes of these experiments. A useful feature of modern flow cytometry is the high-sensitivity signal detection through the use of photomultiplier tubes (PMT) whose response can be tuned by adjusting the voltage. Thus, the voltage for the forward-scatter (FSC), side-scatter (SSC), and gene expression measurements were tuned manually to maximize the dynamic range between autofluorescence signal and maximal expression without losing the details of the population distribution. Once these voltages were determined, they were used for all subsequent measurements. Extremely low signal producing particles were discarded before data storage by setting a basal voltage threshold, thus removing the majority of spurious events. The various instrument settings for data collection are given in Table 2.1.

Laser	Channel	Sensor Voltage
488 nm	Forward-Scatter (FSC)	423 V
488 nm	Side-Scatter (SSC)	537 V
488 nm	Intensity (B1 Filter, 525/50nm)	790 V
488 nm	Trigger (debris threshold)	24.5 V

Table 2.1: **Instrument settings for data collection using the Miltenyi Biotec MACSQuant flow cytometer.** All experimental measurements were collected using these values.

Experimental Measurement

A single data set consisted of seven bacterial strains, all sharing the same operator, with varying repressor copy numbers (R = 0, 22, 60, 124, 260, 1220, and 1740), in addition to an autofluorescent strain, under twelve IPTG concentrations. Data collection took place over two to three hours. During this time, the cultures were held at approximately 4°C by placing the 96-well plate on a MACSQuant ice block. Because the ice block thawed over the course of the experiment, the samples measured last were approximately at room temperature. This means that samples may have grown slightly by the end of the experiment. To confirm that this continued growth did not alter the measured results, a subset of experiments were run in reverse meaning that the fully induced cultures were measured first and the uninduced samples last. The plate arrangements and corresponding fold-change measurements are shown in Fig. 2.16(A) and Fig. 2.16(B), respectively. The measured fold-change values in the reverse ordered plate appear to be drawn from the same distribution as those measured in the forward order, meaning that any growth that might have taken place during the experiment did not significantly affect the results. Both the forward and reverse data sets were used in our analysis.

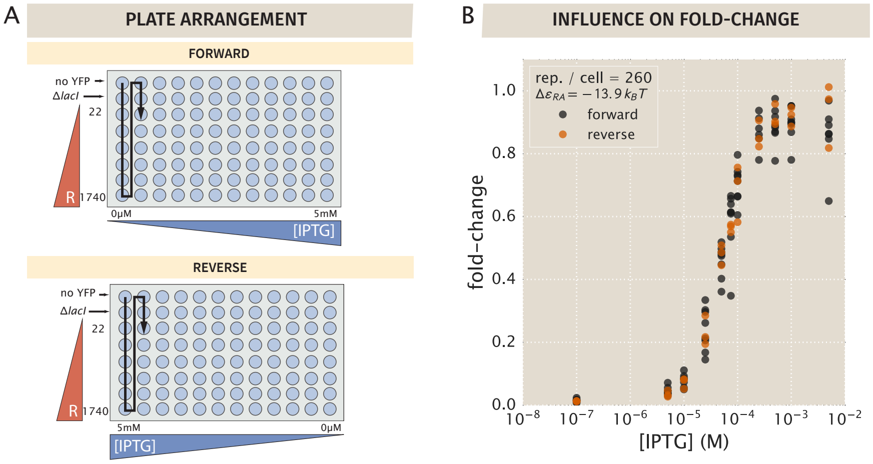


Figure 2.16: **Plate arrangements for flow cytometry.** (A) Samples were measured primarily in the forward arrangement with a subset of samples measured in reverse. The black arrow indicates the order in which samples were processed by the flow cytometer. (B) The experimentally measured fold-change values for the two sets of plate arrangements show that samples measured in the forward arrangement appear to be indistinguishable from those measured in reverse order.

Unsupervised Gating

As explained in the Methods, we used an automatic unsupervised gating procedure to filter the flow cytometry data based on the front and side-scattering values returned by the MACSQuant flow cytometer. We assume that the region with highest density of points in these two channels corresponds to single-cell measurements. Everything extending outside of this region was discarded in order to exclude sources of error such as cell clustering, particulates, or other spurious events.

In order to define the gated region we fit a two-dimensional Gaussian function to the $\log_{10}$ forward-scattering (FSC) and the $\log_{10}$ side-scattering (SSC) data. We then kept a fraction $\alpha \in [0, 1]$ of the data by defining an elliptical region given by

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_\alpha(p), \tag{2.36}$$

where $\boldsymbol{x}$ is the 2×1 vector containing the log(FSC) and log(SSC), $\boldsymbol{\mu}$ is the 2×1 vector representing the mean values of log(FSC) and log(SSC) as obtained from fitting a two-dimensional Gaussian to the data, and $\boldsymbol{\Sigma}$ is the 2×2 covariance matrix also obtained from the Gaussian fit. $\chi^2_\alpha(p)$ is the quantile function for probability p of the chi-squared distribution with two degrees of freedom. Fig. 2.17 shows an example of different gating contours that would arise from different values of α in Eq. (2.36). In this work, we chose $\alpha = 0.4$ which we deemed was a sufficient constraint to minimize the noise in the data. As explained in Supplemental Section 2.8 we compared our high throughput flow cytometry data with single cell microscopy, confirming that the automatic gating did not introduce systematic biases to the analysis pipeline. The specific code where this gating is implemented can be found in GitHub repository.

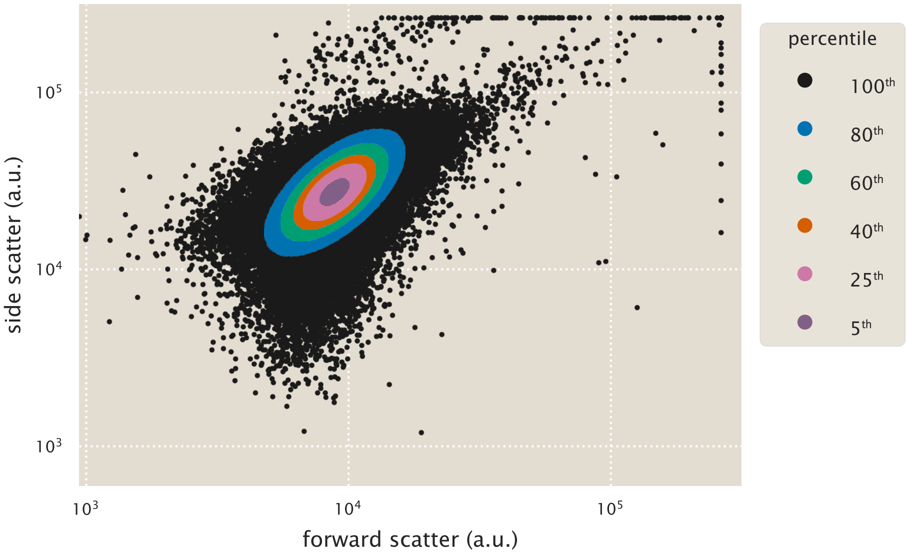


Figure 2.17: **Representative unsupervised gating contours.** Points indicate individual flow cytometry measurements of forward scatter and side scatter. Colored points indicate arbitrary gating contours ranging from 100% ($\alpha = 1.0$) to 5% ($\alpha = 0.05$). All measurements for this work were made computing the mean fluorescence from the 40$^{\text{th}}$ percentile ($\alpha = 0.4$), shown as orange points.

Comparison of Flow Cytometry with Other Methods

Previous work from our lab experimentally determined fold-change for similar simple repression constructs using a variety of different measurement methods (Garcia et al., 2011; Brewster et al., 2014). Garcia and Phillips used the same background strains as the ones used in this work, but gene expression was measured with Miller assays based on colorimetric enzymatic reactions with the LacZ protein (Garcia and Phillips, 2011). Brewster et al. (2014) used a LacI dimer with the tetramerization region replaced with an mCherry tag, where the fold-change was measured as the ratio of the gene expression rate rather than a single snapshot of the gene output.

Fig. 2.18 shows the comparison of these methods along with the flow cytometry method used in this work. The consistency of these three readouts validates the quantitative use of flow cytometry and unsupervised gating to determine the fold-change in gene expression. However, one important caveat revealed by this figure is that the sensitivity of flow cytometer measurements is not sufficient to accurately determine the fold-change for the high repressor copy number strains in O1 without induction. Instead, a method with a large dynamic range such as the Miller assay is needed to accurately resolve the fold-change at such low expression levels.

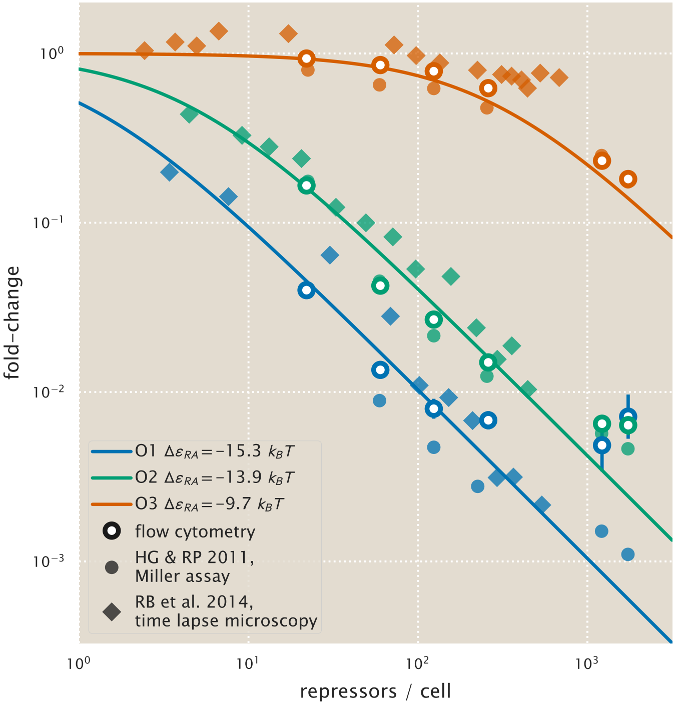


Figure 2.18: **Comparison of experimental methods to determine the fold-change.** The fold-change in gene expression for equivalent simple-repression constructs has been determined using three independent methods: flow cytometry (this work), colorimetric Miller assays (Garcia and Phillips, 2011), and video microscopy (Brewster et al., 2014). All three methods give consistent results, although flow cytometry measurements lose accuracy for fold-change less than 10^{-2}. Note that the repressor-DNA binding energies $\Delta\varepsilon_{RA}$ used for the theoretical predictions were determined in Garcia and Phillips (2011).

2.8 Supplemental Information: Single-cell microscopy

In this section, we detail the procedures and results from single-cell microscopy verification of our flow cytometry measurements. Our previous measurements of fold-change in gene expression have been measured using bulk-scale Miller assays (Garcia and Phillips, 2011) or through single-cell microscopy (Brewster et al., 2014). In this work, flow cytometry was an attractive method due to the ability to screen through many different strains at different concentrations of inducer in a short amount of time. To verify our results from flow cytometry, we examined two bacterial strains with different repressor-DNA binding energies ($\Delta\varepsilon_{RA}$) of $-13.9\ k_BT$ and $-15.3\ k_BT$ with $R = 260$ repressors per cell using fluorescence microscopy and estimated the values of the parameters K_A and K_I for direct comparison between the two methods. For a detailed explanation of the Python code implementation of the processing steps described below, please see this paper's GitHub repository. An outline of our microscopy workflow can be seen in Fig. 2.19.

Strains and Growth Conditions

Cells were grown in an identical manner to those used for measurement via flow cytometry (see Methods). Briefly, cells were grown overnight (between 10 and 13 hours) to saturation in rich media broth (LB) with 100 μg $\cdot$ mL^{-1} spectinomycin in a deep-well 96 well plate at 37°C. These cultures were then diluted 1000-fold into 500 μL of M9 minimal medium supplemented with 0.5% glucose and the appropriate concentration of the inducer IPTG. Strains were allowed to grow at 37°C with vigorous aeration for approximately 8 hours. Prior to mounting for microscopy, the cultures were diluted 10-fold into M9 glucose minimal medium in the absence of IPTG. Each construct was measured using the same range of inducer concentration values as was performed in the flow cytometry measurements (between 100 nM and 5 mM IPTG). Each condition was measured in triplicate in microscopy whereas approximately ten measurements were made using flow cytometry.

Imaging Procedure

During the last hour of cell growth, an agarose mounting substrate was prepared containing the appropriate concentration of the IPTG inducer. This mounting substrate was composed of M9 minimal medium supplemented with 0.5% glucose and 2% agarose (Life Technologies UltraPure Agarose, Cat. No. 16500100). This solution was heated in a microwave until molten followed by addition of the IPTG to the appropriate final concentration. This solution was then thoroughly mixed and a 500 μL aliquot was sandwiched between two glass coverslips and was allowed to solidify.

Once solid, the agarose substrates were cut into approximately 10 mm × 10 mm squares. An aliquot of one to two microliters of the diluted cell suspension was then added to each pad. For each concentration of inducer, a sample of the autofluorescence control, the $\Delta lacI$ constitutive expression control, and the experimental strain was prepared, yielding a total of thirty-six agarose mounts per experiment. These samples

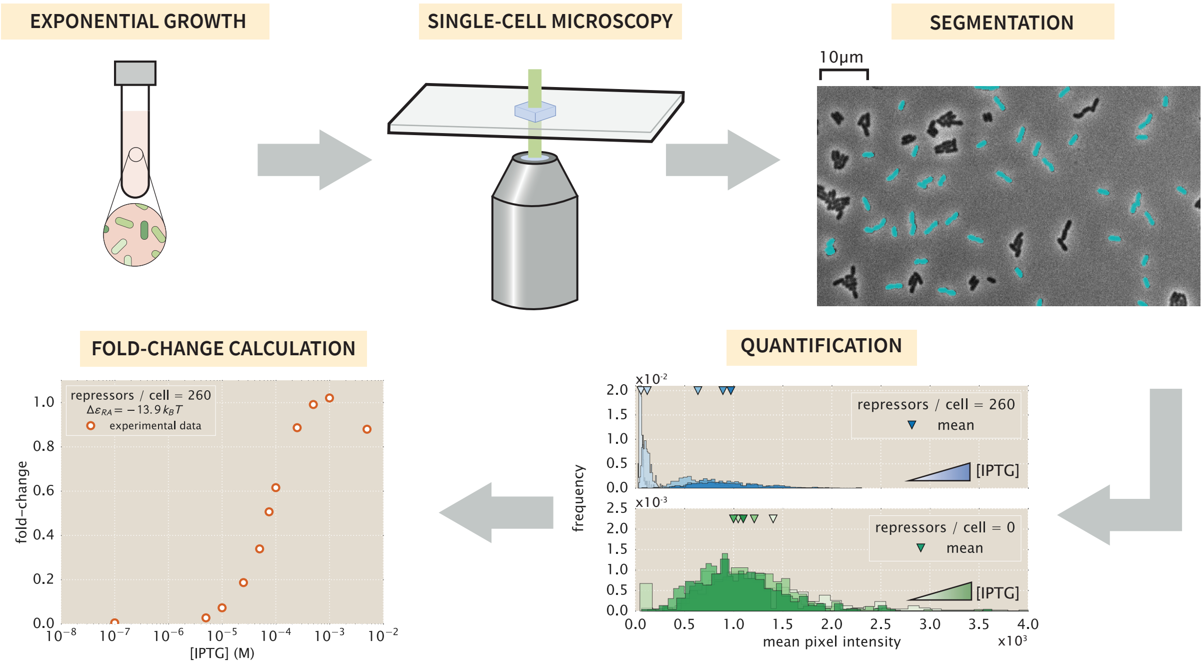


Figure 2.19: **Experimental workflow for single-cell microscopy.** For comparison with the flow cytometry results, the cells were grown in an identical manner to those described in the main text. Once cells had reached mid to late exponential growth, the cultures were diluted and placed on agarose substrates and imaged under 100× magnification. Regions of interest representing cellular mass were segmented and average single-cell intensities were computed. The means of the distributions were used to compute the fold-change in gene expression.

were then mounted onto two glass-bottom dishes (Ted Pella Wilco Dish, Cat. No. 14027-20) and sealed with parafilm.

All imaging was performed on a Nikon Ti-Eclipse inverted fluorescent microscope outfitted with a custom-built laser illumination system and operated by the open-source MicroManager control software (Edelstein et al., 2014). The YFP fluorescence was imaged using a CrystaLaser 514 nm excitation laser coupled with a laser-optimized (Semrock Cat. No. LF514-C-000) emission filter.

For each sample, between fifteen and twenty positions were imaged allowing for measurement of several hundred cells. At each position, a phase contrast image, an mCherry image, and a YFP image were collected in that order with exposures on a time scale of ten to twenty milliseconds. For each channel, the same exposure time was used across all samples in a given experiment. All images were collected and stored in `ome.tiff` format. All microscopy images are available on the CaltechDATA online repository under DOI: 10.22002/D1.229.

Image Processing: Correcting Uneven Illumination

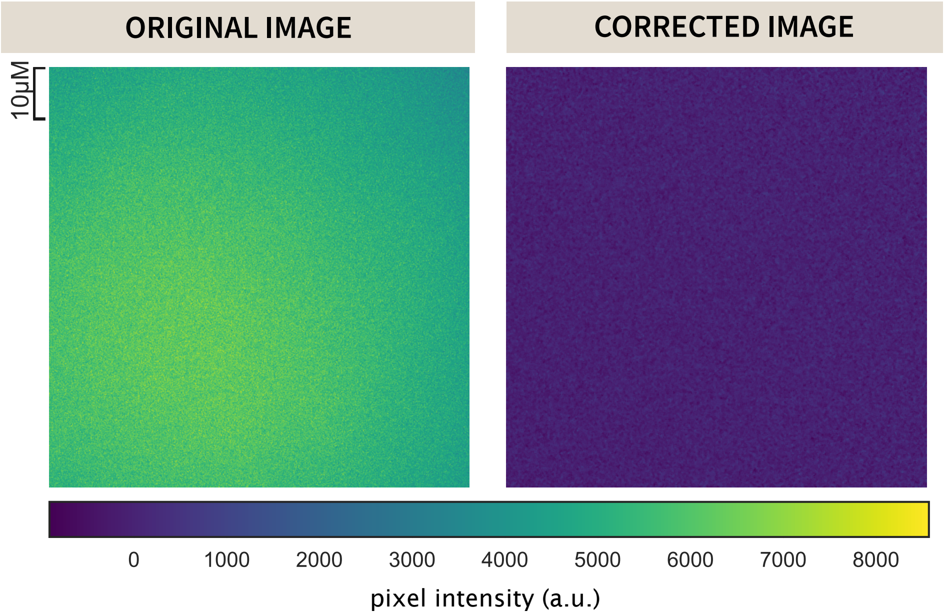


Figure 2.20: **Correction for uneven illumination.** A representative image of the illumination profile of the 512 nm excitation beam on a homogeneously fluorescent slide is shown in the left panel. This is corrected for using equation Eq. (2.37) and is shown in the right panel.

The excitation laser has a two-dimensional gaussian profile. To minimize nonuniform illumination of a single field of view, the excitation beam was expanded to illuminate an area larger than that of the camera sensor. While this allowed for an entire field of view to be illuminated, there was still approximately a 10% difference in illumination across both dimensions. This nonuniformity was corrected for in post-processing by capturing twenty images of a homogeneously fluorescent plastic slide (Autofluorescent Plastic Slides, Chroma Cat. No. 920001) and averaging to generate a map of illumination intensity at any pixel I_YFP. To correct for shot noise in the camera (Andor iXon+ 897 EMCCD), twenty images were captured in the absence of illumination using the exposure time used for the experimental data. Averaging over these images produced a map of background noise at any pixel I_dark. To perform the correction, each fluorescent image in the experimental acquisition was renormalized with respect to these average maps as

$$I_\text{flat} = \frac{I - I_\text{dark}}{I_\text{YFP} - I_\text{dark}} \langle I_\text{YFP} - I_\text{dark} \rangle, \tag{2.37}$$

where I_flat is the renormalized image and I is the original fluorescence image. An example of this correction can be seen in Fig. 2.20.

Image Processing: Cell Segmentation

Each bacterial strain constitutively expressed an mCherry fluorophore from a low copy-number plasmid. This served as a volume marker of cell mass allowing us to segment individual cells through edge detection in fluorescence. We used the Marr-Hildreth edge detector (Marr and Hildreth, 1980) which identifies edges by taking the second derivative of a lightly Gaussian blurred image. Edges are identified as those regions which cross from highly negative to highly positive values or vice-versa within a specified neighborhood. Bacterial cells were defined as regions within an intact and closed identified edge. All segmented objects were then labeled and passed through a series of filtering steps.

To ensure that primarily single cells were segmented, we imposed area and eccentricity bounds. We assumed that single cells projected into two dimensions are roughly 2 μm long and 1 μm wide, so that cells are likely to have an area between 0.5 μm^2 and 6 μm. To determine the eccentricity bounds, we assumed that the a single cell can be approximated by an ellipse with semi-major (a) and semi-minor (b) axis lengths of 0.5 μm and 0.25 μm, respectively. The eccentricity of this hypothetical cell can be computed as

$$\text{eccentricity} = \sqrt{1 - \left(\frac{b}{a}\right)^2}, \tag{2.38}$$

yielding a value of approximately 0.8. Any objects with an eccentricity below this value were not considered to be single cells. After imposing both an area (Fig. 2.21(A)) and eccentricity filter (Fig. 2.21(B)), the remaining objects were considered cells of interest (Fig. 2.21(C)) and the mean fluorescence intensity of each cell was extracted.

Image Processing: Calculation of Fold-Change

Cells exhibited background fluorescence even in the absence of an expressed fluorophore. We corrected for this autofluorescence contribution to the fold-change calculation by subtracting the mean YFP fluorescence of cells expressing only the mCherry volume marker from each experimental measurement. The fold-change in gene expression was therefore calculated as

$$\text{fold-change} = \frac{\langle I_{R>0}\rangle - \langle I_\text{auto}\rangle}{\langle I_{R=0}\rangle - \langle I_\text{auto}\rangle}, \tag{2.39}$$

where $\langle I_{R>0}\rangle$ is the mean fluorescence intensity of cells expressing LacI repressors, $\langle I_\text{auto}\rangle$ is the mean intensity of cells expressing only the mCherry volume marker,

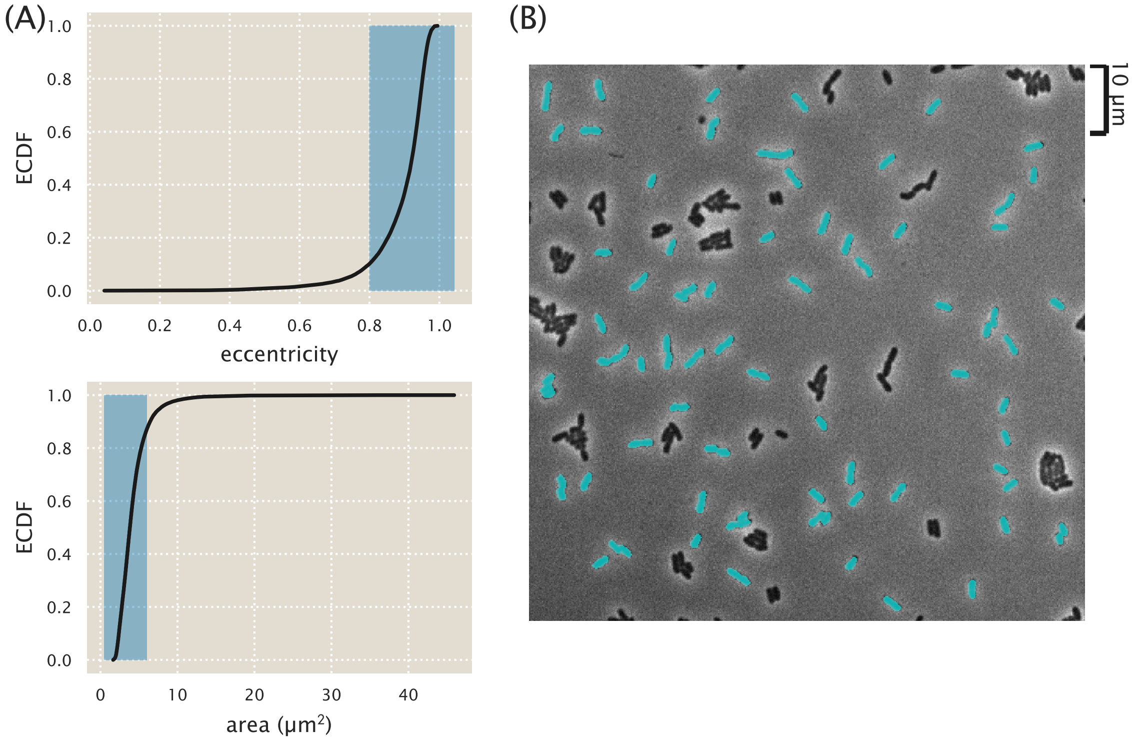


Figure 2.21: **Segmentation of single bacterial cells.** (A) Objects were selected if they had an eccentricity greater than 0.8 and an area between 0.5 μm^2 and 6 μm^2. Highlighted in blue are the regions considered to be representative of single cells. The black lines correspond to the empirical cumulative distribution functions for the parameter of interest. (B) A representative final segmentation mask is shown in which segmented cells are depicted in cyan over the phase contrast image.

and $\langle I_{R=0} \rangle$ is the mean fluorescence intensity of cells in the absence of LacI. These fold-change values were very similar to those obtained through flow cytometry and were well described using the thermodynamic parameters used in the main text. With these experimentally measured fold-change values, the best-fit parameter values of the model were inferred and compared to those obtained from flow cytometry.

Parameter Estimation and Comparison

To confirm quantitative consistency between flow cytometry and microscopy, the parameter values of K_A and K_I were also estimated from three biological replicates of IPTG titration curves obtained by microscopy for strains with R = 260 and operators O1 and O2. Fig. 2.22(A) shows the data from these measurements (orange circles) and the ten biological replicates from our flow cytometry measurements (blue circles), along with the fold-change predictions from each inference. In

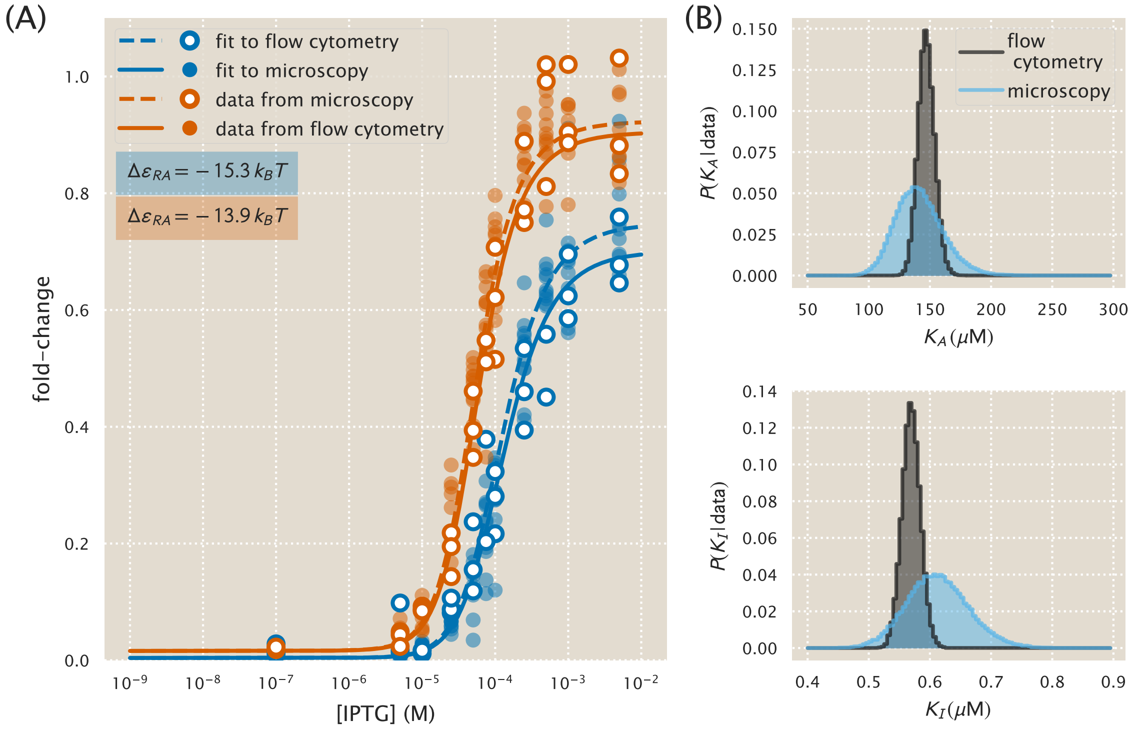


Figure 2.22: **Comparison of measured fold-change between flow cytometry and single-cell microscopy.** (A) Experimentally measured fold-change values obtained through single-cell microscopy and flow cytometry are shown as white filled and solid colored circles, respectively. Solid and dashed lines indicate the predicted behavior using the most likely parameter values of K_A and K_I inferred from flow cytometry data and microscopy data, respectively. The red and blue plotting elements correspond to the different operators O1 and O2 with binding energies $\Delta\varepsilon_{RA}$ of $-13.9\ k_BT$ and $-15.3\ k_BT$, respectively (Garcia and Phillips, 2011). (B) The marginalized posterior distributions for K_A and K_I are shown in the top and bottom panel, respectively. The posterior distribution determined using the microscopy data is wider than that computed using the flow cytometry data due to a smaller fig collection of data sets (three for microscopy and ten for flow cytometry).

comparison with the values obtained by flow cytometry, each parameter estimate overlapped with the 95% credible region of our flow cytometry estimates, as shown in Fig. 2.22(B). Specifically, these values were $K_A = 142^{+40}_{-34}\,\mu$M and $K_I = 0.6^{+0.1}_{-0.1}\,\mu$M from microscopy and $K_A = 149^{+14}_{-12}\,\mu$M and $K_I = 0.57^{+0.03}_{-0.02}\,\mu$M from the flow cytometry data. We note that the credible regions from the microscopy data shown in Fig. 2.22(B) are much broader than those from flow cytometry due to the fewer number of replicates performed.

2.9 Supplemental Information: Fold-change sensitivity analysis

In Fig. 2.6 we found that the width of the credible regions varied widely depending on the repressor copy number R and repressor operator binding energy $\Delta\varepsilon_{RA}$. More precisely, the credible regions were much narrower for low repressor copy numbers R and weak binding energy $\Delta\varepsilon_{RA}$. In this section, we explain how this behavior comes about. We focus our attention on the maximum fold-change in the presence of saturating inducer given by Eq. (2.7). While it is straightforward to consider the width of the credible regions at any other inducer concentration, Fig. 2.6 shows that the credible region are widest at saturation.

The width of the credible regions corresponds to how sensitive the fold-change is to the fit values of the dissociation constants K_A and K_I. To be quantitative, we define

$$\Delta\,\text{fold-change}_{K_A} \equiv \text{fold-change}(K_A, K_I^{\text{fit}}) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}), \qquad (2.40)$$

the difference between the fold-change at a particular K_A value relative to the best-fit dissociation constant $K_A^{\text{fit}} = 139 \times 10^{-6}\,\text{M}$. For simplicity, we keep the inactive state dissociation constant fixed at its best-fit value $K_I^{\text{fit}} = 0.53 \times 10^{-6}\,\text{M}$. A larger difference $\Delta\,\text{fold-change}_{K_A}$ implies a wider credible region. Similarly, we define the analogous quantity

$$\Delta\,\text{fold-change}_{K_I} = \text{fold-change}(K_A^{\text{fit}}, K_I) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}) \qquad (2.41)$$

to measure the sensitivity of the fold-change to K_I at a fixed K_A^{fit}. Fig. 2.23 shows both of these quantities in the limit $c \to \infty$ for different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ and repressor copy numbers R. See our GitHub repository for the code that reproduces these plots.

To understand how the width of the credible region scales with $\Delta\varepsilon_{RA}$ and R, we can Taylor expand the difference in fold-change to first order, $\Delta\,\text{fold-change}_{K_A} \approx \frac{\partial\,\text{fold-change}}{\partial K_A}\left(K_A - K_A^{\text{fit}}\right)$, where the partial derivative has the form

$$\frac{\partial\,\text{fold-change}}{\partial K_A} = \frac{e^{-\beta\Delta\varepsilon_{AI}} \frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n-1}}{\left(1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \qquad (2.42)$$

Similarly, the Taylor expansion $\Delta\,\text{fold-change}_{K_I} \approx \frac{\partial\,\text{fold-change}}{\partial K_I}\left(K_I - K_I^{\text{fit}}\right)$ features the partial derivative

$$\frac{\partial\,\text{fold-change}}{\partial K_I} = -\frac{e^{-\beta\Delta\varepsilon_{AI}} \frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n}}{\left(1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \qquad (2.43)$$

From Eqs. (2.42) and (2.43), we find that both $\Delta\,\text{fold-change}_{K_A}$ and $\Delta\,\text{fold-change}_{K_I}$ increase in magnitude with R and decrease in magnitude with $\Delta\varepsilon_{RA}$. Accordingly, we expect that the O3 strains (with the least negative $\Delta\varepsilon_{RA}$) and the strains with the smallest repressor copy number will lead to partial derivatives with smaller magnitude and hence to tighter credible regions. Indeed, this prediction is carried out in Fig. 2.23.

Lastly, we note that Eqs. (2.42) and (2.43) enable us to quantify the scaling relationship between the width of the credible region and the two quantities R and $\Delta\varepsilon_{RA}$. For example, for the O3 strains, where the fold-change at saturating inducer concentration is ≈ 1, the right-most term in both equations which equals the fold-change squared is roughly 1. Therefore, we find that both $\frac{\partial\,\text{fold-change}}{\partial K_A}$ and $\frac{\partial\,\text{fold-change}}{\partial K_I}$ scale linearly with R and $e^{-\beta\Delta\varepsilon_{RA}}$. Thus the width of the $R = 22$ strain will be roughly 1/1000 as large as that of the $R = 1740$ strain; similarly, the width of the O3 curves will be roughly 1/1000 the width of the O1 curves.

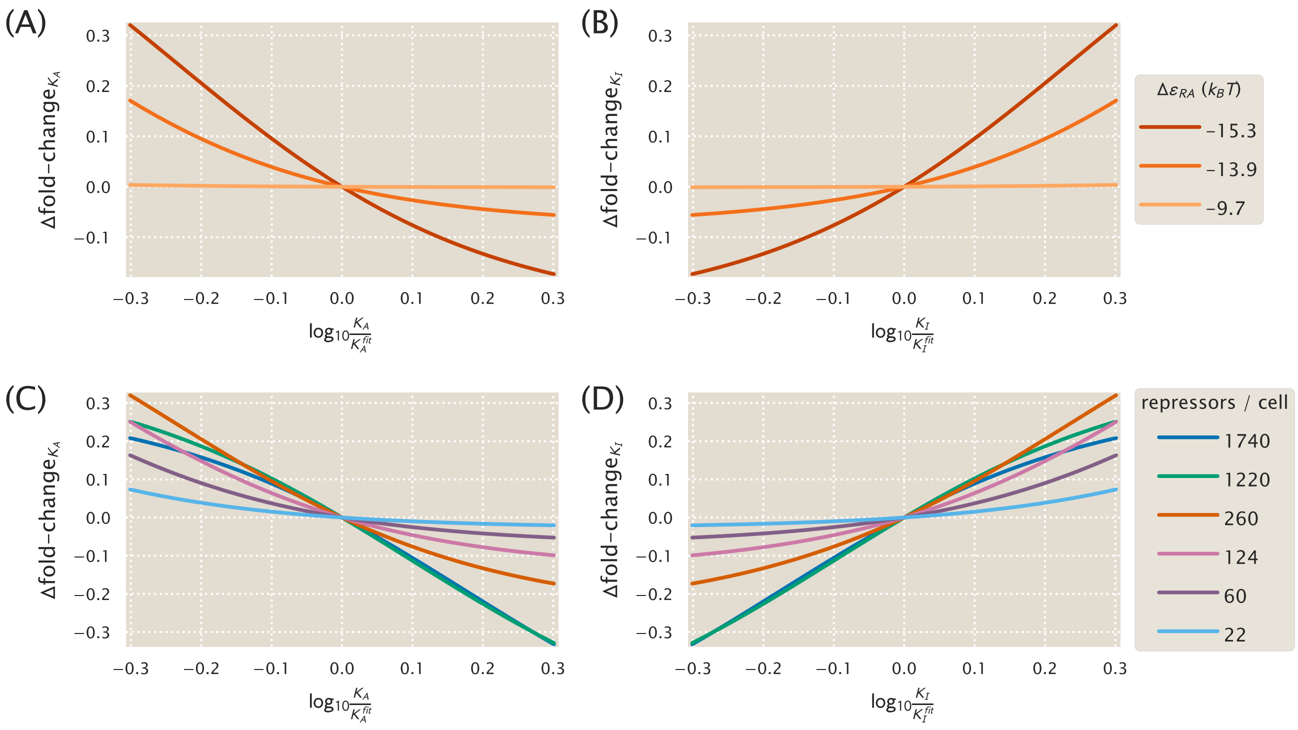


Figure 2.23: **Determining how sensitive the fold-change values are to the fit values of the dissociation constants.**(A) The difference $\Delta\,\text{fold-change}_{K_A}$ in fold change when the dissociation constant K_A is slightly offset from its best-fit value $K_A = 139^{+29}_{-22} \times 10^{-6}$ M, as given by Eq. (2.40). Fold-change is computed in the limit of saturating inducer concentration ($c \longrightarrow \infty$, see Eq. (2.7)) where the credible regions in Fig. 2.6 are widest. The O3 strain ($\Delta\varepsilon_{RA} = -9.7\ k_BT$) is about 1/1000 as sensitive as the O1 operator to perturbations in the parameter values, and hence its credible region is roughly 1/1000 as wide. All curves were made using $R = 260$. (B) As in Panel (A), but plotting the sensitivity of fold-change to the K_I parameter relative to the best-fit value $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M. Note that only the magnitude, and not the sign, of this difference describes the sensitivity of each parameter. Therefore, the O3 strain is again less sensitive than the O1 and O2 strains. (C) As in Panel (A), but showing how the fold-change sensitivity for different repressor copy numbers. The strains with lower repressor copy number are less sensitive to changes in the dissociation constants, and hence their corresponding curves in Fig. 2.6 have tighter credible regions. All curves were made using $\Delta\varepsilon_{RA} = -13.9\ k_BT$. (D) As in Panel (C), the sensitivity of fold-change with respect to K_I is again smallest (in magnitude) for the low repressor copy number strains.

2.10 Supplemental Information: Alternate characterizations of induction

In this section we discuss a different way to describe the induction data, namely, through using the conventional Hill approach. We first demonstrate how using a Hill function to characterize a single induction curve enables us to extract features (such as the midpoint and sharpness) of that single response, but precludes any predictions of the other seventeen strains. We then discuss how a thermodynamic model of simple repression coupled with a Hill approach to the induction response can both characterize an induction profile and predict the response of all eighteen strains, although we argue that such a description provides no insight into the allosteric nature of the protein and how mutations to the repressor would affect induction. We conclude the section by discussing the differences between such a model and the statistical mechanical model used in the main text.

Fitting Induction Curves using a Hill Function Approach

The Hill equation is a phenomenological function commonly used to describe data with a sigmoidal profile (Murphy et al., 2007; Murphy et al., 2010; Rogers et al., 2015). Its simplicity and ability to estimate the cooperativity of a system (through the Hill coefficient) has led to its widespread use in many domains of biology (Frank, 2013). Nevertheless, the Hill function is often criticized as a physically unrealistic model and the extracted Hill coefficient is often difficult to contextualize in the physics of a system (Weiss, 1997). In the present work, we note that a Hill function, even if it is only used because of its simplicity, presents no mechanism to understand how a regulatory system's behavior will change if physical parameters such as repressor copy number or operator binding energy are varied. In addition, the Hill equation provides no foundation to explore how mutating the repressor (e.g., at its inducer-binding interface) would modify its induction profile, although statistical mechanical models have proved capable of characterizing such scenarios (Keymer et al., 2006; Swem et al., 2008; Einav et al., 2016).

Consider the general Hill equation for a single induction profile given by

$$\text{fold-change} = (\text{leakiness}) + (\text{dynamic range})\frac{\left(\frac{c}{K}\right)^n}{1 + \left(\frac{c}{K}\right)^n}, \tag{2.44}$$

where, as in the main text, the leakiness represents the minimum fold-change, the dynamic range represents the difference between the maximum and minimum fold-change, K is the repressor-inducer dissociation constant, and n denotes the Hill coefficient that characterizes the sharpness of the curve ($n > 1$ signifies positive cooperativity, $n = 1$ denotes no cooperativity, and $n < 1$ represents negative cooperativity). Fig. 2.24 shows how the individual induction profiles can be fit (using the same Bayesian methods as described in Supplemental Section 2.11 and the Methods) to this Hill response, yielding a similar response to that shown in Fig. 2.5(D). However, characterizing the induction response in this manner is unsatisfactory because each curve must be fit independently thus removing our predictive power for other repressor copy numbers and binding sites.

The fitted parameters obtained from this approach are shown in Fig. 2.25. These are rather unsatisfactory because they do not clearly reflect the properties of the physical system under consideration. For example, the dissociation constant K between LacI and inducer should not be affected by either the copy number of the repressor or the DNA binding energy, and yet we see upward trends as R is increased or the binding energy is decreased. Here, the K parameter ultimately describes the midpoint of the induction curve and therefore cannot strictly be considered a dissociation constant. Similarly, the Hill coefficient n does not directly represent the cooperativity between the repressor and the inducer as the molecular details of the copy number and DNA binding strength are subsumed in this parameter as well. While the leakiness and dynamic range describe important phenotypic properties of the induction response, this Hill approach leaves us with no means to predict them for other strains. In summary, the Hill equation Eq. (2.44) cannot predict how an induction profile varies with repressor copy number, operator binding energy, or how mutations will alter the induction profile. To that end, we turn to a more sophisticated approach where we use the Hill function to describe the available fraction of repressor as a function of inducer concentration.

Fitting Induction Curves using a Combination Thermodynamic Model and Hill Function Approach

Motivated by the inability in the previous section to characterize all eighteen strains using the Hill function with a single set of parameters, here we combine the Hill approach with a thermodynamic model of simple repression to garner predictive power. More specifically, we will use the thermodynamic model in Fig. 2.2(A) but substitute the statistical model in Fig. 2.2(B) with the phenomenological Hill function Eq. (2.44).

Following Eqs. (2.1)-(2.3), fold-change is given by

$$\text{fold-change} = \left(1 + p_A(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.45}$$

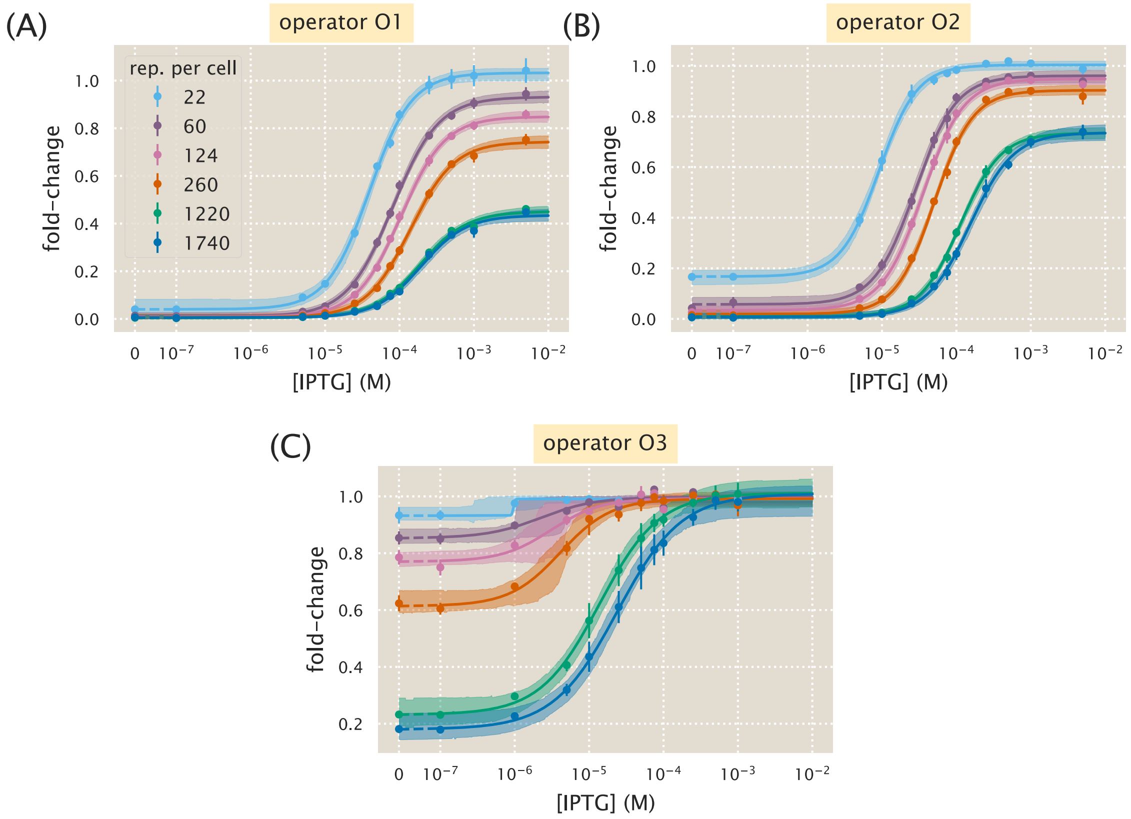


Figure 2.24: **Hill function and MWC analysis of each induction profile.** Data for each individual strain was fit to the general Hill function in Eq. (2.44). (A) strains with O1 binding site, (B) strains with O2 binding site, and (C) strains with O3 binding site. Shaded regions indicate the bounds of the 95% credible region.

where the Hill function

$$p_A(c) = p_A^{\text{max}} - p_A^{\text{range}} \frac{\left(\frac{c}{K_D}\right)^n}{1 + \left(\frac{c}{K_D}\right)^n} \tag{2.46}$$

represents the fraction of repressors in the allosterically active state, with p_A^{max} denoting the fraction of active repressors in the absence of inducer and $p_A^{\text{max}} - p_A^{\text{range}}$ the minimum fraction of active repressors in the presence of saturating inducer. The Hill function characterizes the inducer-repressor binding while the thermodynamic model with the known constants R, N_{NS}, and $\Delta\varepsilon_{RA}$ describes how the induction profile changes with repressor copy number and repressor-operator binding energy.

As in the main text, we can fit the four Hill parameters – the vertical shift and stretch parameters p_A^{max} and p_A^{range}, the Hill coefficient n, and the inducer-repressor dissociation constant K_D – for a single induction curve and then use the fully characterized Eq. (2.45) to describe the response of each of the eighteen strains.

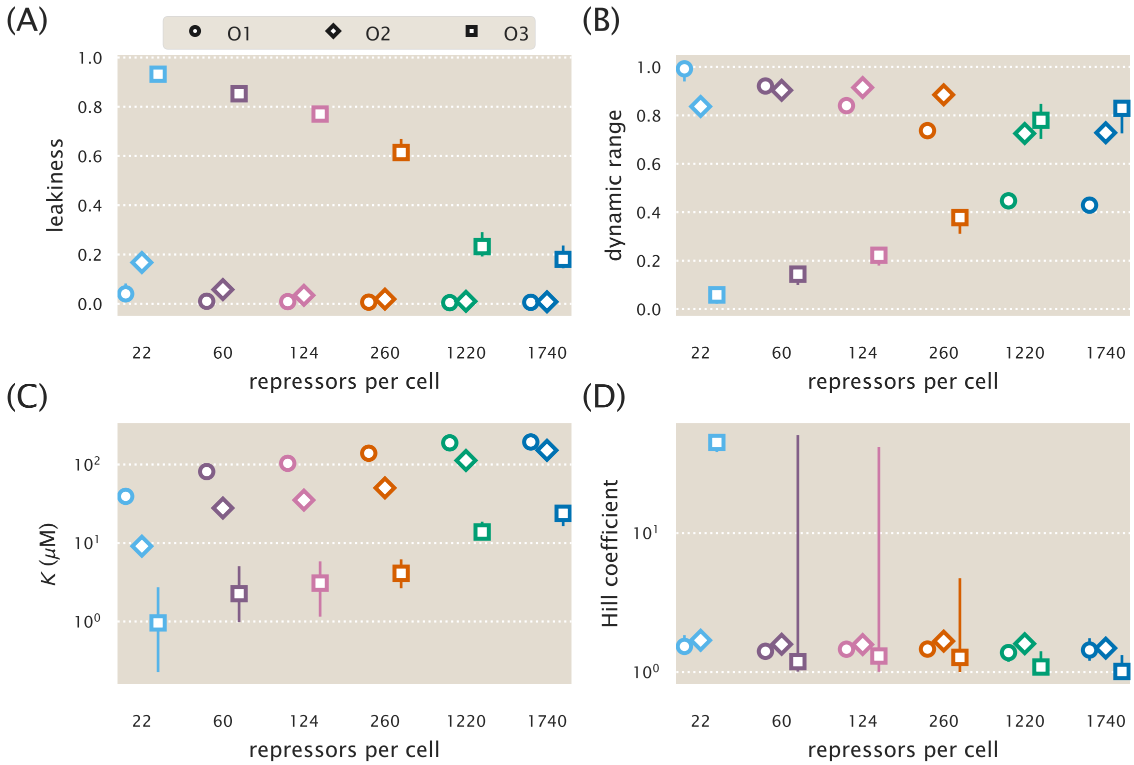


Figure 2.25: **Parameter values for the Hill equation fit to each individual titration.** The resulting fit parameters from the Hill function fits of Fig. 2.24 are summarized. The large parameter intervals for many of the O3 strains are due to the flatter induction profile (as seen by its smaller dynamic range), and the ability for a large range of K and n values to describe the data.

Fig. 2.26 shows this process carried out by fitting the O2 $R = 260$ strain (white circles in Panel (B)) and predicting the behavior of the remaining seventeen strains.

Although the curves in Fig. 2.26 are nearly identical to those in Fig. 2.5 (which were made using the MWC model Eq. (2.5)), we stress that the Hill function approach is more complex than the MWC model (containing four parameters instead of three) and it obscures the relationships to the physical parameters of the system. For example, it is not clear whether the fit parameter $K_D = 4^{+2}_{-1} \times 10^{-6}$ M relays the dissociation constant between the inducer and active-state repressor, between the inducer and the inactive-state repressor, or some mix of the two quantities.

In addition, the MWC model Eq. (2.5) naturally suggests further quantitative tests for the fold-change relationship. For example, mutating the repressor's inducer binding site would likely alter the repressor-inducer dissociation constants K_A and K_I, and it would be interesting to find out if such mutations also modify the allosteric energy

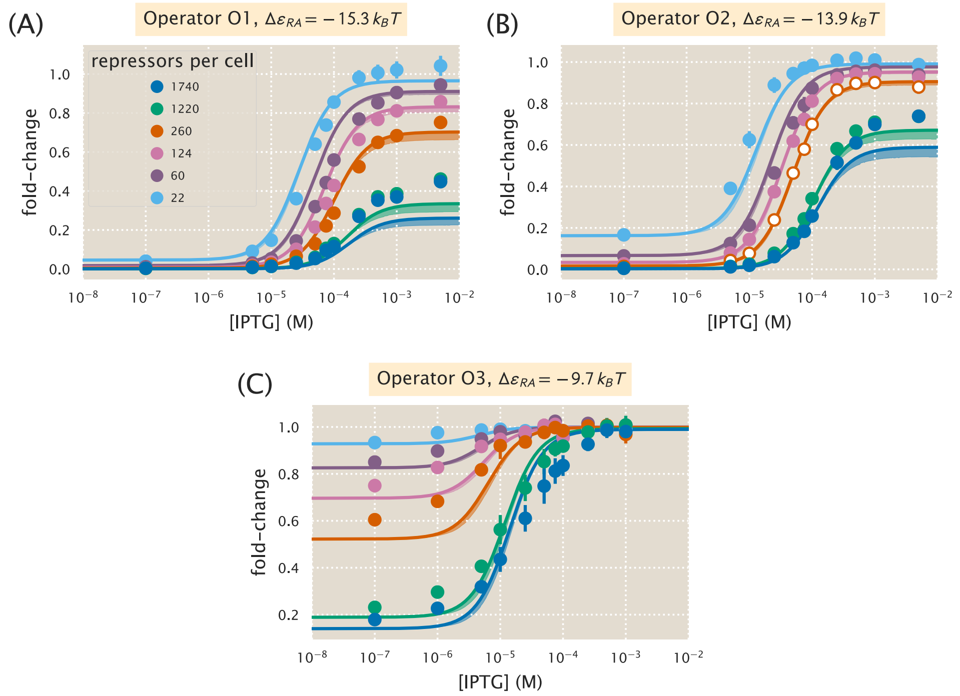


Figure 2.26: **A thermodynamic model coupled with a Hill analysis can characterize induction.** Combining a thermodynamic model of simple repression with the Hill function to characterize the repressor-inducer binding successfully characterizes the induction profiles of all eighteen strains. As in the main text, data was only fit for the O2 $R = 260$ strain using Eqs. (2.45) and (2.46) and the parameters $p_A^{\text{max}} = 0.90^{+0.03}_{-0.01}$, $p_A^{\text{range}} = -0.90^{+0.02}_{-0.03}$, $n = 1.6^{+0.2}_{-0.1}$, and $K_D = 4^{+2}_{-1} \times 10^{-6}$ M. Shaded regions indicate bounds of the 95% credible region.

difference $\Delta\varepsilon_{AI}$ between the repressor's active and inactive conformations. For our purposes, the Hill function Eq. (2.46) falls short of the connection to the physics of the system and provides no intuition about how transcription depends upon such mutations. For these reasons, we present the thermodynamic model coupled with the statistical mechanical MWC model approach in the paper.

2.11 Supplemental Information: Global fit of all parameters

In the main text, we used the repressor copy numbers R and repressor-DNA binding energies $\Delta\varepsilon_{RA}$ as reported by Garcia and Phillips (2011). However, any error in these previous measurements of R and $\Delta\varepsilon_{RA}$ will necessarily propagate into our own fold-change predictions. In this section we take an alternative approach to fitting the physical parameters of the system to that used in the main text. First, rather than fitting only a single strain, we fit the entire data set in Fig. 2.6 along with microscopy data for the synthetic operator Oid (see Supplemental Section 2.12). In addition, we also simultaneously fit the parameters R and $\Delta\varepsilon_{RA}$ using the prior information given by the previous measurements. By using the entire data set and fitting all of the parameters, we obtain the best possible characterization of the statistical mechanical parameters of the system given our current state of knowledge. As a point of reference, we state all of the parameters of the MWC model derived in the text in Table 2.2.

To fit all of the parameters simultaneously, we follow a similar approach to the one detailed in the Methods section. Briefly, we perform a Bayesian parameter estimation of the dissociation constants K_A and K_I, the six different repressor copy numbers R corresponding to the six *lacI* ribosomal binding sites used in our work, and the four different binding energies $\Delta\varepsilon_{RA}$ characterizing the four distinct operators used to make the experimental strains. As in the main text, we fit the logarithms $\tilde{k}_A = -\log \frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\text{M}}$ of the dissociation constants which grants better numerical stability.

As in Eqs. (2.15) and (2.16), we assume that deviations of the experimental fold-change from the theoretical predictions are normally distributed with mean zero and standard deviation σ. We begin by writing Bayes' theorem,

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D) = \frac{P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)}{P(D)}, \tag{2.47}$$

where $\boldsymbol{R}$ is an array containing the six different repressor copy numbers to be fit, $\boldsymbol{\Delta\varepsilon_{RA}}$ is an array containing the four binding energies to be fit, and D is the experimental fold-change data. The term $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D)$ gives the probability distributions of all of the parameters given the data. The term $P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)$ represents the likelihood of having observed our experimental data given some value for each parameter. $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)$ contains all the prior information on the values of these parameters. Lastly, $P(D)$ serves as a normalization constant and hence can be ignored.

Given n independent measurements of the fold-change, the first term in Eq. (2.47) can be written as

$$P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\text{fc}^{(i)}_{\text{exp}} - \text{fc}(\tilde{k}_A, \tilde{k}_I, R^{(i)}, \Delta\varepsilon^{(i)}_{RA}, c^{(i)}))^2}{2\sigma^2}\right], \tag{2.48}$$

where $\mathrm{fc}_{\mathrm{exp}}^{(i)}$ is the i^{th} experimental fold-change and $\mathrm{fc}(\cdots)$ is the theoretical prediction. Note that the standard deviation σ of this distribution is not known and hence needs to be included as a parameter to be fit.

The second term in Eq. (2.47) represents the prior information of the parameter values. We assume that all parameters are independent of each other, so that

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = P(\tilde{k}_A) \cdot P(\tilde{k}_I) \cdot \prod_i P(R^{(i)}) \cdot \prod_j P(\Delta\varepsilon_{RA}^{(j)}) \cdot P(\sigma), \quad (2.49)$$

where the superscript (i) indicates the repressor copy number of index i and the superscript (j) denotes the binding energy of index j. As above, we note that a prior must also be included for the unknown parameter σ.

Because we knew nothing about the values of $\tilde{k}_A$, $\tilde{k}_I$, and σ before performing the experiment, we assign maximally uninformative priors to each of these parameters. More specifically, we assign uniform priors to $\tilde{k}_A$ and $\tilde{k}_I$ and a Jeffreys prior to σ, indicating that K_A, K_I, and σ are scale parameters (Sivia and Skilling, 2006). We do, however, have prior information for the repressor copy numbers and the repressor-DNA binding energies from Garcia and Phillips (2011). This prior knowledge is included within our model using an informative prior for these two parameters, which we assume to be Gaussian. As such, each of the $R^{(i)}$ repressor copy numbers to be fit satisfies

$$P(R^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_{R_i}^2}} \exp\left(-\frac{(R^{(i)} - \bar{R}^{(i)})^2}{2\sigma_{R_i}^2}\right), \quad (2.50)$$

where $\bar{R}^{(i)}$ is the mean repressor copy number and σ_{R_i} is the variability associated with this parameter as reported in Garcia and Phillips (2011). Note that we use the given value of σ_{R_i} from previous measurements rather than leaving this as a free parameter.

Similarly, the binding energies $\Delta\varepsilon_{RA}^{(j)}$ are also assumed to have a Gaussian informative prior of the same form. We write it as

$$P(\Delta\varepsilon_{RA}^{(j)}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_j}^2}} \exp\left(-\frac{(\Delta\varepsilon_{RA}^{(j)} - \Delta\bar{\varepsilon}_{RA}^{(j)})^2}{2\sigma_{\varepsilon_j}^2}\right), \quad (2.51)$$

where $\Delta\bar{\varepsilon}_{RA}^{(j)}$ is the binding energy and σ_{ε_j} is the variability associated with that parameter around the mean value as reported in Garcia and Phillips (2011) .

The σ_{R_i} and σ_{ε_j} parameters will constrain the range of values for $R^{(i)}$ and $\Delta\varepsilon_{RA}^{(j)}$ found from the fitting. For example, if for some i the standard deviation σ_{R_i} is very small, it implies a strong confidence in the previously reported value. Mathematically, the exponential in Eq. (2.50) will ensure that the best-fit $R^{(i)}$ lies within a few standard deviations of $\bar{R}^{(i)}$. Since we are interested in exploring which values could give the best fit, the errors are taken to be wide enough to allow the parameter estimation to freely explore parameter space in the vicinity of the best estimates. Putting all these terms together, we use Markov chain Monte Carlo to sample the posterior distribution $P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D)$, enabling us to determine both the most likely value for each physical parameter as well as its associated credible region (see the GitHub repository for the implementation).

Fig. 2.27 shows the result of this global fit. When compared with Fig. 2.6 we can see that fitting for the binding energies and the repressor copy numbers improves the agreement between the theory and the data. Table 2.3 summarizes the values of the parameters as obtained with this MCMC parameter inference. We note that even though we allowed the repressor copy numbers and repressor-DNA binding energies to vary, the resulting fit values were very close to the previously reported values. The fit values of the repressor copy numbers were all within one standard deviation of the previous reported values provided in Garcia and Phillips (2011). And although some of the repressor-DNA binding energies differed by a few standard deviations from the reported values, the differences were always less than 1 k_BT, which represents a small change in the biological scales we are considering. The biggest discrepancy between our fit values and the previous measurements arose for the synthetic Oid operator, which we discuss in more detail in Supplemental Section 2.12.

Fig. 2.28 shows the same key properties as in Fig. 2.7, but uses the parameters obtained from this global fitting approach. We note that even by increasing the number of degrees of freedom in our fit, the result does not change substantially, due to in general, only minor improvements between the theoretical curves and data. For the O3 operator data, again, agreement between the predicted $[EC_{50}]$ and the effective Hill coefficient remain poor due the theory being unable to capture the steepness of the response curves.

Parameter	**Description**
c	Concentration of the inducer
K_A, K_I	Dissociation constant between an inducer and the repressor in the active/inactive state
$\Delta\varepsilon_{AI}$	The difference between the free energy of repressor in the inactive and active states
$\Delta\varepsilon_P$	Binding energy between the RNAP and its specific binding site
$\Delta\varepsilon_{RA}, \Delta\varepsilon_{RI}$	Binding energy between the operator and the active/inactive repressor
n	Number of inducer binding sites per repressor
P	Number of RNAP
R_A, R_I, R	Number of active/inactive/total repressors
$p_A = \frac{R_A}{R}$	Probability that a repressor will be in the active state
p_{bound}	Probability that an RNAP is bound to the promoter of interest, assumed to be proportional to gene expression
fold-change	Ratio of gene expression in the presence of repressor to that in the absence of repressor
F	Free energy of the system
N_{NS}	The number of non-specific binding sites for the repressor in the genome
$\beta = \frac{1}{k_B T}$	The inverse product of the Boltzmann constant k_B and the temperature T of the system

Table 2.2: **Key model parameters for induction of an allosteric repressor.**

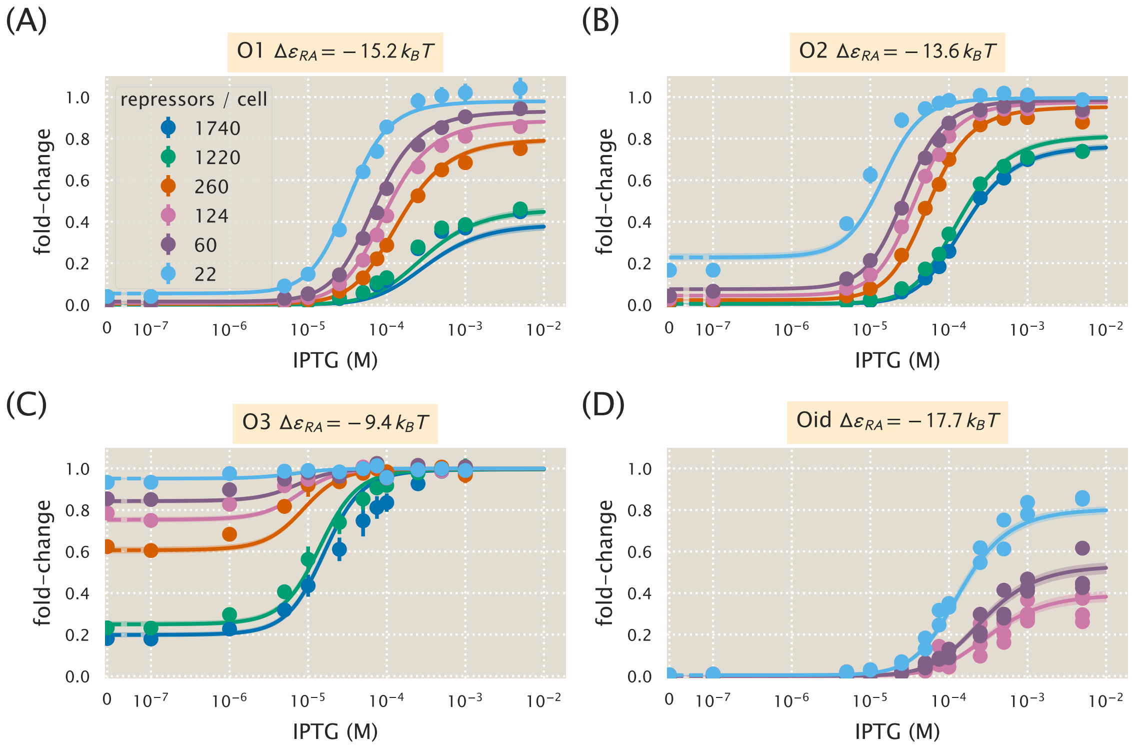


Figure 2.27: **Global fit of dissociation constants, repressor copy numbers and binding energies.** Theoretical predictions resulting from simultaneously fitting the dissociation constants K_A and K_I, the six repressor copy numbers R, and the four repressor-DNA binding energies $\Delta\varepsilon_{RA}$ using the entire data set from Fig. 2.6 as well as the microscopy data for the Oid operator. Error bars of experimental data show the standard error of the mean (eight or more replicates) and shaded regions denote the 95% credible region. Where error bars are not visible, they are smaller than the point itself. For the Oid operator, all of the data points are shown since a smaller number of replicates were taken. The shaded regions are significantly smaller than in Fig. 2.6 because this fit was based on all data points, and hence the fit parameters are much more tightly constrained. The dashed lines at 0 IPTG indicates a linear scale, whereas solid lines represent a log scale.

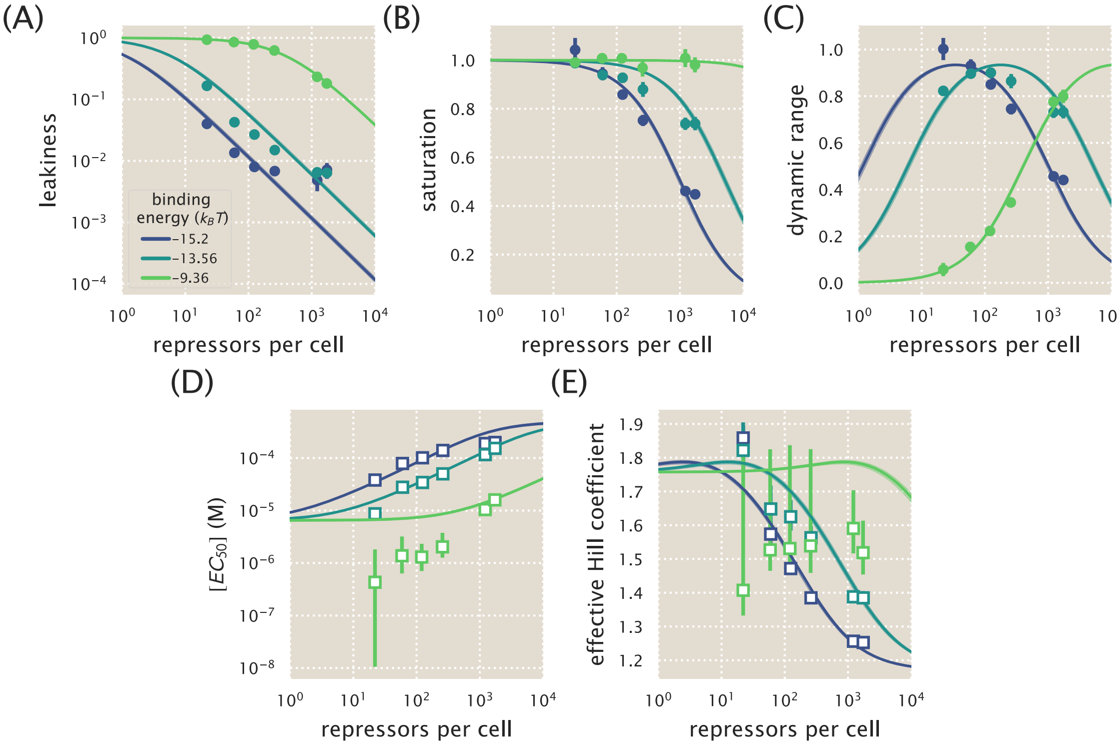


Figure 2.28: **Key properties of induction profiles as predicted with a global fit using all available data.** Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 2.6 in the absence and presence of IPTG. All prediction curves were generated using the parameters listed in 2.3. Both the (D) $[EC_{50}]$ and (E) effective Hill coefficient are inferred by individually fitting all parameters – K_A, K_I, R, $\Delta\varepsilon_{RA}$ – to each operator-repressor pairing in Fig. 2.6(A)-(C) separately to Eq. (2.5) in order to smoothly interpolate between the data points. Note that where error bars are not visible, this indicates that the error bars are smaller than the point itself.

	Reported Values (Garcia and Phillips, 2011)	**Global Fit**
$\tilde{k}_A$	–	$-5.33^{+0.06}_{-0.05}$
$\tilde{k}_I$	–	$0.31^{+0.05}_{-0.06}$
K_A	–	205^{+11}_{-12} μM
K_I	–	$0.73^{+0.04}_{-0.04}$ μM
R_{22}	22 ± 4	20^{+1}_{-1}
R_{60}	60 ± 20	74^{+4}_{-3}
R_{124}	124 ± 30	130^{+6}_{-6}
R_{260}	260 ± 40	257^{+9}_{-11}
R_{1220}	1220 ± 160	1191^{+32}_{-55}
R_{1740}	1740 ± 340	1599^{+75}_{-87}
O1 $\Delta\varepsilon_{RA}$	$-15.3 \pm 0.2\ k_BT$	$-15.2^{+0.1}_{-0.1}\ k_BT$
O2 $\Delta\varepsilon_{RA}$	$-13.9 \pm 0.2\ k_BT$	$-13.6^{+0.1}_{-0.1}\ k_BT$
O3 $\Delta\varepsilon_{RA}$	$-9.7 \pm 0.1\ k_BT$	$-9.4^{+0.1}_{-0.1}\ k_BT$
Oid $\Delta\varepsilon_{RA}$	$-17.0 \pm 0.2\ k_BT$	$-17.7^{+0.2}_{-0.1}\ k_BT$

Table 2.3: **Global fit of all parameter values using the entire data set in Fig. 2.6.** In addition to fitting the repressor inducer dissociation constants K_A and K_I as was done in the text, we also fit the repressor DNA binding energy $\Delta\varepsilon_{RA}$ as well as the repressor copy numbers R for each strain. The middle columns show the previously reported values for all $\Delta\varepsilon_{RA}$ and R values, with ± representing the standard deviation of three replicates. The right column shows the global fits from this work, with the subscript and superscript notation denoting the 95% credible region. Note that there is overlap between all of the repressor copy numbers and that the net difference in the repressor-DNA binding energies is less than 1 k_BT. The logarithms $\tilde{k}_A = -\log \frac{K_A}{1\,\mathrm{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\mathrm{M}}$ of the dissociation constants were fit for numerical stability.

2.12 Supplemental Information: Applicability of theory to the Oid operator sequence

In addition to the native operator sequences (O1, O2, and O3) considered in the main text, we were also interested in testing our model predictions against the synthetic Oid operator. In contrast to the other operators, Oid is one base pair shorter in length (20 bp), is fully symmetric, and is known to provide stronger repression than the native operator sequences considered so far. While the theory should be similarly applicable, measuring the lower fold-changes associated with this YFP construct was expected to be near the sensitivity limit for our flow cytometer, due to the especially strong binding energy of Oid ($\Delta\varepsilon_{RA} = -17.0\ k_BT$) (Garcia et al., 2011). Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry. Supplemental Section 2.8 gives a detailed explanation of how microscopy measurements were used to obtain induction curves.

We follow the approach of the main text and make fold-change predictions based on the parameter estimates from our strain with $R = 260$ and an O2 operator. These predictions are shown in Fig. 2.29(A), where we also plot data taken in triplicate for strains containing $R = 22$, 60, and 124, obtained by single-cell microscopy. We find that the data are systematically below the theoretical predictions. We also considered our global fitting approach (see Supplemental Section 2.11) to see whether we might find better agreement with the observed data. Interestingly, we find that the majority of the parameters remain largely unchanged, but our estimate for the Oid binding energy $\Delta\varepsilon_{RA}$ is shifted to $-17.7\ k_BT$ instead of the value $-17.0\ k_BT$ found by Garcia and Phillips (2011). In Fig. 2.29(B) we again plot the Oid fold-change data but with theoretical predictions using the new estimate for the Oid binding energy from our global fit and find substantially better agreement.

Fig. 2.30 shows the cumulative data from Garcia and Phillips (2011) and Brewster et al. (2014), as well as our data with $c = 0\ \mu$M, which all measured fold-change for the same simple repression architecture utilizing different reporters and measurement techniques. We find that the binding energies from the global fit, including $\Delta\varepsilon_{RA} = -17.7\ k_BT$, compare reasonably well with all previous measurements.

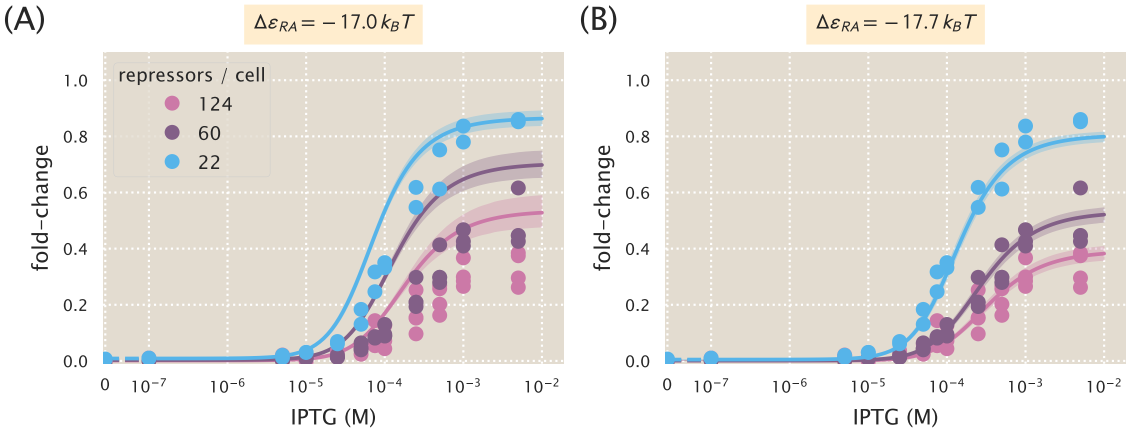


Figure 2.29: **Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers.** (A) Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with $R = 260$. Here we use the previously measured binding energy $\Delta\varepsilon_{RA} = -17.0\ k_BT$ (Garcia and Phillips, 2011). (B) The same experimental data is plotted against the best-fit parameters using the complete O1, O2, O3, and Oid data sets to infer K_A, K_I, repressor copy numbers, and the binding energies of all operators (see Supplemental Section 2.11). Here the major difference in the inferred parameters is a shift in the binding energy for Oid from $\Delta\varepsilon_{RA} = -17.0\ k_BT$ to $\Delta\varepsilon_{RA} = -17.7\ k_BT$, which now shows agreement between the theoretical predictions and experimental data. Shaded regions from the theoretical curves denote the 95% credible region. These are narrower in Panel (B) because the inference of parameters was performed with much more data, and hence the best-fit values are more tightly constrained. Individual data points are shown due to the small number of replicates. The dashed lines at 0 IPTG indicate a linear scale, whereas solid lines represent a log scale.

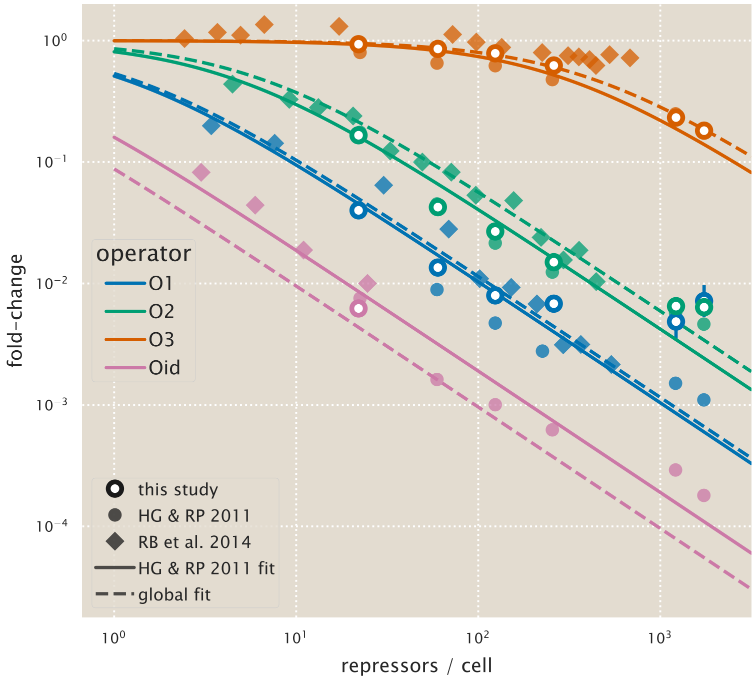


Figure 2.30: **Comparison of fold-change predictions based on binding energies from Garcia and Phillips and those inferred from this work.** Fold-change curves for the different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ are plotted as a function of repressor copy number when IPTG concentration $c = 0$. Solid curves use the binding energies determined from Garcia and Phillips (2011), while the dashed curves use the inferred binding energies we obtained when performing a global fit of K_A, K_I, repressor copy numbers, and the binding energies using all available data from our work. Fold-change measurements from our experiments (outlined circles) Garcia and Phillips (2011) (solid circles), and Brewster et al. (2014) (diamonds) show that the small shifts in binding energy that we infer are still in agreement with prior data. Note that only a single flow cytometry data point is shown for Oid from this study, since the $R = 60$ and $R = 124$ curves from Fig. 2.29 had extremely low fold-change in the absence of inducer ($c = 0$) so as to be indistinguishable from autofluorescence, and in fact their fold-change values in this limit were negative and hence do not appear on this plot.

2.13 Supplemental Information: Comparison of parameter estimation and fold-change predictions across strains

The inferred parameter values for K_A and K_I in the main text were determined by fitting to induction fold-change measurements from a single strain ($R = 260$, $\Delta\varepsilon_{RA} = -13.9\, k_BT, n = 2$, and $\Delta\varepsilon_{AI} = 4.5\, k_BT$). After determining these parameters, we were able to predict the fold-change of the remaining strains without any additional fitting. However, the theory should be independent of the specific strain used to estimate K_A and K_I; using any alternative strain to fit K_A and K_I should yield similar predictions. For the sake of completeness, here we discuss the values for K_A and K_I that are obtained by fitting to each of the induction data sets individually. These fit parameters are shown in Fig. 2.6(D) of the main text, where we find close agreement between strains, but with some deviation and poorer inferences observed with the O3 operator strains. Overall, we find that regardless of which strain is chosen to determine the unknown parameters, the predictions laid out by the theory closely match the experimental measurements. Here we present a comparison of the strain specific predictions and measured fold-change data for each of the three operators considered.

We follow the approach taken in the main text and use Eq. (2.5) to infer values for K_A and K_I by fitting to each combination of binding energy $\Delta\varepsilon_{RA}$ and repressor copy number R. We then use these fitted parameters to predict the induction curves of all other strains. In Fig. 2.31 we plot these fold-change predictions along with experimental data for each of our strains that contains an O1 operator. To make sense of this plot consider the first row as an example. In the first row, K_A and K_I were estimated using data from the strain containing $R = 22$ and an O1 operator (top leftmost plot, shaded in gray). The remaining plots in this row show the predicted fold-change using these values for K_A and K_I. In each row, we then infer K_A and K_I using data from a strain containing a different repressor copy number ($R = 60$ in the second row, $R = 124$ in the third row, and so on). In Fig. 2.32 and Fig. 2.33, we similarly apply this inference to our strains with O2 and O3 operators, respectively. We note that the overwhelming majority of predictions closely match the experimental data.The notable exception is that using the $R = 22$ strain provides poor predictions for the strains with large copy numbers (especially $R = 1220$ and $R = 1740$), though it should be noted that predictions made from the $R = 22$ strain have considerably broader credible regions. This loss in predictive power is due to the poorer estimates of K_A and K_I for the $R = 22$ strain as shown in Fig. 2.6(D).

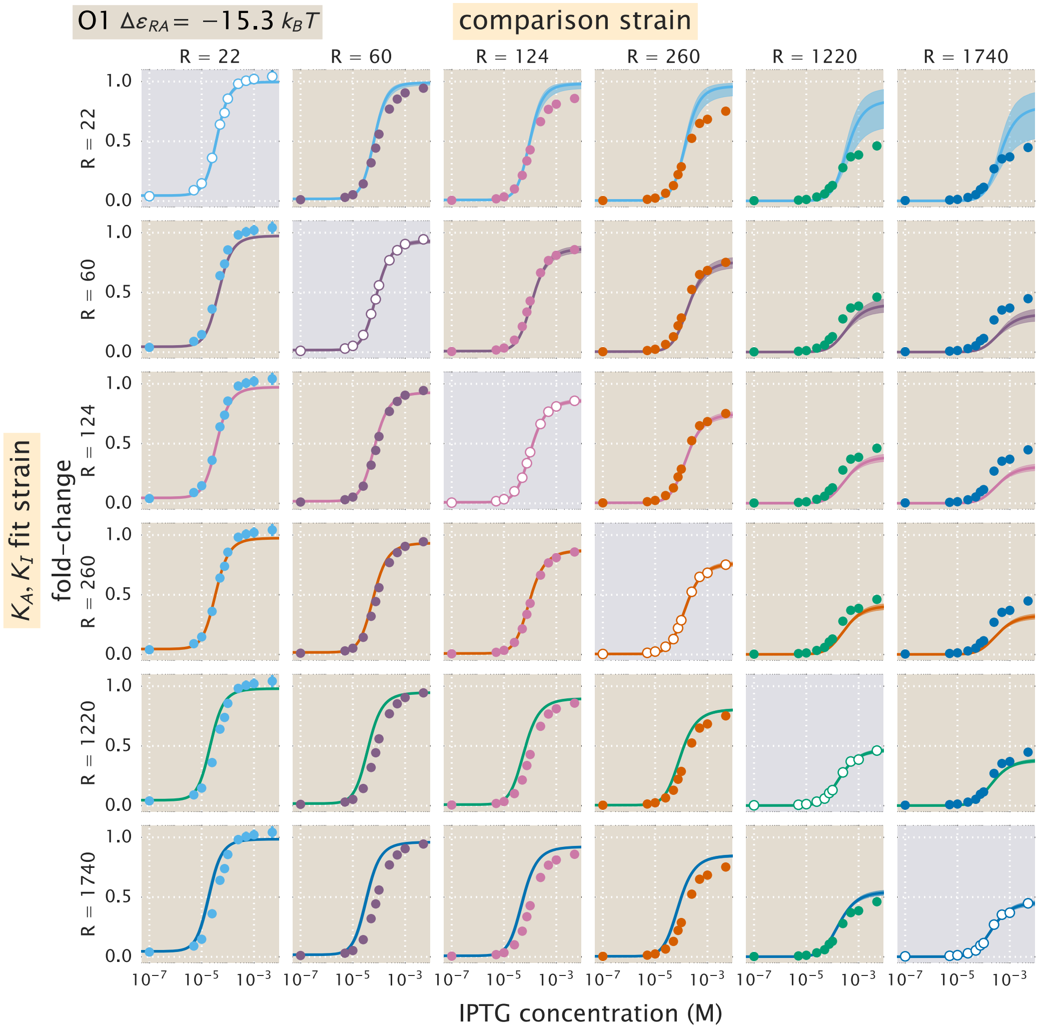


Figure 2.31: **O1 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O1 operator. The solid points correspond to the mean experimental value. The solid lines correspond to Eq. (2.5) using the parameter estimates of K_A and K_I. Each row uses a single set of parameter values based on the strain noted on the left axis. The shaded plots along the diagonal are those where the parameter estimates are plotted along with the data used to infer them. Values for repressor copy number and operator binding energy are from Garcia and Phillips (2011). The shaded region on the curve represents the uncertainty from our parameter estimates and reflects the 95% highest probability density region of the parameter predictions.

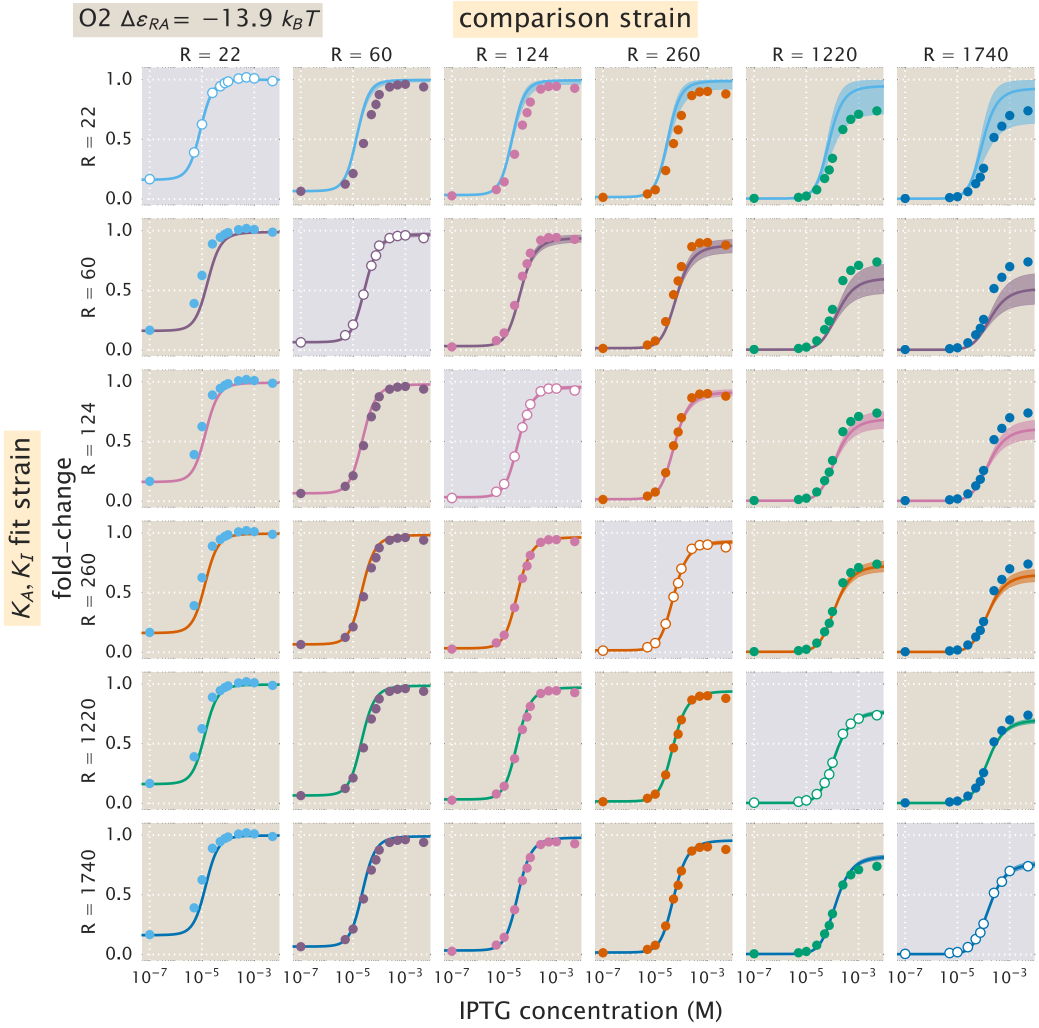


Figure 2.32: **O2 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O2 operator. The plots and data shown are analogous to Fig. 2.31, but for the O2 operator.

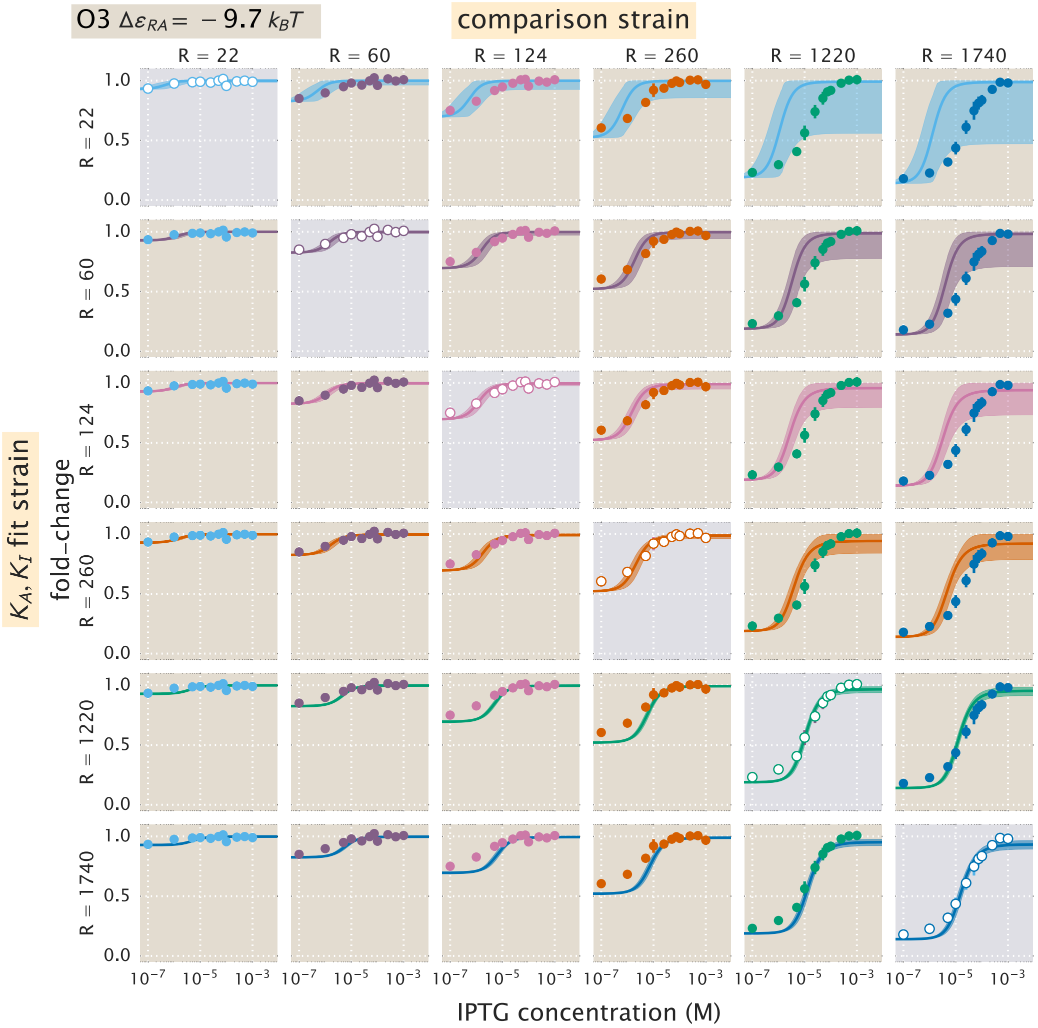


Figure 2.33: **O3 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O3 operator. The plots and data shown are analogous to Fig. 2.31, but for the O3 operator. We note that when using the $R = 22$ O3 strain to predict K_A and K_I, the large uncertainty in the estimates of these parameters (see Fig. 2.6(D)) leads to correspondingly wider credible regions.

2.14 Supplemental Information: Properties of induction titration curves

In this section, we expand on the phenotypic properties of the induction response that were explored in the main text (see Fig. 2.1). We begin by expanding on our discussion of dynamic range and then show the analytic form of the $[EC_{50}]$ for simple repression.

As stated in the main text, the dynamic range is defined as the difference between the maximum and minimum system response, or equivalently, as the difference between the saturation and leakiness of the system. Using Eqs. (2.6)-(2.8), the dynamic range is given by

$$\text{dynamic range} = \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} - \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}. \tag{2.52}$$

The dynamic range, along with saturation and leakiness were plotted with our experimental data in Fig. 2.7(A)-(C) as a function of repressor copy number. Fig. 2.34 shows how these properties are expected to vary as a function of the repressor-operator binding energy. Note that the resulting curves for all three properties have the same shape as in Fig. 2.7(A)-(C), since the dependence of the fold-change upon the repressor copy number and repressor-operator binding energy are both contained in a single multiplicative term, $Re^{-\beta\Delta\varepsilon_{RA}}$. Therefore, increasing R on a logarithmic scale (as in Fig. 2.7(A)-(C)) is equivalent to decreasing $\Delta\varepsilon_{RA}$ on a linear scale (as in Fig. 2.34).

An interesting aspect of the dynamic range is that it exhibits a peak as a function of either the repressor copy number (or equivalently of the repressor-operator binding energy). Differentiating the dynamic range Eq. (2.52) and setting it equal to zero, we find that this peak occurs at

$$\frac{R^*}{N_{NS}} = e^{-\beta(\Delta\varepsilon_{AI}-\Delta\varepsilon_{RA})}\sqrt{e^{\Delta\varepsilon_{AI}}+1}\sqrt{e^{\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n}. \tag{2.53}$$

The magnitude of the peak is given by

$$\text{max dynamic range} = \frac{\left(\sqrt{e^{\Delta\varepsilon_{AI}}+1} - \sqrt{e^{\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n}\right)^2}{\left(\frac{K_A}{K_I}\right)^n - 1}, \tag{2.54}$$

which is independent of the repressor-operator binding energy $\Delta\varepsilon_{RA}$ or R, and will only cause a shift in the location of the peak but not its magnitude.

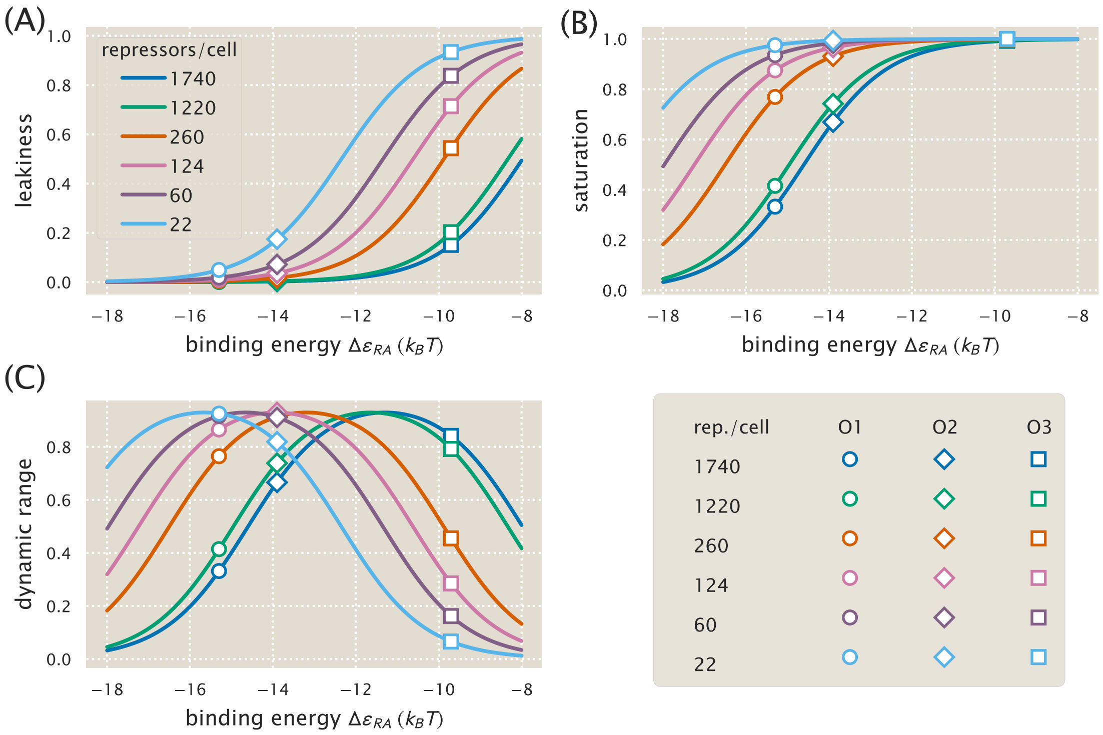


Figure 2.34: **Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number.** Increasing repressor copy number or decreasing the repressor-operator binding energy suppresses gene expression and decreases both the (A) leakiness and (B) saturation. (C) The dynamic range retains its shape but shifts right as the repressor copy number increases. The peak in the dynamic range can be understood by considering the two extremes for $\Delta\varepsilon_{RA}$: for small repressor-operator binding energies, the leakiness is small but the saturation increases with $\Delta\varepsilon_{RA}$; for large repressor-operator binding energies the saturation is near unity and the leakiness increases with $\Delta\varepsilon_{RA}$, thereby decreasing the dynamic range. Repressor copy number does not affect the maximum dynamic range (see Eq. (2.54)). Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively, demonstrating the expected values of the properties using those strains.

We now consider the two remaining properties, the $[EC_{50}]$ and effective Hill coefficient, which determine the horizontal properties of a system - that is, they determine the range of inducer concentration in which the system's response goes from its minimum to maximum values. The $[EC_{50}]$ denotes the inducer concentration required to generate fold-change halfway between its minimum and maximum value and was defined implicitly in Eq. (2.9). For the simple repression system, the $[EC_{50}]$

is given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left(\frac{\left(1+\frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_{RA}}\right) + \left(\frac{K_A}{K_I}\right)^n \left(2e^{-\beta\Delta\varepsilon_{AI}} + \left(1+\frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)\right)}{2\left(1+\frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right) + e^{-\beta\Delta\varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\varepsilon_{AI}}} \right)^{\frac{1}{n}}} - 1. \quad (2.55)$$

Using this expression, we can then find the effective Hill coefficient h, which equals twice the log-log slope of the normalized fold-change evaluated at $c = [EC_{50}]$ (see Eq. (2.10)). In Fig. 2.7(D)-(E) we show how these two properties vary with repressor copy number, and in Fig. 2.35 we demonstrate how they depend on the repressor-operator binding energy. Both the $[EC_{50}]$ and h vary significantly with repressor copy number for sufficiently strong operator binding energies. Interestingly, for weak operator binding energies on the order of the O3 operator, it is predicted that the effective Hill coefficient should not vary with repressor copy number. In addition, the maximum possible Hill coefficient is roughly 1.75, which stresses the point that the effective Hill coefficient should not be interpreted as the number of inducer binding sites, which is exactly 2.

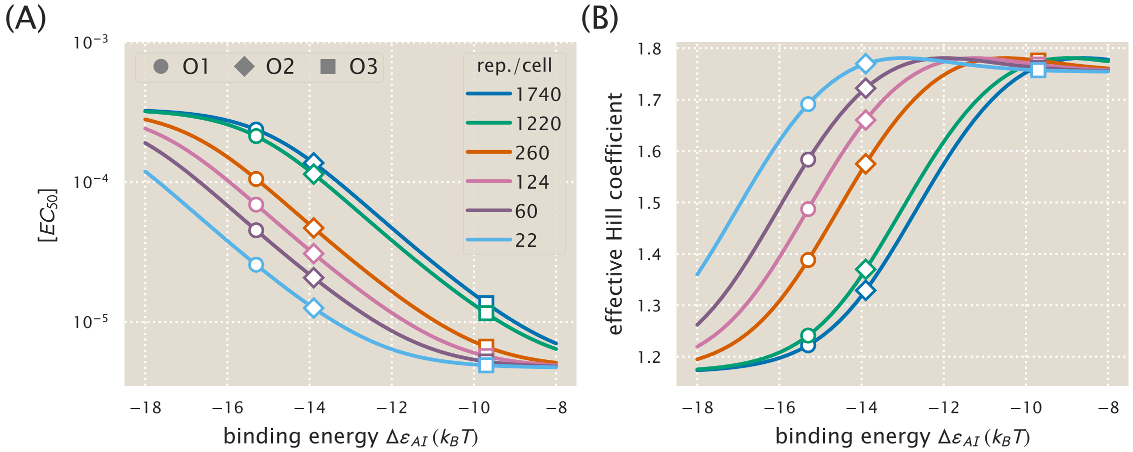


Figure 2.35: **$[EC_{50}]$ and effective Hill coefficient depend strongly on repressor copy number and operator binding energy.** (A) $[EC_{50}]$ values range from very small and tightly clustered at weak operator binding energies (e.g. O3) to relatively large and spread out for stronger operator binding energies (O1 and O2). (B) The effective Hill coefficient generally decreases with increasing repressor copy number, indicating a flatter normalized response. The maximum possible Hill coefficient is roughly 1.75 for all repressor-operator binding energies. Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively.

2.15 Supplemental Information: Applications to other regulatory architectures

In this section, we discuss how the theoretical framework presented in this work is sufficiently general to include a variety of regulatory architectures outside of simple repression by LacI. We begin by noting that the exact same formula for fold-change given in Eq. (2.5) can also describe corepression. We then demonstrate how our model can be generalized to include other architectures, such as a coactivator binding to an activator to promote gene expression. In each case, we briefly describe the system and describe its corresponding theoretical description. For further details, we invite the interested reader to read Bintu et al. (2005) and Marzen et al. (2013).

Corepression

Consider a regulatory architecture where binding of a transcriptional repressor occludes the binding of RNAP to the DNA. A corepressor molecule binds to the repressor and shifts its allosteric equilibrium towards the active state in which it binds more tightly to the DNA, thereby decreasing gene expression (in contrast, an inducer shifts the allosteric equilibrium towards the inactive state where the repressor binds more weakly to the DNA). As in the main text, we can enumerate the states and statistical weights of the promoter and the allosteric states of the repressor. We note that these states and weights exactly match Fig. 2.2 and yield the same fold-change equation as Eq. (2.5),

$$\text{fold-change} \approx \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta \Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_{RA}} \right)^{-1}, \tag{2.56}$$

where c now represents the concentration of the corepressor molecule. Mathematically, the difference between these two architectures can be seen in the relative sizes of the dissociation constants K_A and K_I between the inducer and repressor in the active and inactive states, respectively. The corepressor is defined by $K_A < K_I$, since the corepressor favors binding to the repressor's active state; an inducer must satisfy $K_I < K_A$, as was found in the main text from the induction data (see Fig. 2.5). Much as was performed in the main text, we can make some predictions about the how the response of a corepressor. In Fig. 2.36(A), we show how varying the repressor copy number R and the repressor-DNA binding energy $\Delta\varepsilon_{RA}$ influence the response. We draw the reader's attention to the decrease in fold-change as the concentration of effector is increased.

Activation

We now turn to the case of activation. While this architecture was not studied in this work, we wish to demonstrate how the framework presented here can be extended to include transcription factors other than repressors. To that end, we consider a transcriptional activator which binds to DNA and aids in the binding of RNAP through energetic interaction term ε_{AP}. Note that in this architecture, binding of the activator does not occlude binding of the polymerase. Binding of a coactivator molecule shifts its allosteric equilibrium towards the active state ($K_A < K_I$), where the activator is more likely to be bound to the DNA and promote expression. Enumerating all of the states and statistical weights of this architecture and making the approximation that the promoter is weak generates a fold-change equation of the form

$$\text{fold-change} = \frac{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}} e^{-\beta\varepsilon_{AP}}}{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}}}, \tag{2.57}$$

where A is the total number of activators per cell, c is the concentration of a coactivator molecule, $\Delta\varepsilon_{AA}$ is the binding energy of the activator to the DNA in the active allosteric state, and ε_{AP} is the interaction energy between the activator and the RNAP. Unlike in the cases of induction and corepression, the fold-change formula for activation includes terms from when the RNAP is bound by itself on the DNA as well as when both RNAP and the activator are simultaneously bound to the DNA. Fig. 2.36(B) explores predictions of the fold-change in gene expression by manipulating the activator copy number, DNA binding energy, and the polymerase-activator interaction energy. Note that with this activation scheme, the fold-change must necessarily be greater than one. An interesting feature of these predictions is the observation that even small changes in the interaction energy ($< 0.5\ k_BT$) can result in dramatic increase in fold-change.

As in the case of induction, the Eq. (2.57) is straightforward to generalize. For example, the relative values of K_I and K_A can be switched such that $K_I < K_A$ in which the secondary molecule drives the activator to assume the inactive state represents induction of an activator. While these cases might be viewed as separate biological phenomena, mathematically they can all be described by the same underlying formalism.

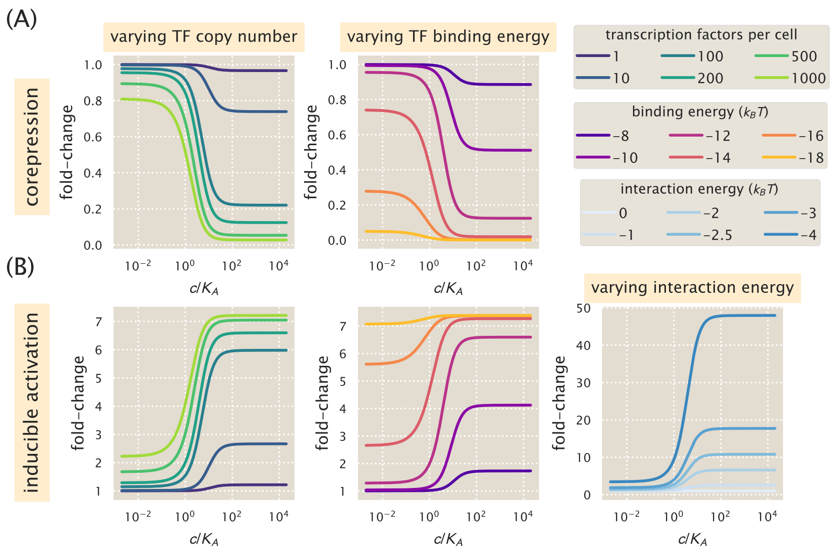


Figure 2.36: **Representative fold-change predictions for allosteric corepression and activation.** (A) Contrary to the case of induction described in the main text, addition of a corepressor decreases fold-change in gene expression. The left and right panels demonstrate how varying the values of the repressor copy number R and repressor-DNA binding energy $\Delta\varepsilon_{RA}$, respectively, change the predicted response profiles. (B) In the case of inducible activation, binding of an effector molecule to an activator transcription factor increases the fold-change in gene expression. Note that for activation, the fold-change is greater than 1. The left and center panels show how changing the activator copy number A and activator-DNA binding energy $\Delta\varepsilon_{AA}$ alter response, respectively. The right panel shows how varying the polymerase-activator interaction energy ε_{AP} alters the fold-change. Relatively small perturbations to this energetic parameter drastically changes the level of activation and plays a major role in dictating the dynamic range of the system.

2.16 Supplemental Information: *E. coli* primer and strain list

Here we provide additional details about the genotypes of the strains used, as well as the primer sequences used to generate them. *E. coli* strains were derived from K12 MG1655. For those containing $R = 22$, we used strain HG104 which additionally has the *lacYZA* operon deleted (positions 360,483 to 365,579) but still contains the native *lacI* locus. All other strains used strain HG105, where both the *lacYZA* and *lacI* operons have both been deleted (positions 360,483 to 366,637).

All 25x+11-yfp expression constructs were integrated at the *galK* locus (between positions 1,504,078 and 1,505,112) while the 3*1x-lacI constructs were integrated at the *ybcN* locus (between positions 1,287,628 and 1,288,047). Integration was performed with λ Red recombineering (Sharan et al., 2009) as described in Garcia and Phillips (2011) using the primers listed in Table 2.4. We follow the notation of Lutz and Bujard (Lutz and Bujard, 1997) for the nomenclature of the different constructs used. Specifically, the first number refers to the antibiotic resistance cassette that is present for selection (2 = kanamycin, 3 = chloramphenicol, and 4 = spectinomycin) and the second number refers to the promoter used to drive expression of either YFP or LacI (1 = $P_{LtetO-1}$, and 5 = *lacUV5*). Note that in 25x+11-yfp, x refers to the LacI operator used, which is centered at +11 (or alternatively, begins at the transcription start site). For the different LacI constructs, 3*1x-lacI, x refers to the different ribosomal binding site modifications that provide different repressor copy numbers and follows from Garcia and Phillips (2011). The asterisk refers to the presence of FLP recombinase sites flanking the chloramphenicol resistance gene that can be used to lose this resistance. However, we maintained the resistance gene in our constructs. A summary of the final genotypes of each strain is listed in Table 2.5. In addition each strain also contained the plasmid pZS4*1-mCherry and provided constitutive expression of the mCherry fluorescent protein. This pZS plasmid is a low copy (SC101 origin of replication) where like with 3*1x-lacI, mCherry is driven by a $P_{LtetO-1}$ promoter.

Primer	**Sequence**	**Comment**
General sequencing primers:		
pZSForwSeq2	TTCCCAACCTTACCAGAGGGC	Forward primer for 3*1x-lacI
251F	CCTTTCGTCTTCACCTCGA	Forward primer for 25x+11-yfp
YFP1	ACTAGCAACACCAGAACAGCCC	Reverse primer for 3*1x-lacI and 25x+11-yfp
Integration primers:		
HG6.1 (*galK*)	gtttgcgcgcagtcagcgatatccattttcgcgaatccggagtgtaagaaACTAGCAACACCAGAACAGCC	Reverse primer for 25x+11-yfp with homology to *galK* locus.
HG6.3 (*galK*)	ttcatattgttcagcgacagcttgctgtacggcaggcaccagctcttccgGGCTAATGCACCCAGTAAGG	Forward primer for 25x+11-yfp with homology to *galK* locus.
galK-control-upstream1	TTCATATTGTTCAGCGACAGCTTG	To check integration.
galK-control-downstream1	CTCCGCCACCGTACGTAAATT	To check integration.
HG11.1 (*ybcN*)	acctctgcggaggggaagcgtgaacctctcacaagacggcatcaaattacACTAGCAACACCAGAACAGCC	Reverse primer for 3*1x-lacI with homology to *ybcN* locus.
HG11.3 (*ybcN*)	ctgtagatgtgtccgttcatgacacgaataagcggtgtagccattacgccGGCTAATGCACCCAGTAAGG	Forward primer for 3*1x-lacI with homology to *ybcN* locus.
ybcN-control-upstream1	AGCGTTTGACCTCTGCGGA	To check integration.
ybcN-control-downstream1	GCTCAGGTTTACGCTTACGACG	To check integration.

Table 2.4: **Primers used in this work.** Lower case sequences denote homology to a chromosomal locus used for integration of the construct into the *E. coli* chromosome. Uppercase sequences refer to the sequences used for PCR amplification.

Strain	Genotype
O1, $R = 0$	HG105::galK⟨⟩25O1+11-yfp
O1, $R = 22$	HG104::galK⟨⟩25O1+11-yfp
O1, $R = 60$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O1, $R = 124$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O1, $R = 260$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O1, $R = 1220$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O1, $R = 1740$	HG105::galK⟨⟩25O1+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
O2, $R = 0$	HG105::galK⟨⟩25O2+11-yfp
O2, $R = 22$	HG104::galK⟨⟩25O2+11-yfp
O2, $R = 60$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O2, $R = 124$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O2, $R = 260$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O2, $R = 1220$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O2, $R = 1740$	HG105::galK⟨⟩25O2+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
O3, $R = 0$	HG105::galK⟨⟩25O3+11-yfp
O3, $R = 22$	HG104::galK⟨⟩25O3+11-yfp
O3, $R = 60$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
O3, $R = 124$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
O3, $R = 260$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS446-lacI
O3, $R = 1220$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1RBS1-lacI
O3, $R = 1740$	HG105::galK⟨⟩25O3+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)
Oid, $R = 0$	HG105::galK⟨⟩25Oid+11-yfp
Oid, $R = 22$	HG104::galK⟨⟩25Oid+11-yfp
Oid, $R = 60$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1147-lacI
Oid, $R = 124$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1027-lacI
Oid, $R = 260$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS446-lacI
Oid, $R = 1220$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1RBS1-lacI
Oid, $R = 1740$	HG105::galK⟨⟩25Oid+11-yfp, ybcN⟨⟩3*1-lacI (RBS1L)

Table 2.5: ***E. coli* strains used in this work.** Each strain contains a unique operator-yfp construct for measurement of fluorescence and R refers to the dimer copy number as measured by Garcia and Phillips (2011).

2.17 Supplemental Information: Effect of chromosomal occupancy by other transcription factors on N_{NS} and the formulation of fold change.

During the review process of the MWC induction work, one reviewer raised several interesting concerns about our derivation of fold change in gene expression and our choice of $N_{NS} = 4.6 \times 10^6$. This choice for N_{NS} reflects the number of nucleotides on the *E. coli* genome, with the parameter itself representing the non-specific background where the repressors and RNA polymerase bind non-specifically on the the chromosome. One particular concern was whether the pool of other transcription factors in the cell might prevent us from treating the entire genome as available for binding by the repressors and RNA polymerase. Another concern was how reasonable it would be to consider N_{NS} as a static parameter as growth conditions are changed. The reviewer was concerned that different growth conditions may lead to a substantial change in genome-wide expression of DNA binding proteins that would then require a redefinition of N_{NS}.

Here we begin by first taking stock of the global transcription factor concentration in *E. coli* using data from a recent proteomic study (Schmidt et al., 2016). We then proceed to show that even after taking this into consideration, the result will only lead to a minor renormalization of the repressor binding energy and would have no effect on our results of Chapter 2. In light of the dependence we observe between N_{NS} and the binding energy, we also consider an alternative formulation of fold change using dissociation constants (Buchler et al., 2003; Kuhlman et al., 2007), but show that this approach will similarly depend on our choice of reference state and is otherwise in agreement with our statistical mechanical formulation of fold change. Lastly, on the reviewers concern over growth conditions and whether a different value of N_{NS} might be needed for each growth condition, we consider the proteomic data noted and find that the DNA binding protein copy number scales with total cellular content and suggests a single choice of N_{NS} may still be reasonable under a variety of growth conditions.

Taking stock of global transcription factor concentration in *E. coli*.

In the work of Schmidt *et al.*, the authors measured the protein copy number across more than half the coding genes (representing greater than 95% by total protein mass; Schmidt et al., 2016). In Figure 2.37(A) the total quantitated protein mass is shown for growth in each of the 22 conditions, ranging from 140 to 370 fg/cell. Of the proteins quantified, we find that 142 proteins are transcription factors or nucleoid-associated proteins that are expected to bind the DNA (based on their

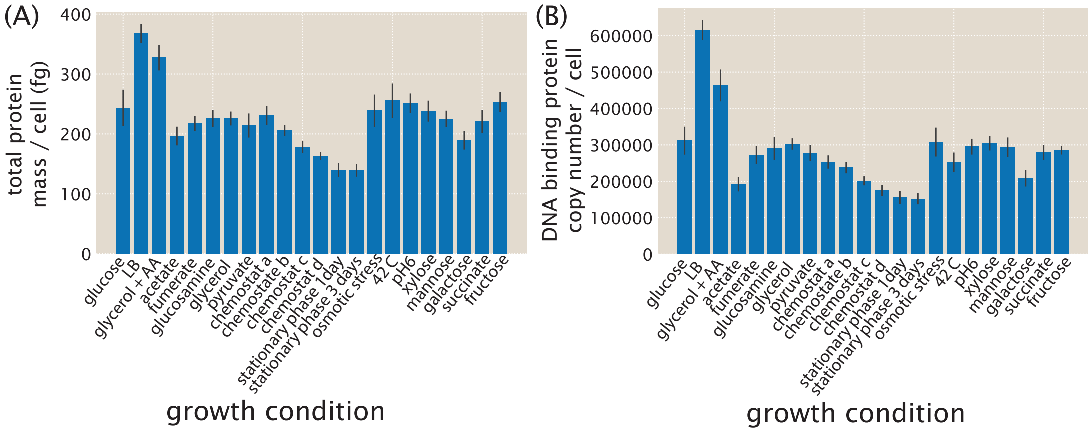


Figure 2.37: **Total cellular protein mass and DNA binding protein copy numbers in *E. coli* across 22 growth conditions.** (A) The total protein mass per cell is shown for each growth condition of Schmidt *et al.*, 2016. These were calculated from all proteins abundances, quantified by mass spectrometry. The values are in line with the expectation that a cell has a mass of about 1,000 fg, with about one-third protein mass. (B) The protein total copy numbers are shown for all DNA binding proteins for each of the 22 growth conditions. DNA binding proteins were identified based on their annotation in the EcoCyc database (Keseler et al., 2010). Error bars are propagated from the reported standard deviations.

annotated function on the database Ecocyc; Keseler et al., 2010). Considering protein copy number instead of mass, we find that there are about 3×10^5 DNA binding proteins per cell when cells are grown in M9 minimal media with 0.5% glucose (the growth condition used in our allostery work). This was found to vary with growth condition (Figure 2.37(B)). For example, growth in LB which is associated with the fastest doubling time has roughly double this copy number.

To make a simple estimate of DNA occupancy from these numbers, let us assume that all transcription factors bind DNA as dimers (since our copy numbers are in monomers per cell, while many transcription factors form complexes in order to bind DNA) and occupy a DNA length of 15 bp (this varies from 7 bp to 38 bp in *E. coli* for transcription factors listed on RegulonDB; Gama-Castro et al., 2016). Considering growth in M9 minimal media with 0.5% glucose, we find that about 2.3 Mbp or about half a genome worth will be occupied (3×10^5 copies / (2 monomers per dimer) × 15 bp per TF).

Effect of genomic occcupany on N_{NS} and fold change.

Now lets see what effect this might have on our expression for fold change. In the most extreme case we could assume that this fraction is totally inaccessible. Ignoring the allosteric nature of the transcription factor for the moment, Garcia and Phillips, 2011, found that fold change was given by,

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}. \tag{2.58}$$

Here R is the repressor copy number, and $\Delta\varepsilon_R$ is the DNA binding energy of the repressor. $\beta = \frac{1}{k_B T}$ where k_B is the Boltzmann constant and T is the temperature. While R has been determined through quantitative western blots (Garcia and Phillips, 2011), the parameter $\Delta\varepsilon_R$ was inferred from experimental measurements of fold-change. If we were to choose a value for N_{NS} different from $N_{NS} = 4.6 \times 10^6$, this will directly effect the inferred value of $\Delta\varepsilon_R$. We can see this by letting $N'_{NS} \equiv \alpha N_{NS}$. Inferring a new binding energy, $\Delta\varepsilon'_{RA}$, that has equally good fit to experimental data, we would require that

$$\Delta\varepsilon'_R - k_B T \ln\left(\frac{R}{\alpha N_{NS}}\right) = \Delta\varepsilon_R - k_B T \ln\left(\frac{R}{N_{NS}}\right). \tag{2.59}$$

Solving Eq. (2.59) for $\Delta\varepsilon'_{RA}$ gives

$$\Delta\varepsilon'_R = \Delta\varepsilon_R + k_B T \ln\left(\frac{N_{NS}}{\alpha N_{NS}}\right) \tag{2.60}$$

$$= \Delta\varepsilon_R - k_B T \ln \alpha. \tag{2.61}$$

If we consider a situation where only half of the DNA is available for binding by RNA polymerase and repressor, (i.e. $N'_{NS} \equiv 0.5 \cdot N_{NS}$), we find that this will only change our inferred value for the binding energy $\Delta\varepsilon'_R$ by ln(2) ≈ 0.7 $k_B T$. With respect to the inferred binding energies, which we used with $N_{NS} = 4.6 \times 10^6$, the only effect is a renormalization of the binding energies (such that they are consistent with the fold change data of Garcia and Phillips, 2011). Otherwise, such a change to N_{NS} will have no effect on our inferences and conclusions more generally. Here we are trying to use one minimal set of parameters across all data generated in the group.

Explicit inclusion of non-specific transcription factors in model of simple repression.

Alternatively, we can take a more explicit approach by including the pool of all DNA binding proteins directly in our model. In Figure 2.38 we show the states and weights for the simple repression architecture. The first three states are the same as those used to calculate fold change in the conventional simple repression model (see Equation 2.58). In addition, we now included a fourth state (or set of states) where other DNA binding proteins might bind to the promoter non-specifically. Since non-specific binding represents our reference energy state (See Model Section of Chapter 2), these additional states will only contribute an entropic term to the partition function. We can calculate p_{bound}, which if we invoked the weak promoter approximation ($\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P} \ll 1$) will be given by

$$p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + L \cdot \frac{C_{ns}}{N_{NS}} + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}. \tag{2.62}$$

L represents the number of ways the other DNA binding proteins may bind the promoter non-specifically, and for simplicity is taken as the length of the promoter region ($L \approx 60$ bp). C_{ns} represents the copy number for this pool of DNA binding proteins, where we have treated it as a single protein species for simplicity. Fold change, which is the ratio of $p_{\text{bound}}(R \geq 0)$ to $p_{\text{bound}}(R = 0)$, will be given by

$$\text{fold-change} = \frac{1 + L \cdot \frac{C_{ns}}{N_{NS}}}{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}} \cdot \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + L \cdot \frac{C_{ns}}{N_{NS}} + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}. \tag{2.63}$$

The RNA polymerase components $\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}$ cancel out and upon some rearrangement, we find that

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R} (1 + L \cdot \frac{C_{ns}}{N_{NS}})^{-1}}. \tag{2.64}$$

Using $C_{ns} \approx 1.5 \cdot 10^5$, which is based on our estimate of the total DNA binding protein copy number found above, we calculate a value of $L \cdot \frac{C_{ns}}{N_{NS}} \approx 2$. Appealing to the results of the previous section, we could consider this as a redefinition of $N'_{NS} = N_{NS} * (1 + L \cdot \frac{C_{ns}}{N_{NS}})$. Again, this would only require a renormalization of the binding energy, but otherwise have no substantial effect on our results.

Statistical mechanical versus thermodynamic formulations of fold change

Our ability to redefine both N_{NS} and $\Delta\varepsilon_R$ but still obtain equivalent formulations of fold change identifies a more subtle point, namely, that the definition of fold change will depend on our choice of reference energies and reference states. It is also common to formulate fold change in the language of dissociations constants (referred to as the the thermodynamic formulation) instead of binding energies (Ackers et al., 1982; Buchler et al., 2003; Kuhlman et al., 2007), but we show below that this similarly entails an arbitrary definition of reference state. We can reconcile the

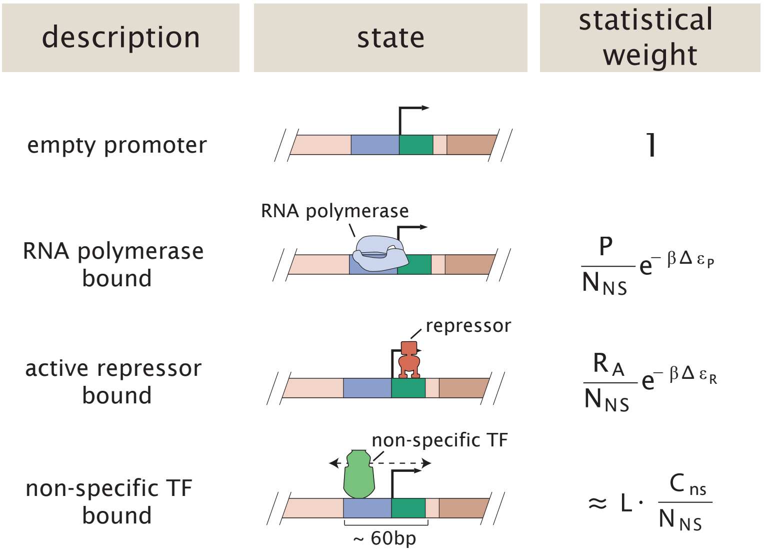


Figure 2.38: **States and Weights for simple repression with pool of non-specific DNA binding proteins.** RNA polymerase (light blue), a repressor, and other non-specific DNA binding proteins compete for binding to a promoter of interest. The difference in energy between a repressor bound to the promoter of interest versus another non-specific site elsewhere on the DNA equals $\Delta\varepsilon_R$; the P RNAP have a corresponding energy difference $\Delta\varepsilon_P$ relative to non-specific binding on the DNA. In addition, there are C_{ns} DNA binding proteins per cell that may bind the promoter of length $L \approx 60$ bp. These proteins are assumed to bind non-specifically and therefore only contribute an entropic term. N_{NS} represents the number of non-specific binding sites on the genome.

statistical mechanical formulation and the dissociation constant approach (i.e. the biochemical formulation), when an appropriate choice of reference state is made.

Following the approach of Buchler et al., 2003 and Kuhlman et al., 2007, we can define p_{bound} by

$$p_{bound} = \frac{[P]}{K_P} \cdot \frac{1}{1 + \frac{[R]}{K_R}}, \tag{2.65}$$

where $[P]$ and $[R]$ represent the molar concentrations of RNA polymerase and repressor within the cell, respectively. K_P and K_R represent the dissociation constants of polymerase and repressor bound to DNA, respectively. Using Equation 2.65, fold change will be given by

$$\text{fold-change} = \frac{1}{1 + \frac{[R]}{K_R}}. \tag{2.66}$$

Following the convention in biochemistry, the dissociation constant K_R is defined to describe the reaction,

$$[\text{R–O}] \rightleftharpoons [\text{R}] + [\text{O}] \tag{2.67}$$

with K_R given by the ratio of off and on rates between the repressor and operator DNA (denoted by [O]), and which is related to the standard free energy ΔG° by,

$$\Delta G^\circ = -N_A k_B T \ln K_R. \tag{2.68}$$

N_A is Avogadro's number. Note that in contrast to the free energy $\Delta\varepsilon_R$, ΔG° is usually reported in units of energy per mol.

We can relate the formulation of fold change above to our statistical mechanical result shown in Equation 2.58 by comparing the terms in the denominator. Specifically, for Equation 2.58 and Equation 2.65 to be equivalent, we require

$$\frac{[R]}{K_R} = \frac{R}{N_{NS}} \cdot e^{\frac{-\Delta\varepsilon_R}{k_B T}} = \frac{[R]V_{\text{cell}}}{N_{NS}} \cdot e^{\frac{-\Delta\varepsilon_R}{k_B T}}. \tag{2.69}$$

Here V_{cell} refers to the volume of the cell and allows us to convert between copy number and cellular concentration. What we find is that in this instance, the dissociation constant used by (Buchler et al., 2003) is related to the binding energy by,

$$K_R = \frac{N_{NS}}{V_{\text{cell}}} \cdot e^{\frac{\Delta\varepsilon_R}{k_B T}} \tag{2.70}$$

which includes the effect of the repressor bound to the genomic DNA through the term N_{NS}. This dissociation constant differs from what one might measure through conventional *in vitro* biochemical assays. An important point to notice is that even in the thermodynamic formulation there is some reference state given by the choice of the standard concentration $c_\circ$ and reference energy. In the statistical mechanical formulation we are somewhat more explicit about the pool of transcription factors, which are moving between the promoter site and the N_{NS} sites on the genomic DNA.

Effect of growth condition on cellular concentration of transcription factors. Lastly, we consider the reviewer's concern over what effect a change in growth condition might have on the abundance of DNA binding proteins and fraction of occupied DNA. In Figure 2.37A we found that the total protein mass per cell varies more than two fold across the different growth conditions. If we instead consider how the relative abundance of DNA binding proteins vary across growth conditions, we find that the fraction of DNA binding proteins appears to scale with total protein (5-7% of the protein by copy number; see Figure 2.39).

Since we only have a measure of protein abundance and not on the cellular DNA content across each of these growth conditions, it is difficult to know whether the fraction of occupied DNA might varied. However, when cells are grown in LB media, other work has found that there are about 3-4 chromosomal copies per cell, while in M9 minimal media with 0.5% glucose, there are 1-2 chromosomal copies per cell. In Figure 2.37B we found that the total copy number of DNA binding proteins was similarly doubled when cells were grown in LB media, as compared to growth in M9 minimal media with 0.5% glucose. This would suggest that the fraction of occupied DNA might not dramatically differ even though their total protein mass (and also their growth rates) is quite different.

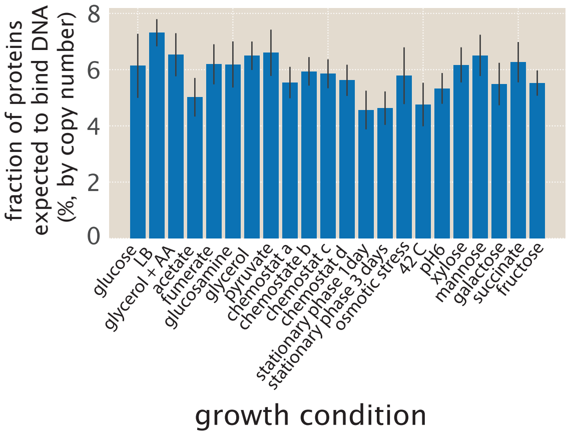


Figure 2.39: **Percent of proteins that are DNA binding proteins in *E. coli* across 22 growth conditions.** The percent of total protein copy number that are DNA binding proteins were calculated from the Schmidt *et al.* data (Schmidt et al., 2016).

References

Ackers, G. K., Johnson, A. D., and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences* 79.4, pp. 1129–33.

Aghaeepour, N., Finak, G., The FlowCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., and Scheuermann, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods* 10.3, pp. 228–238.

Auerbach, A. (2012). Thinking in Cycles: MWC is a Good Model for Acetylcholine Receptor-Channels. *The Journal of Physiology* 590.1, pp. 93–8.

Berg, J., Willmann, S., and Lässig, M. (2004). Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology* 4.1, p. 42.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development* 15.2, pp. 116–124.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005a). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics and Development* 15.2, pp. 116–124.

Boedicker, J. Q., Garcia, H. G., Johnson, S., and Phillips, R. (2013a). DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Physical Biology* 10.6, p. 066005.

Boedicker, J. Q., Garcia, H. G., and Phillips, R. (2013b). Theoretical and Experimental Dissection of DNA Loop-Mediated Repression. *Physical Review Letters* 110.1, p. 018101.

Boulton, S. and Melacini, G. (2016). Advances in NMR Methods To Map Allosteric Sites: From Models to Translation. *Chemical Reviews* 116.11, pp. 6267–6304.

Brewster, R. C., Jones, D. L., and Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology* 8.12.

Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. *Cell* 156.6, pp. 1312–1323.

Brophy, J. A. N. and Voigt, C. A. (2014). Principles of genetic circuit design. *Nature Methods* 11.5, pp. 508–520.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–41.

Canals, M., Lane, J. R., Wen, A., Scammells, P. J., Sexton, P. M., and Christopoulos, A. (2012). A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation. *Journal of Biological Chemistry* 287.1, pp. 650–659.

Chroma Technology Corporation (2016). Chroma Spectra Viewer.

Daber, R., Sharp, K., and Lewis, M. (2009). One Is Not Enough. *Journal of Molecular Biology* 392.5, pp. 1133–1144.

Daber, R., Sochor, M. A., and Lewis, M. (2011). Thermodynamic analysis of mutant lac repressors. *Journal of Molecular Biology* 409.1, pp. 76–87.

Edelstein, A. D., Tsuchida, M. A., Amodaj, N., Pinkard, H., Vale, R. D., and Stuurman, N. (2014). Advanced methods of microscope control using μManager software. *Journal of Biological Methods* 1.2, pp. 10–10.

Einav, T., Mazutis, L., and Phillips, R. (2016). Statistical Mechanics of Allosteric Enzymes. EN. *The Journal of Physical Chemistry B* 121 (15).

Eldar, A. and Elowitz, M. (2010). Functional roles for noise in genetic circuits. *Nature* 467.7312, 167–173.

Fernández-Castané, A., Vine, C. E., Caminal, G., and López-Santín, J. (2012). Evidencing the role of lactose permease in IPTG uptake by *Escherichia coli* in fed-batch high cell density cultures. *Journal of Biotechnology* 157.3, pp. 391–398.

Forsén, S. and Linse, S. (1995). Cooperativity: over the Hill. *Trends in Biochemical Sciences* 20.12, pp. 495 –497.

Frank, S. (2013). Input-output relations in biological systems: measurement, information and the Hill equation. *Biology Direct* 8.1, p. 31.

Gama-Castro, S. et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* 44.D1, pp. D133–D143.

Garcia, H. G. and Phillips, R. (2011). Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–8.

Garcia, H. G., Lee, H. J., Boedicker, J. Q., and Phillips, R. (2011). Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. *Biophysical Journal* 101.3, 535–544.

Garcia, H. G., Sanchez, A., Boedicker, J. Q., Osborne, M., Gelles, J., Kondev, J., and Phillips, R. (2012). Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Reports* 2.1, pp. 150–161.

Gardino, A. K., Volkman, B. F., Cho, H. S., Lee, S.-Y., Wemmer, D. E., and Kern, D. (2003). The NMR Solution Structure of BeF3-Activated Spo0F Reveals the Conformational Switch in a Phosphorelay System. *Journal of Molecular Biology* 331.1, pp. 245–254.

Gerland, U. and Hwa, T. (2002). On the Selection and Evolution of Regulatory DNA Motifs. *Journal of Molecular Evolution* 55.4, pp. 386–400.

Huang, Z. et al. (2011). ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Research* 39, p. D663.

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346.6216, pp. 1533–1536.

Kao-Huang, Y., Revzin, A., Butler, A. P., O'Conner, P., Noble, D. W., and Hippel, P. H. von (1977). Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound Escherichia coli lac repressor in vivo. *Proceedings of the National Academy of Sciences* 74.10, pp. 4228–32.

Keseler, I. M. et al. (2010). EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Research* 39.Database, pp. D583–D590.

Keymer, J. E., Endres, R. G., Skoge, M., Meir, Y., and Wingreen, N. S. (2006). Chemosensing in *Escherichia coli*: two regimes of two-state receptors. *Proceedings of the National Academy of Sciences* 103.6, pp. 1786–91.

Klumpp, S. and Hwa, T. (2008). Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences* 105.51, pp. 20245–50.

Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences* 104.14, pp. 6043–6048.

Levantino, M., Spilotros, A., Cammarata, M., Schirò, G., Ardiccioni, C., Vallone, B., Brunori, M., and Cupane, A. (2012). The Monod-Wyman-Changeux allosteric model accounts for the quaternary transition dynamics in wild type and a recombinant mutant human hemoglobin. *Proceedings of the National Academy of Sciences* 109.37, pp. 14894–9.

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., and Lu, P. (1996). Crystal Structure of the Lactose Operon Repressor and its Complexes with DNA and Inducer. *Science* 271.5253, pp. 1247–54.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 157.3, pp. 624–635.

Lindsley, J. E. and Rutter, J. (2006). Whence cometh the allosterome? *Proceedings of the National Academy of Sciences* 103.28, pp. 10533–5.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* 73A.4, pp. 321–332.

Lutz, R. and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research* 25.6, pp. 1203–10.

Maecker, H. T. et al. (2005). Standardization of cytokine flow cytometry assays. *BMC Immunology* 6.1, p. 13.

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society B: Biological Sciences* 207.1167, pp. 187–217.

Martins, B. M. C. and Swain, P. S. (2011). Trade-Offs and constraints in allosteric sensing. *PLoS Computational Biology* 7.11, pp. 1–13.

Marzen, S., Garcia, H. G., and Phillips, R. (2013). Statistical mechanics of Monod-Wyman-Changeux (MWC) models. *Journal of Molecular Biology* 425.9, pp. 1433–1460.

Milo, R., Hou, J. H., Springer, M., Brenner, M. P., and Kirschner, M. W. (2007). The relationship between evolutionary and physiological variation in hemoglobin. *Proceedings of the National Academy of Sciences* 104.43, pp. 16998–17003.

Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* 107.52, pp. 22534–9.

Monod, J., Changeux, J.-P., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology* 6, pp. 306–329.

Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the Nature of Allosteric Transitions: A Plausible Model. *Journal of Molecular Biology* 12, pp. 88–118.

Moon, T. S., Lou, C., Tamsir, A., Stanton, B. C., and Voigt, C. A. (2012). Genetic programs constructed from layered logic gates in single cells. *Nature* 491.7423, pp. 249–253.

Murphy, K. F., Balázsi, G., and Collins, J. J. (2007). Combinatorial promoter design for engineering noisy gene expression. *Proceedings of the National Academy of Sciences* 104.31, pp. 12726–12731.

Murphy, K. F., Adams, R. M., Wang, X., Balázsi, G., and Collins, J. J. (2010). Tuning and controlling gene expression noise in synthetic gene networks. *Nucleic Acids Research* 38.8, pp. 2712–2726.

Oehler, S., Amouyal, M., Kolkhof, P., Wilcken-Bergmann, B. von, and Müller-Hill, B. (1994). Quality and position of the three lac operators of *E. coli* define efficiency of repression. *The EMBO Journal* 13.14, pp. 3348–3355.

Oehler, S., Alberti, S., and Müller-Hill, B. (2006). Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Research* 34.2, pp. 606–612.

O’Gorman, R. B., Rosenberg, J. M., Kallai, O. B., Dickerson, R. E., Itakura, K., Riggs, A. D., and Matthews, K. S. (1980). Equilibrium binding of inducer to lac repressor-operator DNA complex. *Journal of Biological Chemistry* 255.21, pp. 10107–10114.

Phillips, R. (2015). Napoleon Is in Equilibrium. *Annual Review of Condensed Matter Physics* 6.1, pp. 85–111.

Poelwijk, F. J., deVos, M. G. J., and Tans, S. J. (2011). Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell* 146.3, pp. 462–470.

Rogers, J. K., Guzman, C. D., Taylor, N. D., Raman, S., Anderson, K., and Church, G. M. (2015). Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research* 43.15, pp. 7648–7659.

Rohlhill, J., Sandoval, N. R., and Papoutsakis, E. T. (2017). Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated *Escherichia coli* Growth on Methanol. *ACS Synthetic Biology*, Advance online publication.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene Regulation at the Single-Cell Level. *Science* 307.5717, pp. 1962–1965.

Rydenfelt, M., Garcia, H. G., Cox, R. S., and Phillips, R. (2014). The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS ONE* 9.12, pp. 1–31.

Saiz, L. and Vilar, J. M. G. (2008). Ab initio thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Research* 36.3, p. 726.

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27.10, pp. 946–950.

Schmidt, A. et al. (2015). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology* 34.1, pp. 104–110.

Schmidt, A. et al. (2016). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology* 34 (1), pp. 104–111.

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (2010). Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* 330.6007, pp. 1099–102.

Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U. (2003). Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences* 100.13, pp. 7702–7707.

Sharan, S. K., Thomason, L. C., Kuznetsov, S. G., and Court, D. L. (2009). Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols* 4.2, pp. 206–223.

Shis, D. L., Hussain, F., Meinhardt, S., Swint-Kruse, L., and Bennett, M. R. (2014). Modular, Multi-Input Transcriptional Logic Gating with Orthogonal LacI/GalR Family Chimeras. *ACS Synthetic Biology* 3.9, pp. 645–651.

Sivia, D. and Skilling, J. (2006). Data analysis: a Bayesian tutorial. OUP Oxford.

Sochor, M. A. (2014). *In vitro* transcription accurately predicts lac repressor phenotype *in vivo* in *Escherichia coli*. *PeerJ* 2, e498.

Sourjik, V. and Berg, H. C. (2002). Receptor sensitivity in bacterial chemotaxis. *Proceedings of the National Academy of Sciences* 99.1, pp. 123–127.

Swem, L. R., Swem, D. L., Wingreen, N. S., and Bassler, B. L. (2008). Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio harveyi. *Cell* 134.3, pp. 461–473.

Thomason, L. C., Costantino, N., and Court, D. L. (2007). *E. coli* genome manipulation by P1 transduction. *Current Protocols in Molecular Biology* Chapter 1, Unit 1.17–1.17.8.

Tungtur, S., Skinner, H., Zhan, H., Swint-Kruse, L., and Beckett, D. (2011). In vivo tests of thermodynamic models of transcription repressor function. *Biophysical Chemistry* 159.1, pp. 142–151.

Velyvis, A., Yang, Y. R., Schachman, H. K., and Kay, L. E. (2007). A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences* 104.21, pp. 8815–20.

Vilar, J. M. G. and Leibler, S. (2003). DNA Looping and Physical Constraints on Transcription Regulation. *Journal of Molecular Biology* 331.5, pp. 981–989.

Vilar, J. M. G. and Saiz, L. (2013). Reliable Prediction of Complex Phenotypes from a Modular Design in Free Energy Space: An Extensive Exploration of the *lac* Operon. *ACS Synthetic Biology* 2.10, pp. 576–586.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., and Kegel, W. K. (2014). Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters* 113.25, pp. 1–5.

Weiss, J. N. (1997). The Hill equation revisited: uses and misuses. *The FASEB Journal* 11.11, pp. 835–41.

Zeldovich, K. B. and Shakhnovich, E. I. (2008). Understanding Protein Evolution: From Protein Physics to Darwinian Selection. *Annual Review of Physical Chemistry* 59.1, pp. 105–127.

Chapter 3

CHARACTERIZATION OF THE SEQUENCE-DEPENDENT OCCUPANCY OF LACI.

This work was performed in collaboration with S. L. Barnes, W. T. Ireland, J. B. Kinney, and R. Phillips. Author contribution note: for this chapter, I (NB) assisted in Sort-Seq sample processing, strain construction, fold-change measurements, and helped perform data analysis.

3.1 Introduction

High-throughput sequencing has delivered on the promise that we can sequence the genome of nearly any species at will. The amount of genome data available is already enormous and will only continue to grow. However, this mass of data is nearly useless without appropriate methods of analyzing it. Despite decades of research, genomic data still defies our efforts to "read" it. When faced with an entirely new genome, we can guess that a stretch of DNA contains a gene, translate that hypothetical gene into an amino acid sequence, and then guess at the structure of the protein coded for by that gene. In some cases, we can also guess at the locations of transcription start sites and transcription factor binding sites, but these guesses tell us little about the actual usage of these putative sites. A more detailed understanding of how sequence elements control genomic activity is needed in order to improve the accuracy of such predictions. An important avenue for developing this level of understanding is to propose models that map sequence to function and perform experiments that test these models.

A crucial example of the need to map sequence to function is transcriptional regulation. It has proven difficult to determine the precise regulatory mechanisms of individual genes, let alone entire gene networks. Over half of the genes in *E. coli*, which is arguably the best-understood model organism, lack any regulatory annotation (see RegulonDB (Gama-Castro et al., 2016)). Those operons whose regulation is well described (e.g. the *lac*, *rel*, and *mar* operons (Oehler et al., 1990; Gerdes et al., 2005; Alekshun and Levy, 1997)) required decades of work, often involving laborious genetic and biochemical experiments (Minchin and Busby, 2009). A wide variety of new techniques have been proposed and implemented to simplify the process of determining how a gene is regulated. ChIP-based methods such as ChIP-chip and ChIP-seq make it possible to determine the genome-wide binding locations of individual transcription factors of interest. Massively parallel reporter assays (MPRAs) have made it possible to read out transcription factor binding position and occupancy *in vivo* with base-pair resolution, and provide a means for analyzing non-binding features such as "insulator" sequences (Levo et al., 2017; Melnikov et al., 2012; Levy et al., 2017). *In vitro* methods such as protein-binding microarrays (Berger et al., 2006), SELEX (Fields et al., 1997; Jolma et al., 2013), MITOMI (Maerkl and Quake, 2007; Shultzaberger et al., 2012), and binding assays performed in high-throughput sequencing flow cells (Jung et al., 2017; Nutiu et al., 2011) have made it possible to measure transcription factor affinity to a broad array of possible binding sites and develop detailed records of transcription factor sequence specificities.

In spite of this progress, it remains difficult to integrate the various aspects of transcriptional regulation revealed by such experiments into a cohesive understanding of a given promoter or transcription factor. While *in vitro* methods may provide accurate measurements of transcription factor sequence specificities and binding affinities, including insight into the effects of flanking sequences (Dror et al., 2015; Levo et al., 2015), they cannot fully account for the *in vivo* consequences of binding site context and interactions with other proteins. Current *in vivo* methods for determining transcription factor binding affinities, such as bacterial one-hybrid (Christensen et al., 2011; Xu and Noyes, 2015), require a restructuring of the promoter so that it no longer resembles its genomic counterpart. Additionally, while computational efforts to "read" the genome provide a promising avenue for understanding transcriptional regulation in its native context, efforts to computationally ascertain the locations of transcription factor binding sites frequently produce false positives (Weirauch et al., 2013; Djordjevic et al., 2003). Furthermore, a common assumption underlying many of these methods is that transcription factor occupancy in the vicinity of a promoter implies regulation, but it has been shown that occupancy cannot accurately predict the effect of a transcription factor on gene regulation (Garcia et al., 2012; Wunderlich and Mirny, 2009).

An ideal technique would be capable of interrogating multiple aspects of transcriptional regulation at once, from locating transcription factor binding sites to identifying the sequence specificity of these binding sites. As previously noted, massively parallel reporter assays have shown a great deal of promise for this reason. In Ref. Brewster et al., 2012, we showed that the MPRA Sort-Seq (Kinney et al., 2010), combined with a simple linear model for protein-DNA binding specificity, can be used to accurately predict the binding energies of multiple RNAP binding site mutants, serving as a jumping off point for the use of such models as a quantitative tool in synthetic biology. Here we adopt a similar philosophy to explore whether this technique can be more broadly applied to other regulatory components such as transcription factor binding sites. Specifically, we use Sort-Seq to map sequence to binding energy for the repressor-binding site interaction, and we rigorously characterize the variables that must be considered in order to obtain an accurate sequence-binding energy map. Then, we show how such a mapping can be used to characterize how sequence controls protein binding and, ultimately, gene expression. As concrete applications of this approach, we show that our sequence-energy mapping can be used to precisely design a series of binding sites with a hierarchy of precisely controlled binding energies. With this suite of different binding energies in hand, we then show how those binding sites can be used to design a wide range of induction responses with different phenotypic properties such as leakiness, dynamic range and $[EC_{50}]$. Finally, we use Sort-Seq to also consider the consequence of single amino acid perturbations to our mapping of DNA sequence specificity. This broad collection of case studies provides a rigorous test of the quantitative mapping between regulatory sequence and function offered by the Sort-Seq approach.

3.2 Results

In order to map regulatory sequence to binding energy *in vivo*, we applied Sort-Seq (Kinney et al., 2010) to synthetically constructed promoters with binding sites for RNA polymerase (RNAP) and *lac* repressor (LacI). As shown in Fig. 3.1A, Sort-Seq works by first generating a library of cells, each of which contains a mutated promoter that drives expression of GFP from a low copy plasmid (5-10 copies per cell; Lutz and Bujard, 1997) and provides a read-out of transcriptional state. We use fluorescence-activated cell sorting (FACS) to sort that library of cells into multiple bins gated by their fluorescence level and then sequence the mutated plasmids from each bin. Binding by LacI to the promoter occludes binding by RNAP (Ackers et al., 1982; Buchler et al., 2003), and mutations to both binding site sequences will influence what bin each cells is sorted into.

One of the important aspects demonstrated by Kinney et al., 2010, is that we can use the large sequence data set from Sort-Seq (0.5-2 million sequences) to perform information-based modeling and extract quantitative information from the data. In particular, it is possible to infer energy matrix models that describe the sequence-dependent energy of interaction between transcription factors and their binding sites (Kinney et al., 2010; Ireland and Kinney, 2016). Here we set out to test the accuracy of the models that come from Sort-Seq experiments in the context of the simple repression architecture (Bintu et al., 2005), with repression by LacI as noted above.

In order to be more representative of the range in both transcription factor and protein-DNA binding energies observed in *E. coli* more generally, but also to test the capabilities of the approach more broadly, we constructed a set of strains with a range of repressor copy numbers and DNA binding energies (both key parameters of the simple repression architecture, as we will find in the next section). We performed a set of separate Sort-Seq experiments in *E. coli* strains with mean LacI dimer copy numbers ranging from 22-1740 copies per cell (Fig. 3.1B). We varied the binding site sequence of the LacI binding site in our promoter library, using the three natural sites found at the *lac* operon (O1 with binding energy, -15.3 k_BT; O2, the second strongest,-13.9 k_BT; and O3 the weakest at -9.7 k_BT (Garcia and Phillips, 2011)).

Sequence-dependent thermodynamic model of the simple repression architecture

We begin by defining the thermodynamic model of simple repression that we will apply to our Sort-Seq data. This is identical to the model we considered in Chapter 2, though here we will also define energy matrices that describe the sequence-dependent interaction energies of RNAP and LacI to their binding sites.

We consider a cell with P copies of RNAP per cell and R copies of LacI per cell, and begin by enumerating all possible states of the promoter and their corresponding statistical weights. As shown in Fig. 3.2, the promoter can either be empty, occupied by RNAP, or occupied by LacI. In addition to these specific binding sites, we assume that there are $N_{NS} = 4.6 \times 10^6$ non-specific binding sites elsewhere on the chromosome where RNAP and LacI may bind non-specifically. We define our reference energy such that all specific binding energies are measured relative to the average non-specific binding energy. For simplicity, our model explicitly ignores the complexity of the distribution of non-specific binding affinities in the genome and makes the assumption that a single parameter can capture the energy difference between our binding site of interest and the average site in the reservoir.

Thermodynamic models of transcription assume that gene expression is proportional to the probability that the RNAP is bound to the promoter p_{bound}, and as we have

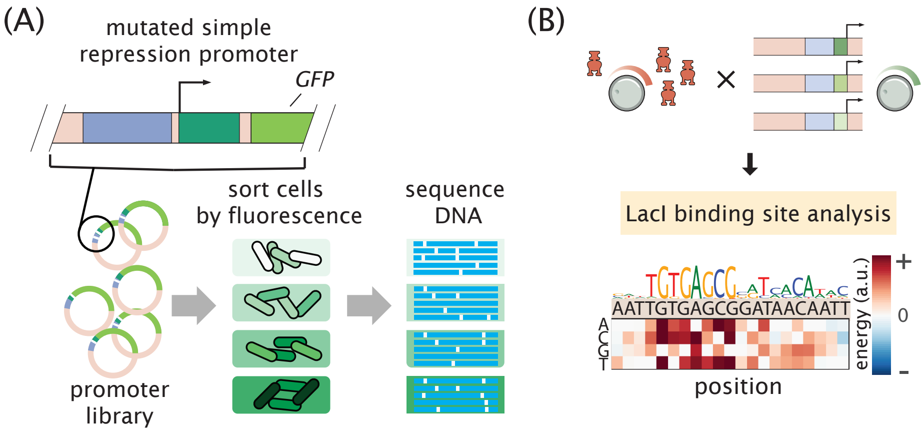


Figure 3.1: **Process flow for using Sort-Seq to obtain energy matrices.** (A) A simple repression motif was designed in which a LacI repressor binding site is placed immediately downstream of the RNAP site. When RNAP binds, it initiates transcription and the GFP reporter gene is expressed. The RNAP and LacI binding sites were both randomly mutated at a rate of approximately 10% and the resulting plasmid library was transformed into cells such that each cell contains a different mutant. We then sort the cell population into bins based on fluorescence level, and then sequence the cells in each bin to map sequence to expression. (B) We analyze simple repression constructs using each of the three natural *lac* operators, O1, O2, and O3, and performed Sort-Seq in *E. coli* strains with mean copy numbers per cell of 22 ± 4, 60 ± 20, 124 ± 30, 260 ± 40, 1220 ± 160, and 1740 ± 340 (using strains from Garcia and Phillips, 2011). The Sort-Seq data was used to infer energy matrices that describe the sequence-dependent repression by LacI. An example energy matrix and sequence logo (Stormo, 2000) are shown for LacI, with the convention that the wild-type nucleotides have zero energy.

found in Chapters 1 and 2, this is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{1 + \frac{p_A(c) R}{N_{NS}} e^{-\beta \Delta \varepsilon_R} + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}, \tag{3.1}$$

with $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant and T is the temperature of the system. Here we have included the allosteric aspect of LacI through the term, $p_A(c)$, which indicates the fraction of active LacI in the presence of inducer. c denote the concentration of inducer present in the cell ($p_A(c) = 1$ when no inducer is present).

We describe the sequence-dependent binding energies for RNAP, $\Delta\varepsilon_P$, and LacI, $\Delta\varepsilon_R$, using linear energy matrix models. The define the binding energy associated with each protein i, $\Delta\varepsilon_i$ ($i = P$ for RNAP, and $i = R$ for LacI), by

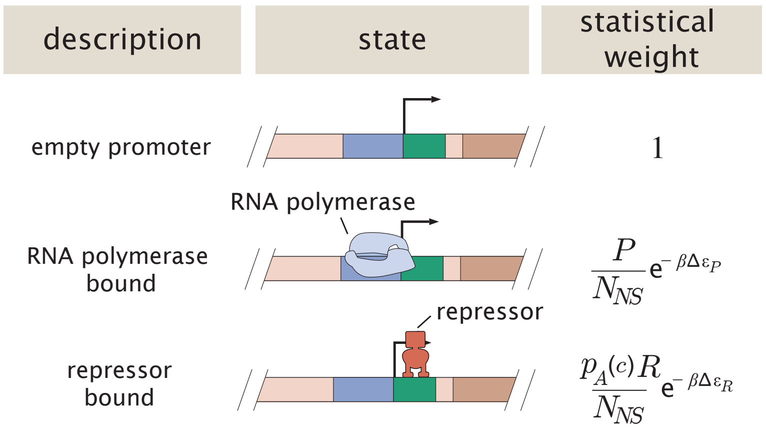


Figure 3.2: States and weights for the simple repression motif. There are P RNA polymerase (blue) and a R repressors (red) per cell that compete for binding to a promoter of interest. The difference in energy between a repressor bound to the promoter of interest versus another non-specific positions elsewhere on the DNA equals $\Delta\varepsilon_R$; the P RNA polymerase have a corresponding energy difference $\Delta\varepsilon_P$ relative to non-specific binding on the DNA. N_{NS} represents the number of non-specific binding sites for both RNA polymerase and repressor.

$$\Delta\varepsilon_i = \alpha\varepsilon_{i,\text{mat}} + \Delta\varepsilon_{i,\text{wt}}, \tag{3.2}$$

where ε_{mat} is the energy value obtained by summing the matrix elements associated with a sequence (further defined below), α_i is a scaling factor that converts the matrix values into k_BT units, and $\Delta\varepsilon_{i,\text{wt}}$ is the binding energy associated with the wild-type operator.

Energy matrices treat each base pair position j along a binding site as contributing a certain amount to the binding energy. Mathematically the energy matrix is described by a 4xL matrix, where each column j of matrix parameters will represent the energies for each nucleotide $i = A, C, G,$ or T (= 1, 2, 3, or 4) associated with position j of the binding site. For example, index ($i = 2, j = 3$) represents the energy parameter for nucleotide C at position 3. The binding energy from an energy matrix will then be given by,

$$\varepsilon_{i,\text{mat}} = \sum_{j=1}^{L} E(S_j), \tag{3.3}$$

where $E(S_j)$ represents the energy contribution of nucleotide S_j, at position j of the binding site. These models can be extended to allow for non-additive contributions from each position, though linear models appear to be sufficient to describe transcription factor binding in bacteria in general (Berg and Hippel, 1987; Benos et al., 2002; Brewster et al., 2012). By convention, we have fixed the values of the matrix positions associated with the wild-type sequence to 0 k_BT, so that $\varepsilon_{i,\text{mat}} = 0$ for a wild-type sequence. Thus, $\alpha_i \varepsilon_{i,,\text{mat}}$ can be interpreted as the change in binding energy relative to the wild-type energy caused by specific mutations in the sequence of interest.

Inferring models of the simple repression architecture using Sort-Seq

We use the MPAthic software to infer the parameters of the energy matrices and thermodynamic parameters of p_{bound} (Kinney et al., 2010; Ireland and Kinney, 2016). The software uses Markov-Chain Monte Carlo (MCMC) to determine the set of parameters that maximize the mutual information between the distribution of sequences in the binned sequence data and the model's predictions. More specifically, the inference approach samples the probability distribution

$$p(\{\theta\}|\{S, f\}) \propto 2^{N \cdot I(f, \{\text{model predictions}\})}. \tag{3.4}$$

Here $\{\theta\}$ is the set of model parameters that define our model (e.g. entries in the energy matrices), $\{S, f\}$ represents our data set of sequences S and the sorted bin f where they were found. N is the number of sequences in the data, and $I(f, \{\text{model predictions}\}))$ is the mutual information between the distribution of binned sequences and the model's predictions, which we discuss further below. A more detailed description of the inference approach can be found in Appendix A.

Due to the computational burden of fitting a large number of parameters by MCMC (all parameters of the energy matrices for RNAP and LacI, and the thermodynamic parameters), we find it convenient to first infer the energy matrices (in arbitrary units) for LacI and RNAP from the Sort-Seq data. Fig. 3.3A summarizes the result for one of the LacI energy matrices (using a O1 binding site library, and E. coli strain with $R = 1740$ LacI per cell). Mutual information is estimated from the joint probability distribution between model prediction and binned sequence data, which is estimated by performing kernel density estimation. Note that in this instance, we are estimating a joint distribution to calculate the mutual information between sequence bin and energy prediction, $I(f, \text{energy (a.u.)})$. We repeat this procedure to generate an energy matrix for the RNAP binding site.

With our energy matrices in hand, we then use the Sort-Seq sequence data to

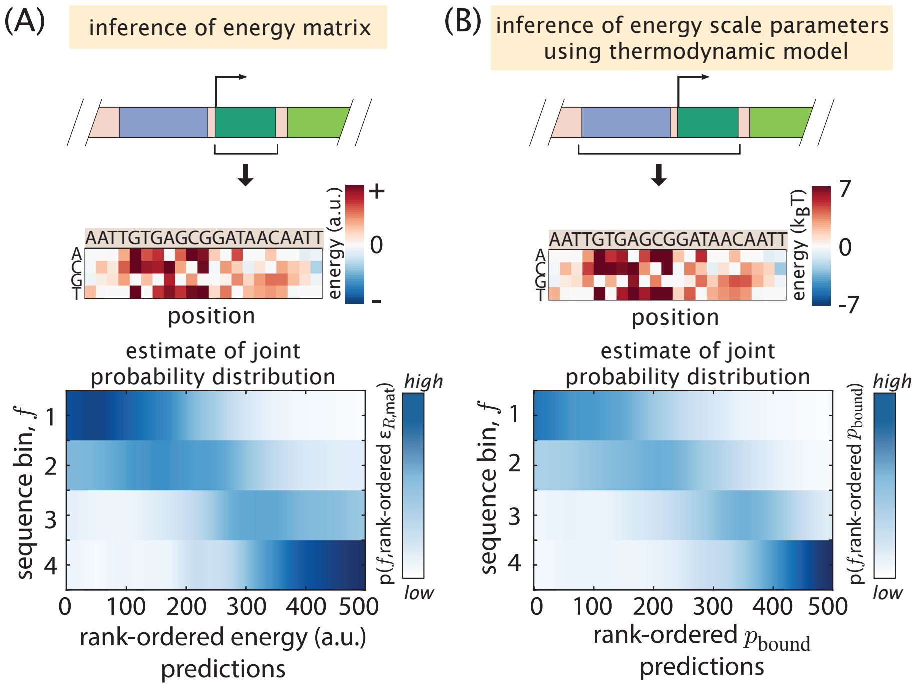


Figure 3.3: **Inference of LacI energy matrices.** (A) Using the aligned sequence data for the LacI binding site, information-based modeling was performed with the MPAthic software (Ireland and Kinney, 2016) to determine the parameters of the LacI energy matrix (in arbitrary units). By convention, the energies are defined such that the O1 wild-type sequence has zero energy. Kernel density estimation was performed to estimate the joint probability distribution between sequence bin f and rank-ordered energy predictions from the inferred matrix. (B) Sort-Seq data was fit to the thermodynamic Eq. (3.1), where binding energies were calculated from the separately inferred energy matrices for LacI and RNAP. The entire promoter sequence from each mutated sequence was used in this inference. This allowed determination of the scaling factors for binding by LacI and the energy matrix shown in absolute k_BT energy units. A joint probability distribution between sequence bin f and rank-ordered predictions of p_{bound} is shown using the inferred model. Data is from the Sort-Seq experiment using an O1 LacI binding site and performed in a strain with R = 1740 repressor copies per cell.

determine the scaling parameters of Eq. (3.2), by fitting the data against the thermodynamic model defined by Eq. (3.1). In this second fitting procedure, we perform a parallel tempering MCMC, which essentially performs multiple MCMC at different ‘temperatures’ and improves sampling of the thermodynamic parameter space (see further detail in Appendix A). One example energy matrix is shown in Fig. 3.3B, which is now reported in units of k_BT. Note that not all parameters of Eq. (3.1) can be determined from this inference approach (Kinney et al., 2010), though it is sufficient to determine the scaling parameters needed to calculate binding by LacI in absolute energy units.

In Fig. 3.4A we summarize the energy matrices for LacI for the strains with the highest repressor copy number, $R = 1740$. Here we plot the energy matrices that determine $\varepsilon_{R,,\mathrm{mat}}$, which allows us to compare the sequence specificity of each matrix. We find that the energy matrices from the O1 and O2 binding site data are quite similar, while the matrix from the O3 binding site data to be somewhat less consistent (Pearson correlation coefficients: $r = 0.91$ between O1 and O2; $r = 0.69$ between O1 and O3).

The entire set of LacI matrices generated from the Sort-Seq experiments are summarized in Fig. 3.4B. Here we calculate the correlation of each matrix (relative to the $R = 1740$, O1 energy matrix), and overlay these values on a plot of the expected fold-change as a function repressor copy number. Fold-change here refers to the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor, and while not directly measured, provides a useful reference for the extent of repression expected by LacI in each Sort-Seq experiment. We find each matrix from the O1 and O2 binding site data sets to be quite consistent. Notably however, those from the O3 binding site data sets are less similar. Given the low repression expected by LacI in strains with an O3 binding site, this result may be due to the Sort-Seq data containing less information content associated with binding of LacI. Though it is also useful to note that we also find some correlation among matrices based on the same binding site library ($r > 0.94$ across O1 matrices; $r > 0.91$ across O2 matrices, and $r > 0.80$ across O3 matrices).

Sort-Seq energy matrices provide accurate prediction of LacI binding energy

In order to test the binding energy predictions that are provided by our LacI energy matrices, we constructed a set of simple repression constructs where the O1 binding site was mutated at 1, 2, or 3 positions (summarized in Table 3.2). These were placed into our *E. coli* strains containing different LacI copy numbers ($R = 22 \pm 4$, 60 ± 20, 124 ± 30, 260 ± 40, 1220 ± 160, and 1740 ± 340, where errors denote standard deviation of at least three replicates as measured in Ref. (Garcia and Phillips, 2011)), and measured expression as a function of transcription factor concentration for each of the designed LacI binding sites.

Here we find it more convenient to use the fold-change in gene expression instead of expression alone. As we noted earlier, fold-change is defined as the ratio of gene expression in the presence of repressor relative to expression in the absence of

LacI binding site sequence	prediction (k_BT)
1 bp mutations	
AATTGTGAGCGGAGAACAATT	-11.9
AATTGTGAGCGCATAACAATT	* -15.6
AATTGTGAGCGGATCACAATT	-15.2
AATTGTGAGCGGAAAACAATT	* -11.5
AATTGCGAGCGGATAACAATT	-10.0
AATTGTGAGGGGATAACAATT	-12.2
AATTGTGAGCGGATATCAATT	* -12.8
AATTGTGAGCAGATAACAATT	* -9.8
AATTGTGAGAGGATAACAATT	-6.3
2 bp mutations	
AAATGTGAGCGGGTAACAATT	-14.6
AATTGTGAGCGGGTAACAACT	-13.6
AAATGTGAGCGGATAACAACT	-13.3
AATTGTGAGCGAGTAACAATT	-14.0
ATTTGTGAGCGGAGAACAATT	-11.9
CATTGTGAGCGCATAACAATT	-15.3
AATTGTGAGCGGAACACAATT	-11.7
AATTGTGAGCGGAATACAATT	-9.6
AATTGCGAGCGGATAACAAAT	-10.5
AATTGTGAGGGGATAACAATC	-14.1
3 bp mutations	
AAATGTGAGCGAGTAACAATT	-13.6
AATTGTGAGCGAATAACAACC	-14.6
AAATGTGAGCGAATAACAACT	-12.2
AATTGTGAGCGAGTAACAACT	-12.6
ATTTGTGAGCGAAGAACAATT	-10.8
AATTGTGAGCGGAACACAATG	-12.3
AATTGTGAGCGGGATACAATT	-9.5
AATTGCGAGCGGATAACAAAG	-11.2
AATTGTGAGGGTATAACAATC	-13.5

Table 3.1: **Summary of designed O1 mutant binding site sequences and the predicted LacI binding energies for each.** The binding energies represent the average across the different energy matrices inferred from the Sort-Seq data using the O1 binding site. Sequences noted with an asterisk were also used to test whether the sequence-energy mapping could be used to design different induction responses in Fig. 3.6.

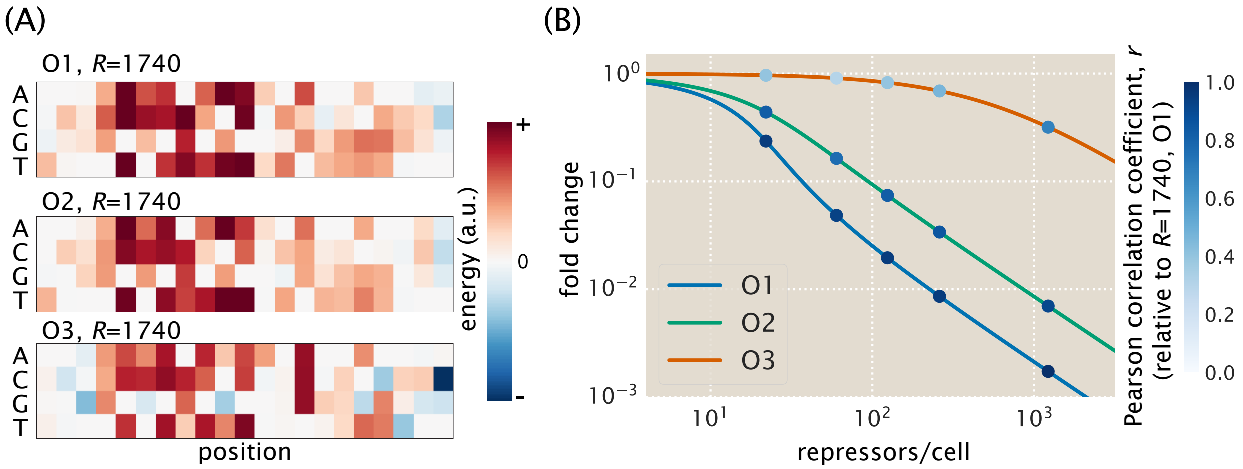


Figure 3.4: **Energy matrices for the natural *lac* operators from Sort-Seq data.** (A) Energy matrix models are shown for the LacI binding site from experiments performed with O1, O2, and O3 libraries, and in strains with $R = 1740$ repressor copies per cell. All energy matrices are plotted such that an O1 binding site sequence will have zero energy. (B) Pearson correlation coefficients were calculated relative to the energy matrix found using the O1 library in a strain with $R = 1740$ repressor copies per cell. Each marker represents the correlation coefficient for a matrix from a separate Sort-Seq experiment. Data is overlayed on a plot of expected expression fold-change (calculated assuming 10 plasmid copies per cell (Weinert et al., 2014)) to provide a reference for the expected influence of LacI on expression under each particular Sort-Seq experiment.

repressor (i.e. constitutive expression), namely,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}, \tag{3.5}$$

where p_{bound} was defined in Eq. (3.1). In Chapter 2 we found that under the weak promoter approximation, this reduces to the form

$$\text{fold-change} \approx \left(1 + p_A(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}\right)^{-1}. \tag{3.6}$$

For now we are only concerned with the case where no inducer is present in the growth media (i.e. where $p_A(c) = 1$). Using our LacI energy matrix to predict $\Delta\varepsilon_R$, we find that we can make parameter-free predictions of fold-change for each LacI binding site sequence as a function of the repressor copy number associated with each of our *E. coli* strains.

We use flow cytometry to measure fluorescence of each strain, using an experimental pipeline that is identical to the approach used in Chapter 2. Briefly, cells were grown to exponential phase in M9 minimal media with 0.5% glucose. Following a 1:10 dilution in fresh media, the fluorescence was measured by flow cytometry and automatically gated to include only single-cell measurements. We then calculated fold-change from the mean fluorescence level of each strain relative to a strain where LacI has been deleted. In Fig. 3.5A we show fold-change measurements for a subset of the 1 bp, 2 bp, and 3 bp mutants, overlaid with the parameter-free curves using our LacI energy matrix predictions of $\Delta\varepsilon_R$.

Since we performed fold-change measurements for each O1 mutant at several repressor copy numbers, it was also possible to use these measurements to directly estimate the LacI binding energies for each binding site sequence. In Fig. 3.5B we compare the measured binding energies against those predicted by our LacI energy matrix. For single base pair mutations most predictions are accurate to within 1 k_BT, with many predictions differing from the measured values by less than 0.5 k_BT. Though we do note that one of the sequences whose predicted binding was -6.3 k_BT, was instead found to have a binding energy of about -10.5 k_BT. Predictions are less accurate for 2 bp or 3 bp mutations, although the majority of these predictions are still within 1.5 k_BT of the measured value.

While not completely unexpected, we find that the quality of matrix predictions degrades as we deviate farther from the O1 wild-type sequence that was used in the Sort-Seq experiment to generate the energy matrix. To evaluate predictions for a broader range of deviations from the energy matrix, we made predictions from both the O1 energy matrix and the energy matrix for O2, which has five mutations relative to O1. This allowed us to access predictions for binding sites that are mutated by several base pairs relative to the matrix. As shown in Fig. 3.5C, we find that predictions remain relatively accurate for mutants that differ by up to 4 bp relative to the wild-type sequence, with median deviations of $\sim$ 1.5 k_BT or less from the measured binding energy. For a system with $R = 60$ LacI dimers, this mismatch in binding energy would imply that a prediction of fold-change would be inaccurate by approximately 0.10 - 0.35 (depending on the mutant binding site). By contrast, the median mismatch of $\sim 0.5 k_BT$ shown for 1 bp mutants implies that our fold-change predictions are only inaccurate by 0.04 - 0.12, highlighting that predicted binding energies for single-point mutations will in general be more reliable.

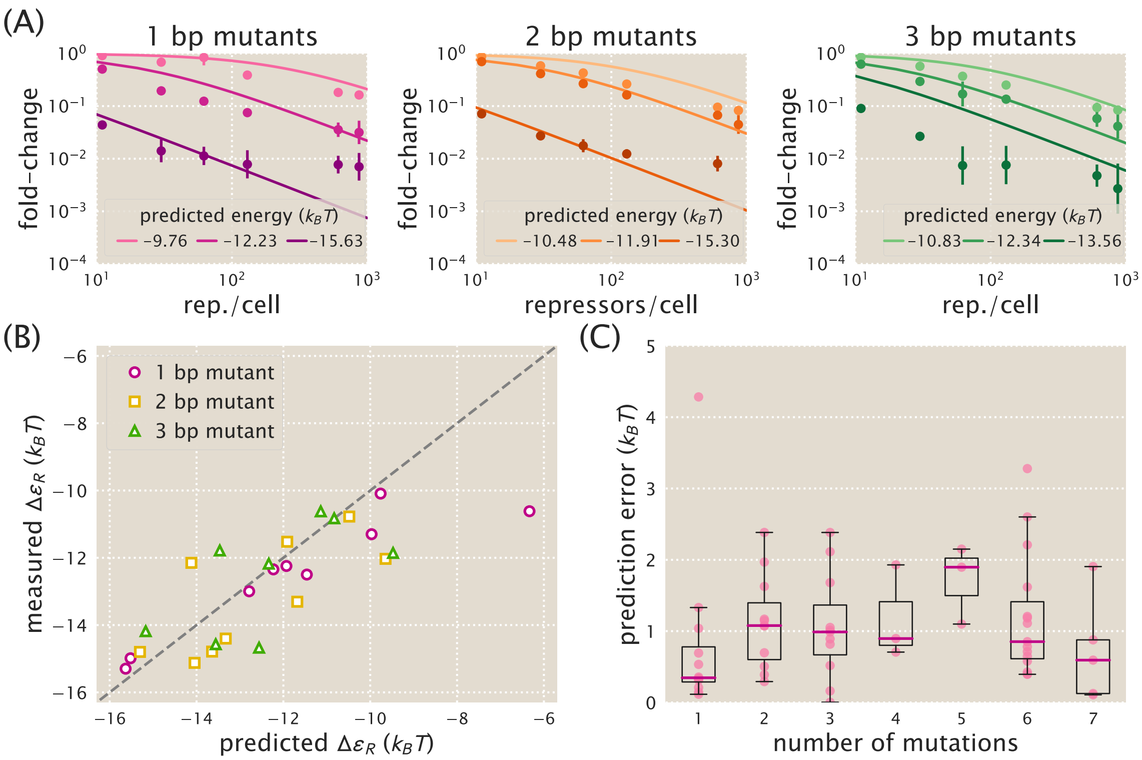


Figure 3.5: **Fold-change data reflects expected values from predicted fold-change curves.** (A) Fold-change data were obtained for each of the mutant operators by measuring their respective fluorescence levels at multiple LacI copy numbers. The solid lines in each plot represent the expected fold-change curve for each binding energy as predicted by the O1 energy matrix. A subset of data sets are shown for the 1 bp (left), 2 bp (middle), and 3 bp (right) mutants. Approximately 30 mutants were measured in total, with five replicate measurements performed for each strain. Predicted energies are based on the average predictions from the different O1 energy matrices. (B) The measured binding energy values $\Delta\varepsilon_R$ (y axis) are plotted against binding energy values predicted from an energy matrix derived from the O1 operator (x axis). While the quality of the binding energy predictions does appear to degrade as the number of mutations relative to O1 is increased, the O1 energy matrix is still able to approximately predict the measured values. (C) Binding energies for each mutant were predicted using both the O1 and O2 energy matrices and compared against measured binding energy values. The amount of error associated with each of these predictions is plotted here against the number of mutations relative to the wild-type sequence whose energy matrix was used to make the prediction. For sequences with 4 or fewer mutations, the median prediction error is consistently lower than 1.5 k_BT.

Regulatory sequence can be used to tune the simple repression induction curve.

A common desire in synthetic biology is to design regulatory circuits that provide specific input-output characteristics. An often used strategy is to try many ‘parts’ until the desired expression response is obtained (Kosuri et al., 2013). Previous work however has also shown that rather than rely on such trial and error approaches, it also is possible to use thermodynamic models of regulation to accurately predict specific input-output characteristics (Bintu et al., 2005; Garcia and Phillips, 2011). Such models also provide non-obvious insight into what characteristics can be designed. Indeed, in Chapter 2 we showed how repressor copy number and repressor-DNA binding energy could be used to tune the induction response. Above we have also shown how we can use regulatory sequence, through the design of specific LacI binding site sequences, to further control the level of gene expression.

As a next step, we were interested in whether our sequence-energy mapping could also be used to precisely design different induction responses. In Chapter 2 we found that the allosteric response of LacI to the inducer IPTG was well described through the Monod-Wyman-Changeux (MWC) model, with LacI in equilibrium between two conformations, termed the inactive and active states. In our formulation of fold-change as a function of inducer concentration (Eq. (3.1)), $p_A(c)$ is well described by

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^2}{\left(1 + \frac{c}{K_A}\right)^2 + e^{-\beta \Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^2}, \tag{3.7}$$

where c is the concentration of inducer, K_A and K_I are the dissociation constants of the inducer and repressor when the repressor is in its active or inactive state, respectively, and $\Delta\varepsilon_{AI}$ is the difference in free energy between the repressor's active and inactive states. In Chapter 2 we found that for induction of LacI by IPTG, $K_A = 139\ \mu\text{M}$, $K_I = 0.53\ \mu\text{M}$, and $\Delta\varepsilon_{AI} = 4.5\ k_BT$.

We note that an induction response can be described by four key phenotypic parameters. The leakiness is the minimum fold-change when no inducer is present, given by fold-change($c \to 0$). The saturation is the maximum fold-change when inducer is present at saturating concentrations, given by fold-change($c \to \infty$). The dynamic range is the difference between the saturation and leakiness, and represents the magnitude of the induction response. Figure 3.6A shows how these three phenotypic parameters vary with $\Delta\varepsilon_R$ given the values of K_A, K_I, and $\Delta\varepsilon_{AI}$ listed above and the repressor copy number $R = 260$. Lastly, the $[EC_{50}]$ of an induction response denotes the inducer concentration required to generate a response that is halfway between the minimum and maximum value.

We can see that there are inherent trade-offs between phenotypic parameter values. For instance, in this particular system one cannot tune $\Delta\varepsilon_R$ to obtain a small dynamic range (e.g. a dynamic range of 0.1) while also having an intermediate leakiness value (e.g. a leakiness of 0.4). Rather, one must design an induction response by choosing from the available phenotypes, or else alter the system by tuning additional parameters such as K_A and K_I, which requires mutating the protein itself.

To show how energy matrices can be used to design specific induction responses, we used the phenotypic trade-offs shown in Figure 3.6A to choose four different values of $\Delta\varepsilon_R$ that would provide distinct outputs. These values were $\Delta\varepsilon_R \approx -16\ k_BT$, which would provide a minimal leakiness level but not reach full saturation; $\Delta\varepsilon_R \approx -13\ k_BT$, which would maximize dynamic range; $\Delta\varepsilon_R \approx -11.5\ k_BT$, which would maximize saturation but have an intermediate dynamic range; and $\Delta\varepsilon_R \approx -10\ k_BT$, which is close to the threshold between specific binding and nonspecific binding, and would provide a narrow dynamic range. Four of the single base-pair mutants designed in the previous section had predicted binding energies that matched these approximate values (noted with an asterisk in Table 3.2). Induction responses for each of these mutants were deterined by growing cultures in the presence of varying IPTG concentrations and measuring the fold-change at each concentration. Figure 3.6B shows how the induction data compare against fold-change curves plotted using $\Delta\varepsilon_R$ values predicted from the energy matrix, and fold-change as defined in Eq. (3.1) and Eq. (3.7). The measured induction responses were found to match the theoretical predictions quite well, though for the sequence with a predicted energy of $\Delta\varepsilon_R \approx -11.5\ k_BT$, we find that the $[EC_{50}]$ is shifted toward a higher IPTG concentration. This is at least in part due to a higher measured binding energy (-12.5 k_BT instead of -11.5 k_BT) than predicted by our LacI energy matrix.

Sort-Seq can be used to probe both the DNA and amino acid interactions

So far we have examined how energy matrices provide us with a quantitative mapping between DNA sequence and binding energy, and how this can allow us to predict specific input-output characteristics. In this final section we show how we can also use energy matrices to investigate the effects of amino acid mutations on a transcription factor's sequence specificity. Specifically, we make individual amino acids changes to the repressor's DNA-binding domain and through additional Sort-Seq experiments, observe how those mutations modify the LacI energy matrix. This approach in particular makes it possible to determine how changing the amino acid composition of the DNA-binding domain alters DNA sequence preference.

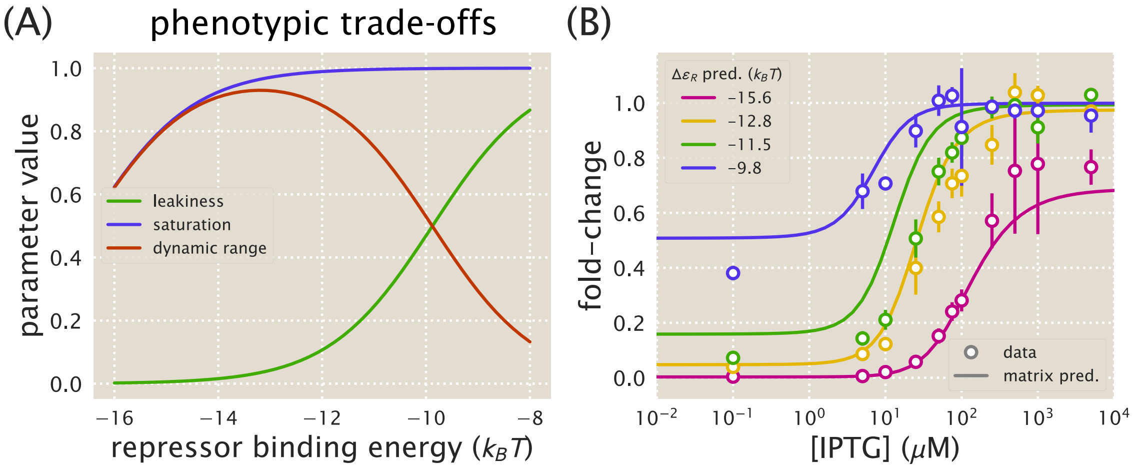


Figure 3.6: **Energy matrix predictions can be used to design precise phenotypic responses** (A) Phenotypic parameters (leakiness, saturation, and dynamic range) exhibit trade-offs as $\Delta\varepsilon_R$ is varied. Maximizing saturation or minimizing leakiness can only be achieved by reducing the dynamic range below its maximum. (B) Operators with different values of $\Delta\varepsilon_R$ were chosen to have varying induction responses based on the phenotypic trade-offs shown in Part A. The induction responses predicted based on energy matrix predictions (solid lines) generally agree well with IPTG induction data obtained for each of the binding sites in a background strain with $R = 260$.

We performed Sort-Seq using strains containing one of three LacI mutants, Y20I, Q21A, or Q21M, where the first letter indicates the wild-type amino acid, the number indicates the amino acid position, and the last letter indicates the identity of the mutated amino acid. These mutants have previously been found to alter LacI-DNA binding properties without entirely disrupting the repressor's ability to bind DNA (Milk et al., 2010; Daber et al., 2011). We note that we use a slightly different version of LacI from the one used in Refs. (Milk et al., 2010; Daber et al., 2011), so that the residue numbers in our version of LacI are shifted upward by 3 bp.

Sequence logos for each LacI mutant are shown in Figure 3.7, along with the wild-type sequence logo for comparison. As with the wild-type repressor, for each of the mutant repressors we find that the left half-site of the sequence logo has a higher information content. For both Y20I and Q21M, the same sequence is preferred in the left half-site as the wild-type sequence logo. This contrasts with the results from Ref. (Milk et al., 2010), in which it was found that Y20I prefers an adenine at sequence position 7, rather than the guanine preferred at this position by the wild-type repressor. As in Ref. (Milk et al., 2010), we find that an adenine is preferred at sequence position 8 for the Q21A mutant.

Some more subtle features can be observed when comparing the right half-sites. Within the right half-site, the most important base positions consistently appear to be 12, 13, 16 and 17. All mutants, along with the wild-type repressor, prefer cytosine and adenine at sequence positions 16 and 17. The wild-type, Q21A, and Q21M mutants all prefer an adenine and a tyrosine at positions 12 and 13, while the Y20I mutant prefers tyrosine and cytosine. For all mutants, the preferred bases at positions 16 and 17 are symmetrical to the corresponding bases in the left half-site (positions 4 and 5). By contrast, position 12 is consistently not symmetrical to position 8 in the right half-site, and position 13 for Y20I is not symmetrical to position 7 in the right half-site. Thus we see that the *lac* repressor's notable preference for a pseudo-symmetric binding site is preserved in each of the mutants we tested.

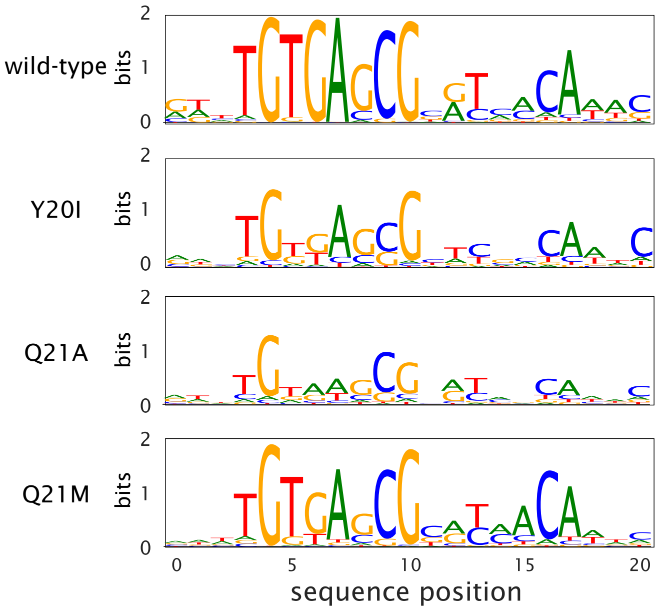


Figure 3.7: **Point mutations to LacI DNA-binding domain cause subtle changes to sequence specificity.** Mutations were made to residues 20 and 21 of LacI, both of which lie within the DNA-binding domain. The mutations Y20I and Q21A weaken the repressor-operator binding energy, while the mutation Q21M strengthens the binding energy. Y20I exhibits minor changes to specificity in low-information regions of the binding site, and Q21A experiences a change to specificity within a high-information region of the binding site. Specifically, Q21A prefers A at operator position 7 while the wild-type repressor prefers G at this position.

3.3 Discussion

We have shown how the massively parallel reporter assay, Sort-Seq (Kinney et al., 2010), can be used to generate a mapping between regulatory sequence and transcription factor binding energy using linear energy matrix models. By using a simple thermodynamic model, we find that this mapping provides further control over the input-output gene expression characteristics through finer control of the LacI DNA-binding energy. This work follows from a previous effort in our group to test the validity of such energy matrix models that describe binding of RNAP (Brewster et al., 2012). Here we explore whether the approach can be applied more broadly to other regulatory components. Specifically, we first used Sort-Seq to map sequence to binding energy by inferring energy matrices for the repressor LacI. We perform this work in the context of a simple repression architecture, which represents a widespread bacterial regulatory architecture (Rydenfelt et al., 2014) that is commonly employed in synthetic biology (Voigt and Brophy, 2014; Khalil and Collins, 2010; Purnick and Weiss, 2009). We then demonstrate the validity of our model by designing roughly 30 mutant LacI binding site sequences, where we then demonstrate control over fold-change in gene expression, and show how such regulatory sequences can be used to optimize the inducible response of LacI by IPTG. Lastly, we show how Sort-Seq can also be used to probe the amino acid-DNA interactions. Here we perform Sort-Seq in several *E. coli* strains containing mutant LacI proteins and find only minor perturbations to the LacI sequence specificity following single amino-acid changes to the LacI DNA-binding domain.

While we focused on the regulatory component of LacI, we believe it will be possible to use regulatory sequence to predict gene expression more broadly across the bacterial genome and to other synthetic regulatory constructs, assuming that a thermodynamic model is in hand that can adequately describe the regulatory architecture. It is clear from our work that although we could accurately design regulatory sequences with a predictable fold-change, there were a variety of instances with notable discrepancies between the measured and predicted fold-change. This may suggest the need to consider more complex models than our linear energy matrices that incorporate non-additive contributions (Benos et al., 2002). Deep-learning algorithms may provide an alternative approach to model the DNA-protein interactions (Zeng et al., 2016; Zhou et al., 2017). Another consideration is that while Sort-Seq was performed on plasmids, our designed promoters were integrated on the chromosome, and aspects related to chromosomal context and DNA compaction are not considered in our model (Kuhlman and Cox, 2012). Landing pad technologies for chromosomal integration (Kuhlman and Cox, 2010; Zhang et al., 2016; St-Pierre et al., 2013) could enable massively parallel reporter assays to be performed on chromosomes instead of on plasmids, and enable more accurate descriptions of chromosomally integrated promoters. Nonetheless, even where predicted fold-change did not match the observed fold-change, we still find a clear correlation between the predicted and measured LacI binding energies, and we have shown how regulatory sequence and a thermodynamic model can be used to guide our design of optimized inducible regulatory systems.

3.4 Methods

Sort-Seq libraries

To generate promoter libraries for Sort-Seq, mutagenized oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA). These consisted of single-stranded DNA containing the *lacUV5* promoter and LacI operator plus 15 bp on each end for PCR amplification. Either both, the *lacUV5* promoter and LacI binding site, or only the LacI binding site was mutated with a ten percent mutation rate per nucleotide. These oligonucleotides were amplified by PCR and inserted back into the pZS25-operator-YFP construct using Gibson Assembly. This plasmid is maintained in low copy (5-10 copies per cell) with the SC101 origin of replication (Lutz and Bujard, 1997). To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in SOC media, cells were diluted with 50 mL LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies and were assessed by plating 100 μL of cells diluted 1:10^4 onto an LB plate containing kanamycin.

DNA Constructs for fold-change measurements

Simple repression motifs used in Sort-Seq experiments and fold-change measurements were adapted from those in Garcia *et al.*(Garcia and Phillips, 2011). Briefly, the LacI operator (O1, O2, or O3) and YFP reporter gene were cloned into a pZS25 background directly downstream of a *lacUV5* promoter, driving expression of the YFP gene when the operator is not bound by LacI. This plasmid contains a kanamycin resistance gene for selection. Mutant LacI operator constructs were generated by PCR amplification of the *lacUV5* O1-YFP plasmid using primers containing the point mutations as well as sufficient overlap for re-circularizing the amplified DNA by Gibson Assembly.

A second construct was generated to provide expression of the *lacI* gene. Here, *lacI* was cloned into a pZS3*1 background that provides constitutive expression of LacI from a $P_{\mathrm{LtetO-1}}$ promoter (Lutz and Bujard, 1997). This plasmid contains a chloramphenicol resistance gene for selection. To produce strains with different mean copy number of LacI that differ from the wild-type value of about 11 tetramers per cell, the ribosomal binding site for the *lacI* gene was mutated as described in (Salis et al., 2009) using site-directed mutagenesis (Quickchange II; Stratagene, San Diego, CA) and further detailed in (Garcia and Phillips, 2011).

Bacterial Strains

E. coli strains used in this work were derived from K12 MG1655. To generate strains with different LacI copy number, the *lacI* constructs were integrated into a strain that additionally has the entire *lacI* and *lacZYA* operons removed from the chromosome. These were integrated at the *ybcN* chromosomal location. This resulted in strains containing mean LacI copy numbers of $R = 30, 62, 130, 610,$ and 870, which were measured previously by quantitative western blots (Garcia and Phillips, 2011).

For Sort-Seq experiments, plasmid promoter libraries were constructed as described below and then transformed into the strains with different LacI copy number. For fold-change measurements, only the native O1 operator and associated mutants were considered. These simple repression constructs were chromosomally integrated at the *galK* chromosomal location. Generation of the final strains containing a simple repression motif and a specific LacI copy number were achieved by P1 transduction. For each LacI titration experiment, we also generated a strain where the entire *lacI* and *lacZYA* operons were removed, but with only the operator-YFP construct integrated. This provided us with a fluorescence expression measurement corresponding to $R = 0$, which is necessary for calculation of fold-change.

Sort-Seq fluorescence sorting

For each Sort-Seq experiment, cells were grown to saturation in lysogeny broth (LB) and then diluted 1:10,000 into minimal M9 + 0.5% glucose for overnight growth. Once these cultures reached an OD 0.2-0.3 the cells were washed three times with PBS by centrifugation at 4000 rpm for 10 minutes and at 4°C. They were then diluted two-fold with PBS to reach an approximate OD of 0.1-0.15. These cells were then passed through a 40 μm cell strainer to eliminate any large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used to obtain initial fluorescence histograms of 500,000 events per library, which were used to set four binning gates that each covered 15% of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were regrown overnight in 10 mL of LB media, under kanamycin selection.

Sort-Seq sequencing and data analysis

Overnight cultures from each sorted bin were miniprepped (Qiagen, Germany), and PCR was used to amplify the mutated region from each plasmid for Illumina sequencing. The primers contained Illumina adapter sequences as well as barcode sequences that enable pooling of the sorted samples. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech or NGX Bio (San Fransisco, CA). Single-end 100bp or paired-end 150bp flow cells were used, with about 500,000 sequences collected per library bin. After performing a quality check and filtering for sequences whose PHRED score was greater than 20 for each base pair, the total number of useful reads per bin was approximately 300,000 to 500,000 per million reads requested. Energy weight matrices for binding by LacI and RNAP were inferred using Bayesian parameter estimation with a error-model-averaged likelihood as previously described (Kinney et al., 2010; Kinney and Atwal, 2014).

Fold-change measurements by flow cytometry

Fold-change measurements were collected as previously described [cite Induction paper] on a MACSquant Analyzer 10 Flow Cytometer (Miltenyi Biotec, Germany). Briefly, YFP fluorescence measurements were collected using 488nm laser excitation, with a 525/50 nm emission filter. Settings in the instrument panel for the laser were as follows: trigger on FSC (linear, 423V), SSC (linear, 537 V), and B1 laser (hlog, 790V). Before each experiment the MACSquant was calibrated using MACSQuant Calibration Beads (Miltenyi Biotec, CAT NO. 130-093-607). Following growth of cells to OD 0.2-0.3, they were diluted ten fold in ice-cold minimal M9 + 0.5% glucose. Cells were then automatically sampled from a 96-well plate kept at approximately 4° - 10°C using MACS Chill 96 Rack (Miltenyi Biotec, CAT NO. 130-094-459) at a flow rate of 2,000 - 6,000 measurements per second.

The fold-change in gene expression was calculated by taking the ratio of the mean YFP expression of the population of cells in the presence of LacI repressor to that in the absence of LacI repressor. Since the measured fluorescence intensity of each cell also includes autofluorescence which is present even in the absence of YFP, we account for this background by computing the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{3.8}$$

where $\langle I_{R>0} \rangle$ is the average cell YFP intensity in the presence of repressor, $\langle I_{R=0} \rangle$ is the average cell YFP intensity in the absence of repressor, and $\langle I_{\text{auto}} \rangle$ is the average cell autofluorescence intensity.

Data curation

All data was collected, stored, and preserved using the Git version control software in combination with off-site storage and hosting website GitHub.

Acknowledgements

Access to the Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer was graciously provided by the Pamela Björkman lab at Caltech. We thank David Tirrell, Bradley Silverman, and Seth Lieblich for access and training for use of their Beckman Coulter MoFlo XDP cell sorter.

3.5 Supplemental Information: Summary of designed O1 binding site mutant results.

Identifier	LacI binding sequence	O1 matrix prediction	O2 matrix prediction	Measured energy
mut005	AATTGTGAGCGGAGAACAATT	-11.929881	-13.428262	-12.243772
mut007	AATTGTGAGCGCATAACAATT	-15.633221	-14.197103	-15.296422
mut008	AATTGTGAGCGGATCACAATT	-15.520049	-14.133914	-14.986353
mut009	AATTGTGAGCGGAAAACAATT	-11.459789	-12.924778	-12.498838
mut010	AATTGCGAGCGGATAACAATT	-9.968247	-11.878477	-11.299124
mut011	AATTGTGAGGGGATAACAATT	-12.230209	-13.455658	-12.344994
mut012	AATTGTGAGCGGATATCAATT	-12.787483	-13.642761	-12.996080
mut013	AATTGTGAGCAGATAACAATT	-9.760610	-12.692912	-10.091807
mut014	AATTGTGAGAGGATAACAATT	-6.331624	-8.997448	-10.615486
mut102	AATTGTGAGCGGGTAACAACT	-13.641728	-13.896787	-14.788271
mut103	AAATGTGAGCGGATAACAACT	-13.328345	-13.584199	-14.401196
mut104	AATTGTGAGCGAGTAACAATT	-14.044856	-14.070952	-15.122752
mut105	ATTTGTGAGCGGAGAACAATT	-11.911801	-13.428375	-11.523189
mut107	CATTGTGAGCGCATAACAATT	-15.302753	-14.016493	-14.797621
mut108	AATTGTGAGCGGAACACAATT	-11.679837	-12.712688	-13.305983
mut109	AATTGTGAGCGGAATACAATT	-9.647010	-12.138189	-12.030819
mut110	AATTGCGAGCGGATAACAAAT	-10.481933	-11.487112	-10.774666

mut111	AATTGTGAGGGGATAACAATC	-14.118290	-14.046511	-12.149832
mut201	AAATGTGAGCGAGTAACAATT	-13.558126	-13.874477	-14.571139
mut204	AATTGTGAGCGAGTAACAACT	-12.559931	-13.505622	-14.673368
mut205	ATTTGTGAGCGAAGAACAATT	-10.830003	-13.037210	-10.827536
mut207	CATTGTGAGCGCATAACATTT	-15.171401	-14.057285	-14.182531
mut208	AATTGTGAGCGGAACACAATG	-12.337016	-13.053090	-12.175545
mut209	AATTGTGAGCGGGATACAATT	-9.473663	-12.254301	-11.857128
mut210	AATTGCGAGCGGATAACAAAG	-11.139112	-11.827513	-10.621422
mut211	AATTGTGAGGGTATAACAATC	-13.464516	-13.934262	-11.784251

Table 3.2: **Summary of O1 mutant binding site results.** The table summarizes the results of the binding site mutants. Matrix predictions are based on the average predicted binding energy from the different matrices available, based on either the set of O1 based matrices, or O2 matrices (i.e. Sort-Seq experiments where the promoter contained either an O1 or O2 sequence). The measured energy column notes the value obtained from fitting the measured fold-change data to the fold-change theory and inferring the binding energy. Note that the predicted energies may change slightly as more analysis is applied to the Sort-Seq data and matrix inference approach.

References

Ackers, G. K., Johnson, A. D., and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences* 79.4, pp. 1129–33.

Alekshun, M. N. and Levy, S. B. (1997). Regulation of chromosomally mediated multiple antibiotic resistance: the *mar* regulon. *Journal of Molecular Biology* 41.10, pp. 2067–2075.

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein–DNA interactions: how good an approximation is it. *Nucleic Acids Research* 30.20, pp. 4442–4451.

Berg, O. G. and Hippel, P. H. von (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* 193.4, pp. 723–743.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Iii, P. W. E., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology* 24.11, pp. 1429–1435.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development* 15.2, pp. 116–124.

Brewster, R. C., Jones, D. L., and Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology* 8.12.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–41.

Christensen, R. G., Gupta, A., Zuo, Z., Schriefer, L. A., Wolfe, S. A., and Stormo, G. D. (2011). A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. 39.12, pp. 1–9.

Daber, R., Sochor, M. A., and Lewis, M. (2011). Thermodynamic analysis of mutant lac repressors. *Journal of Molecular Biology* 409.1, pp. 76–87.

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research* 13.11, 2381–2390.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research* 25.9, pp. 1268–1280.

Fields, D. S., He, Y.-y., Al-uzri, A. Y., and Stormo, G. D. (1997). Quantitative Specificity of the Mnt Repressor. *Journal of Experimental Zoology* 271, pp. 178–194.

Gama-Castro, S. et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* 44.D1, pp. D133–D143.

Garcia, H. G. and Phillips, R. (2011). Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–8.

Garcia, H. G., Sanchez, A., Boedicker, J. Q., Osborne, M., Gelles, J., Kondev, J., and Phillips, R. (2012). Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Reports* 2.1, pp. 150–161.

Gerdes, K., Christensen, S. K., and Løbner-Oleson, A. (2005). Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology* 2.5, pp. 371–382.

Ireland, W. T. and Kinney, J. B. (2016). MPAthic: quantitative modeling of sequence-function relationships for massively parallel assays. *bioRxiv*, p. 054676.

Jolma, A. et al. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell* 152.1-2, pp. 327–339.

Jung, C. et al. (2017). Massively Parallel Biophysical Analysis of CRISPR- Cas Complexes on Next Generation Sequencing Chips. *Cell* 170.1, 35–47.e13.

Khalil, A. S. and Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics* 11.5, pp. 367–379.

Kinney, J. B. and Atwal, G. S. (2014). Parametric Inference in the Large Data Limit Using Maximally Informative Models. *Neural Computation* 26.4, pp. 637–653.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* 107.20, 9158–9163.

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences* 110.34, pp. 14024–14029.

Kuhlman, T. E. and Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research* 38.6, e92–e92.

Kuhlman, T. E. and Cox, E. C. (2012). Gene location and DNA density determine transcription factor distributions in Escherichia coli. *Molecular Systems Biology* 8.1, p. 610.

Levo, M. et al. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* 25.7, pp. 1018–1029.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Molecular Cell* 65.4, 604–617.e6.

Levy, L. et al. (2017). A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial sigma54 Promoters. *Cell Reports* 21.3, pp. 845–858.

Lutz, R. and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research* 25.6, pp. 1203–10.

Maerkl, S. J. and Quake, S. R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315.January, pp. 233–238.

Melnikov, A. et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* 30.3, pp. 271–277.

Milk, L., Daber, R., and Lewis, M. (2010). Functional rules for lac repressor-operator associations and implications for protein-DNA interactions. *Protein Science* 19.6, pp. 1162–1172.

Minchin, S. D. and Busby, S. J. (2009). Analysis of mechanisms of activation and repression at bacterial promoters. *Methods* 47.1, pp. 6–12.

Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G. P., and Burge, C. B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology* 29.7, pp. 659–664.

Oehler, S., Eismann, E. R., Krämer, H., and Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal* 9.4, pp. 973–979.

Purnick, P. E. M. and Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* 10.6, pp. 410–422.

Rydenfelt, M., Garcia, H. G., Cox, R. S., and Phillips, R. (2014). The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS ONE* 9.12, pp. 1–31.

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27.10, pp. 946–950.

Shultzaberger, R. K., Maerkl, S. J., Kirsch, J. F., and Eisen, M. B. (2012). Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA – Binding Domain. *PLoS Genetics* 8.3.

St-Pierre, F., Cui, L., Priest, D. G., Endy, D., Dodd, I. B., and Shearwin, K. E. (2013). One-Step Cloning and Chromosomal Integration of DNA. *ACS Synthetic Biology* 2.9, pp. 537–541.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16.1, pp. 16–23.

Voigt, C. A. and Brophy, J. A. N. (2014). Principles of genetic circuit design. *Nature Methods* 11.5, pp. 508–520.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., and Kegel, W. K. (2014). Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters* 113.25, pp. 1–5.

Weirauch, M. T. et al. (2013). Evaluation of methods for modeling transcription-factor sequence specificity. *Nature Biotechnology* 31.2, pp. 126–134.

Wunderlich, Z. and Mirny, L. A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics* 25.10, pp. 434–440.

Xu, D. J. and Noyes, M. B. (2015). Understanding DNA-binding specificity by bacteria hybrid selection. *Briefings in Functional Genomics* 14.1, pp. 3–16.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* 32.12, pp. i121–i127.

Zhang, H., Susanto, T. T., Wan, Y., and Chen, S. L. (2016). Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences* 113.15, pp. 4182–4187.

Zhou, B., Pei, J., Lai, L., and Sun, T. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 18.1, p. 277.

Chapter 4

A SYSTEMATIC APPROACH FOR DISSECTING THE MOLECULAR MECHANISMS OF TRANSCRIPTIONAL REGULATION IN BACTERIA

A version of this chapter originally appeared as N. M. Belliveau, S. L.Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, R. Phillips (2017). A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. bioRxiv, 239335. http://doi.org/10.1101/239335. It is also in preparation for publication in a peer-reviewed journal.

4.1 Introduction

The sequencing revolution has left in its wake an enormous challenge: the rapidly expanding catalog of sequenced genomes is far outpacing a sequence-level understanding of how the genes in these genomes are regulated. This ignorance extends from viruses to bacteria to archaea to eukaryotes. Even in *E. coli*, the model organism in which transcriptional regulation is best understood, we still have no indication if or how more than half of the genes are regulated (Fig. 1; see also RegulonDB (Gama-Castro et al., 2016) or EcoCyc (Keseler et al., 2013)). In other model bacteria such as *Bacillus subtilis, Caulobacter crescentus, Vibrio harveyii*, or *Pseudomonas aeruginosa*, far fewer genes have established regulatory mechanisms (Münch et al., 2003; Cipriano et al., 2013; Kılıç et al., 2013).

New tools are needed for studying regulatory architecture in these and other bacteria. Although an arsenal of genetic and biochemical methods have been developed for dissecting promoter function at individual bacterial promoters (reviewed in Minchin and Busby, 2009), these methods are not readily parallelized. As a result, they will likely not lead to a comprehensive understanding of full regulatory genomes anytime soon. RNA sequencing, chromatin immunoprecipitation, and other high-throughput techniques are increasingly being used to study gene regulation in *E. coli* (Grainger et al., 2005; Bonocora and Wade, 2015; Latif et al., 2016; Zheng et al., 2004; Singh et al., 2014; Vvedenskaya et al., 2015b; Wade, 2015), but these methods are incapable of revealing either the nucleotide-resolution location of all functional transcription factor binding sites, or the way in which interactions between DNA-bound transcription factors and RNA polymerase modulate transcription.

In recent years a variety of massively parallel reporter assays have been developed for dissecting the functional architecture of transcriptional regulatory sequences in bacteria, yeast, and metazoans. These technologies have been used to infer biophysical models of well-studied loci, characterize synthetic promoters constructed from known binding sites, and search for new transcriptional regulatory sequences (Kinney et al., 2010; Melnikov et al., 2012; Kheradpour et al., 2013; Patwardhan et al., 2012; Sharon et al., 2012; Kosuri et al., 2013; Arnold et al., 2013; Maricque et al., 2017). CRISPR assays have also shown promise for identifying longer range enhancer-promoter interactions in mammalian cells (Fulco et al., 2016). However, no approach for using massively parallel reporter technologies to decipher the functional mechanisms of previously uncharacterized regulatory sequences has yet been established.

Here we describe a systematic and scalable approach for dissecting the functional architecture of previously uncharacterized bacterial promoters at nucleotide resolution using a combination of genetic, functional, and biochemical measurements. First, a massively parallel reporter assay (Sort-Seq; Kinney et al., 2010) is performed on a promoter in multiple growth conditions in order to identify functional transcription factor binding sites. DNA affinity chromatography and mass spectrometry (Mittler et al., 2009; Mirzaei et al., 2013) are then used to identify the regulatory proteins that recognize these sites. In this way one is able to identify both the functional transcription factor binding sites and cognate transcription factors in previously unstudied promoters. Subsequent massively parallel assays are then performed in gene-deletion strains to provide additional validation of the identified regulators. In many cases, the reporter data thus generated can further be used to infer quantitative models of transcriptional regulation.

In what follows, we first describe the application of this approach to four previously annotated promoters: *lacZYA*, *relBE*, *marRAB*, and *yebG*. This illustrates the overarching logic of our method and provides a benchmark for how well these methods work. We then describe this strategy applied to the previously uncharacterized promoters of *purT*, *xylE*, and *dgoRKADT*. These results demonstrate the ability to go from complete regulatory ignorance to an explicit quantitative model of a promoter's input-output behavior.

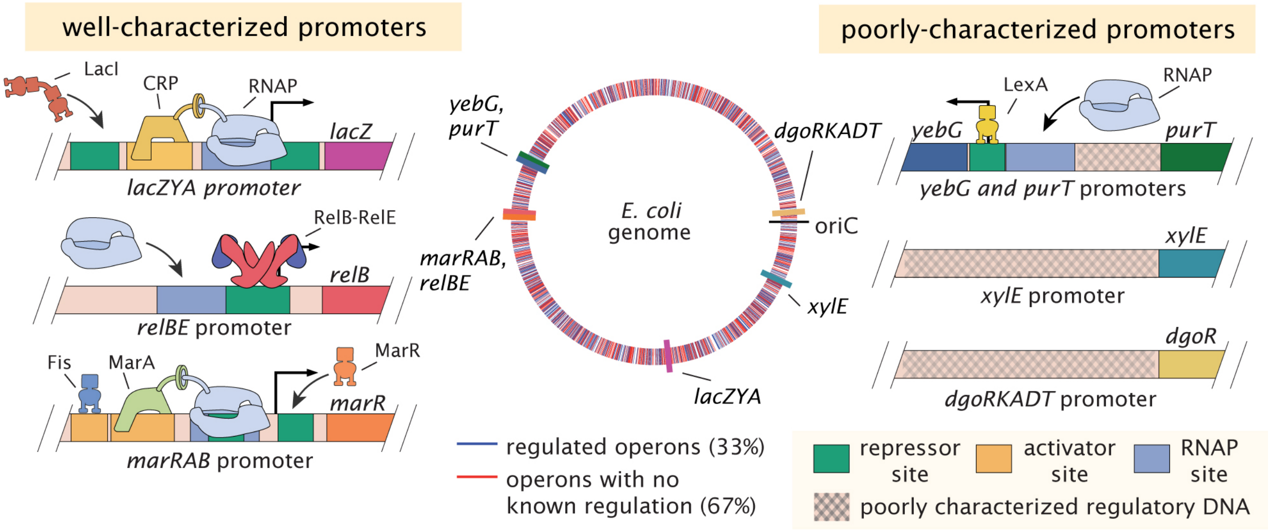


Figure 4.1: **Summary of transcriptional regulatory knowledge in *E. coli*.** left panel: Well-characterized promoters considered in this work. The schematics highlight the known regulatory architectures for the annotated promoters of *marRAB*, *relBE*, and *lacZYA*. The center plot identifies the genomic location of different operons in *E. coli*. Operons with annotated TF binding sites are shown in blue, while those lacking regulatory descriptions are shown in red (Gama-Castro et al., 2016). The genomic location of the promoters considered in this work are labeled. Right panel: promoters associated with the operons of *yebG* and the poorly-characterized operons *purT*, *xylE*, and *dgoRKADT*. The promoters of *yebG* and *purT* are oriented in opposite directions. Repressor binding sites are shown in green, activator binding sites in yellow, and RNA polymerase (RNAP) binding sites in blue. The poorly characterized regulatory DNA is noted by a hashed pattern.

4.2 Results

To dissect how a promoter is regulated, we begin by performing Sort-Seq (Kinney et al., 2010). As shown in Fig. 4.2A, Sort-Seq works by first generating a library of cells, each of which contains a mutated promoter that drives expression of GFP from a low copy plasmid (5-10 copies per cell; Lutz and Bujard, 1997) and provides a read-out of transcriptional state. We use fluorescence-activated cell sorting (FACS) to sort cells into multiple bins gated by their fluorescence level and then sequence the mutated plasmids from each bin. We found it sufficient to sort the libraries into four bins and generated data sets of about 0.5-2 million sequences across the sorted bins (Fig. 4.8A-D). Putative binding sites were identified by examining expression shift plots which show the average change in fluorescence when each position is mutated (Fig. 4.2B, top plot). Mutations to the DNA will disrupt binding of transcription factors (Mustonen et al., 2008), so regions with a positive shift are suggestive of binding by a repressor, while a negative shift suggests binding by an activator or

RNA polymerase (RNAP).

The identified binding sites are further interrogated by performing information-based modeling with the Sort-Seq data. Here we generate energy matrix models (Kinney et al., 2010; Ireland and Kinney, 2016) that describe the sequence-dependent energy of interaction of a transcription factor at the putative binding site. For each matrix, we use a convention that the wild-type sequence is set to have an energy of zero (see example energy matrix in Fig. 4.2B). Mutations that enhance binding are identified in blue, while mutations that weaken binding are identified in red. We also use the energy matrices to generate sequence logos (Berg and Hippel, 1987; Schneider and Stephens, 1990; Stormo, 2000) which provides a useful visualization of the sequence-specificity (see above matrix in Fig. 4.2B).

We next perform DNA affinity chromatography experiments using DNA oligonucleotides containing the binding sites identified by Sort-Seq. Here we apply a stable isotopic labeling of cell culture (SILAC) approach (Ong et al., 2002), which enables us to perform a second reference chromatography experiment that is simultaneously analyzed by mass spectrometry to identify the target transcription factor. As shown in Fig. 4.2C, we begin by preparing two cell lysates: one with cells supplemented with natural lysine and the other with a heavy isotopic form of lysine. We then perform chromatography using magnetic beads with the tethered oligonucleotides. Our reference experiment is performed identically, except that the binding site has been mutated away (and is generally performed using the light lysate). The abundance of each protein is determined by mass spectrometry and used to calculate protein enrichment ratios, with the target transcription factor expected to exhibit a ratio greater than one. Most proteins detected will exhibit a protein enrichment near one due to non-specific binding in both purifications.

The energy matrix models and results from each DNA affinity chromatography experiment provide insight into the identity of each regulatory factor and hypotheses about potential regulatory mechanisms. In some instances we are able to test these hypothesis further with additional information-based modeling of thermodynamic models on our Sort-Seq data. Finally, to confirm binding by an identified regulator we perform Sort-Seq experiments in gene deletion strains, which no longer show the positive or negative shift in expression along the binding site.

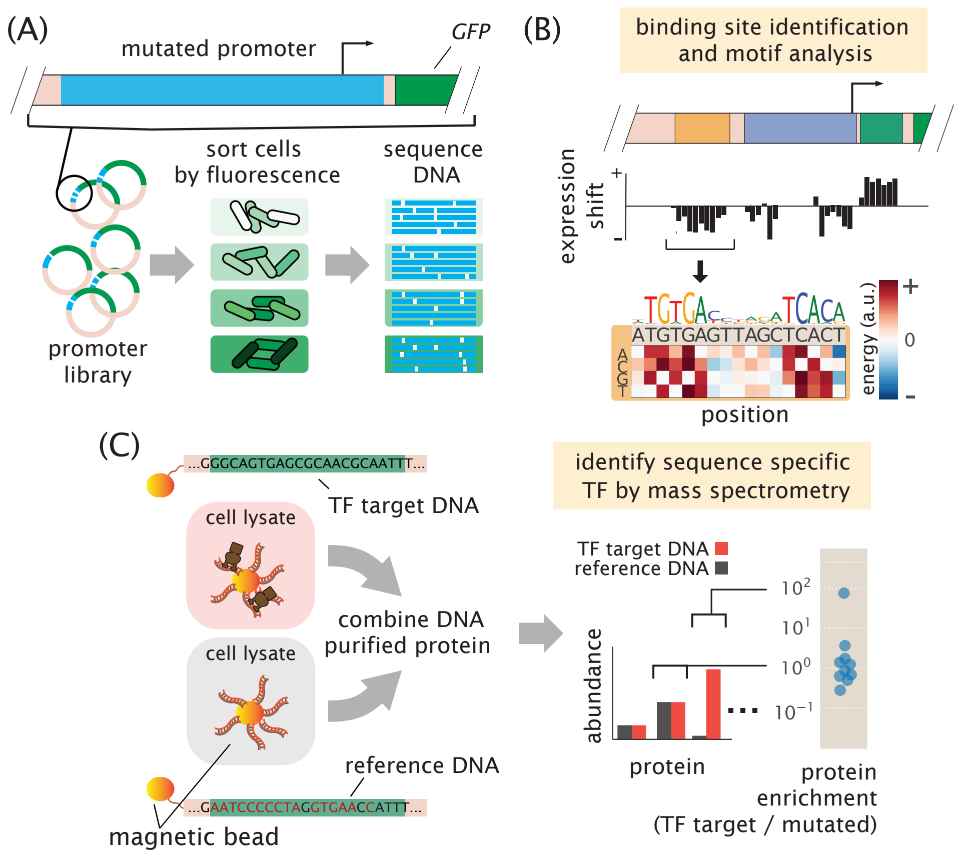


Figure 4.2: **Overview of approach to characterize transcriptional regulatory DNA, using Sort-Seq and mass spectrometry.** (A) Schematic of Sort-Seq. A promoter plasmid library is placed upstream of GFP and is transformed into cells. The cells are sorted into four bins by FACS and after regrowth, plasmids are purified and sequenced. The entire intergenic region associated with a promoter is included on the plasmid and a separate downstream ribosomal binding site sequence is used for translation of the *GFP* gene. (B) Regulatory binding sites are identified by calculating the average expression shift due to mutation at each position. The schematic shows the expression shift on a promoter region containing an activator (orange), RNAP (blue), and repressor (green) binding site. Quantitative models can be inferred to describe the associated DNA-protein interactions. An example energy matrix that describes the binding energy between an as yet unknown activator to the DNA is shown. By convention, the wild-type nucleotides have zero energy, with blue squares identifying mutations that enhance binding (negative energy), and where red squares reduce binding (positive energy). *(Caption continued on next page)*

Figure 4.2: *(continued from previous page)* (C) DNA affinity chromatography and mass spectrometry is used to identify the putative transcription factor (TF) for an identified repressor site. DNA oligonucleotides containing the target binding site are tethered to magnetic beads and used to purify the target transcription factor from cell lysate. Protein abundance is determined by mass spectrometry and a protein enrichment is calculated as the ratio in abundance relative to a second reference experiment where the target sequence is mutated away. SILAC labeling enables simultaneous measurement of both purifications by mass spectrometry.

Sort-Seq recovers the known regulatory features of well-characterized promoters

To first demonstrate Sort-Seq as a tool to discover regulatory binding sites *de novo* we began by looking at the promoters of *lacZYA* (*lac*), *relBE* (*rel*), and *marRAB* (*mar*) (Oehler et al., 1990; Gerdes et al., 2005; Alekshun and Levy, 1997), which are shown in the left panel of Fig. 4.1. These promoters have been studied extensively and provide a useful testbed of distinct regulatory motifs to test our approach. To proceed we constructed libraries for each promoter by mutating their known regulatory binding sites. We also considered two different mutation frequencies in our libraries. For *lac*, our library had a mutation rate of approximately three percent per bp, while *mar* and *rel* had a rate of roughly nine percent per bp. For a 20 bp binding site, this corresponds to an average of less than one mutation per sequence at the low mutation rate, and about two mutations at the high mutation rate (See Supplemental Section 4.5 and Fig. 4.8E,F for additional characterization).

We begin by considering the *lac* promoter. It contains three *lac* repressor (LacI) binding sites, two of which we consider here, and a cyclic AMP receptor (CRP) binding site. It exhibits the classic catabolic switch-like behavior that results in diauxie when *E. coli* is grown in the presence of glucose and lactose sugars (Loomis et al., 1967; Oehler et al., 1990; Busby and Ebright, 1999). We performed Sort-Seq with cells grown in M9 minimal media and at 37°C. The expression shifts at each position are shown in Fig. 4.3A, with annotated binding sites from RegulonDB noted above the plot. The expression shifts reflect the expected regulatory role of each binding site, showing positive shifts for LacI and negative shifts for CRP and RNAP. The difference in magnitude at the two LacI binding sites likely reflect the different binding energies between these two binding site sequences, with LacI O3 having an *in vivo* dissociation constant that is almost three orders of magnitude weaker than the LacI O1 binding site (Oehler et al., 1990; Garcia and Phillips, 2011).

Next we consider the *rel* promoter that transcribes the toxin-antitoxin pair RelE and RelB. It is one of about 36 toxin-antitoxin systems found on the chromosome, with important roles in cellular physiology including cellular persistence (Gerdes et al., 2005; Yamaguichi and Inouye, 2011; Maisonneuve and Gerdes, 2014). When the toxin, RelE, is in excess of its cognate binding partner, the antitoxin RelB, the toxin causes cellular paralysis through cleavage of mRNA (Griffin et al., 2013). Interestingly, the antitoxin protein also contains a DNA binding domain and is a repressor of its own promoter (Gotfredsen and Gerdes, 1998; Overgaard et al., 2009; Cataudella et al., 2012; Cataudella et al., 2013). We performed Sort-Seq with cells grown in M9 minimal media and at 37°C. The expression shifts are shown in Fig. 4.3B and were consistent with binding by RNAP and RelBE. In particular, a positive shift was observed at the binding site for RelBE, and the RNAP binding site showed mainly a negative shift in expression.

The third promoter, *mar*, is associated with multiple antibiotic resistance since its operon codes for the transcription factor MarA, which activates a variety of genes including the major multi-drug resistance efflux pump, ArcAB-tolC, and increases antibiotic tolerance (Alekshun and Levy, 1997). The *mar* promoter is itself activated by MarA, SoxS, and Rob (via the so-called marbox binding site), and further enhanced by Fis, which binds upstream of this marbox (Martin and Rosner, 1997). Under standard laboratory growth it is under repression by MarR (Aono et al., 1998). We found that the promoter's fluorescence was quite dim in M9 minimal media and instead grew libraries in lysogeny broth (LB) at 30°C (Seoane and Levy, 1995). Again, the different features in the expression shift plot (Fig. 4.3C) appeared to be consistent with the noted binding sites. One exception was that the downstream MarR binding site was not especially apparent. Both positive and negative expression shifts were observed along its binding site, which may be due to overlap with other features present including the native ribosomal binding site. There have also been reported binding sites for CRP (Ruiz and Levy, 2010; Zheng, 2004), Cra (Shimada et al., 2011), CpxR/CpxA (Weatherspoon-Griffin et al., 2014), and AcrR (Lee et al., 2014). However these studies either required overexpression of the associated transcription factor, were computationally identified, or demonstrated through *in vitro* assays and were not observed under the growth condition considered here.

While each promoter qualitatively showed the expected regulatory behavior in each expression shift plot, we were also interested in whether we could recover the quantitative sequence specificity of each transcription factor from our data. We inferred energy matrices and associated sequence logos for the binding sites of RNAP, LacI, CRP, RelBE, MarA, and Fis. These are shown in Fig. 4.3A-C and Fig. 4.9, and agreed with sequence logos generated from known genomic binding sites for these transcription factors (Pearson correlation coefficient r=0.5-0.9; see Supplemental Section 4.6).

For the repressors RelBE and MarR, there was no data available that characterized their sequence specificity with which to compare against our Sort-Seq results and validate binding by these transcription factors. We therefore repeated Sort-Seq in strains where the transcription factors coding regions were deleted. As noted earlier, we expect that the positive or negative expression shift associated with binding should no longer be observed. The associated expression shift plots are shown in Fig. 4.3D and Fig. 4.3E, for the promoters of *rel* in a $\Delta relBE$ strain and *mar* in a $\Delta marR$ strain, respectively. In each instance we no longer see the increase or decrease in expression that was associated with each transcription factor.

Identification of transcription factors with DNA affinity chromatography and quantitative mass spectrometry.

For our purpose of completely dissecting a promoter, it was next important to show that DNA affinity chromatography could indeed be used to identify transcription factors in *E. coli*. In particular, a challenge arises in identifying transcription factors due to their very low abundance. In *E. coli* the cumulative distribution in protein copy number shows that more than half have a copy number less than 100 per cell, with 90% having copy number less than 1,000 per cell. This is several orders of magnitude below that of many other cellular proteins (Li et al., 2014).

We began by applying the approach to known binding sites for LacI and RelBE. For LacI, which is present in *E. coli* in about 10 copies per cell, we used the strongest binding site sequence, Oid (*in vivo* $K_d \approx 0.05$ nM), and the weakest natural operator sequence, O3 (*in vivo* $K_d \approx 110$ nM) (Oehler et al., 1990; Oehler et al., 2006; Kuhlman et al., 2007; Garcia and Phillips, 2011). In Fig. 4.4A we plot the protein enrichments from each transcription factor identified by mass spectrometry. LacI was found with both DNA targets, with fold enrichment greater than 10 in each case, and significantly higher than most of the proteins detected (indicated by the shaded region, which represents the 95% probability density region of all proteins detected, including non-DNA binding proteins). Purification of LacI with about 10 copies per cell using the weak O3 binding site sequence is near the limit of what would be

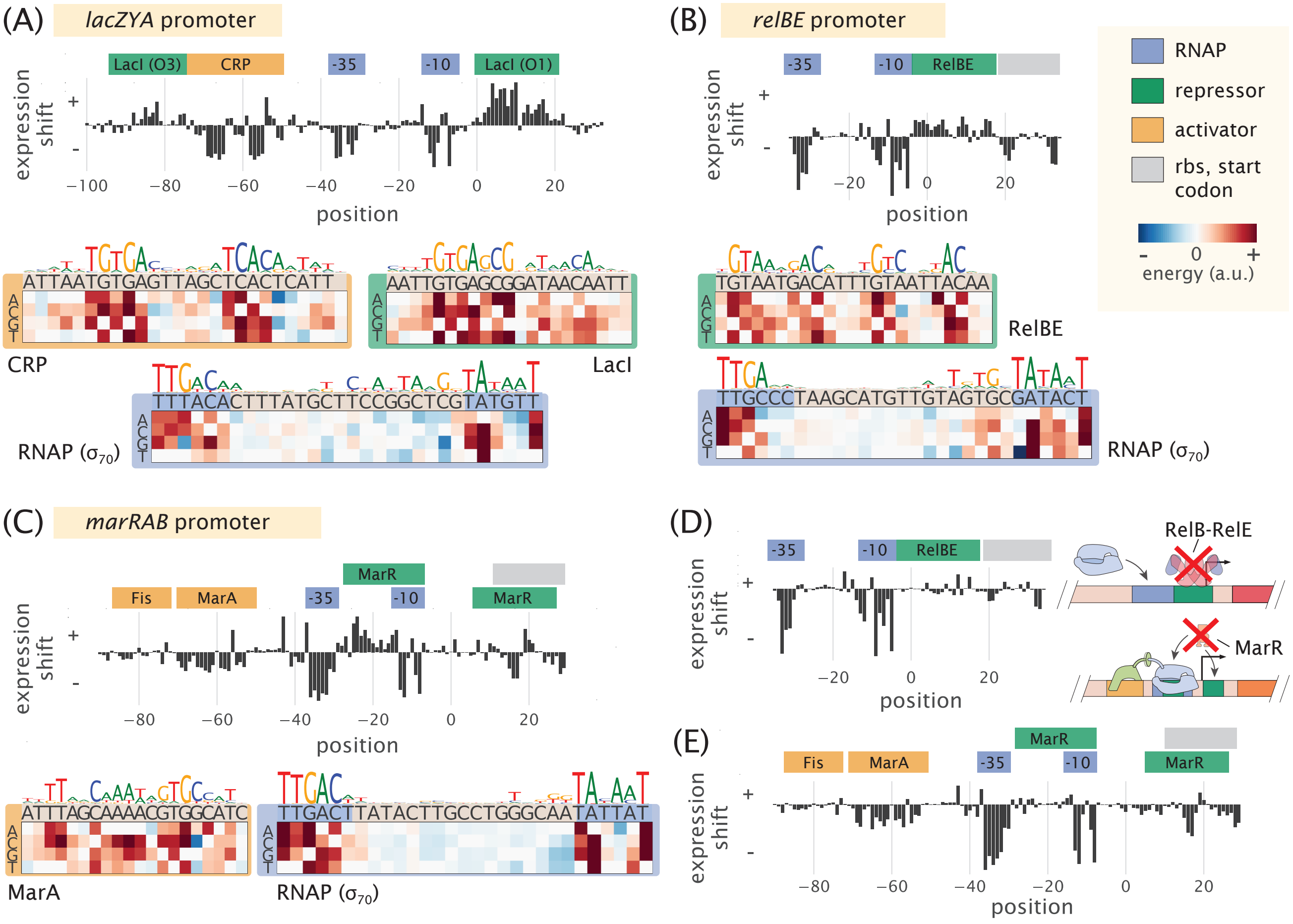


Figure 4.3: **Sort-Seq identifies the regulatory landscape of the *lac*, *rel*, and *mar* promoters**. (A) Sort-Seq of the *lac* promoter. Cells were grown in M9 minimal media with 0.5% glucose. Expression shifts are shown, with annotated binding sites for CRP (activator), RNAP (-10 and -35 subsites), and LacI (repressor) noted. Energy matrices and sequence logos are shown for each binding site. (B) Sort-Seq of the *rel* promoter. Cells were also grown in M9 minimal media with 0.5% glucose. The expression shifts identify the binding sites of RNAP and RelBE (repressor), and energy matrices and sequence logos are shown for these. (C) Sort-Seq of the *mar* promoter. Here cells were grown in Lysogeny broth (LB) at 30°C. The expression shifts identify the known binding sites of Fis and MarA (activators), RNAP, and MarR (repressor). Energy matrices and sequence logos are shown for MarA and RNAP. (D) Expression shifts for the *rel* promoter, but in a Δ*rel* genetic background. Cells were grown in conditions identical to (B) but do not show a positive expression shift across the entire RelBE binding site. (E) Expression shifts for the *mar* promoter, but in a Δ*marR* genetic background. The positive expression shift observed where MarR is expected to bind is no longer observed. Binding site annotations are identified in blue for RNAP sites, green for repressor sites, yellow for activator sites, and gray for ribosomal binding site and start codons. These annotations refer to the binding sites noted on RegulonDB that were observed in the Sort-Seq data.

necessary for most *E. coli* promoters.

To ensure this success was not specific to LacI, we also applied chromatography to the RelBE binding site. RelBE provides an interesting case since the strength of binding by RelB to DNA is dependent on whether RelE is bound in complex to RelB (with at least a 100 fold weaker dissociation constant reported in the absence of RelE (Li et al., 2008; Overgaard et al., 2008)). As shown in Fig. 4.4B, we found over 100 fold enrichment of both proteins by mass spectrometry. To provide some additional intuition into these results we also considered the predictions from a statistical mechanical model of DNA binding affinity (See Supplemental Section 4.7). As a consequence of performing a second reference purification, we find that fold enrichment should mostly reflect the difference in binding energy between the DNA sequences used in the two purifications, and be much less dependent on whether the protein was in low or high abundance within the cell. This appeared to be the case when considering other *E. coli* strains with LacI copy numbers between about 10 and 1,000 copies per cell (Fig. 4.10C). Further characterization of the measurement sensitivity and dynamic range of this approach is noted in Supplemental Section 4.8.

Sort-Seq discovers regulatory architectures in unannotated regulatory regions. Given that more than half of the promoters in *E. coli* have no annotated transcription factor binding sites in RegulonDB, we narrowed our focus by using several high-throughput studies to identify candidate genes to apply our approach (Marbach et al., 2012; Schmidt et al., 2016). The work by Schmidt *et al.* (Schmidt et al., 2016) in particular measured the protein copy number of about half the *E. coli* genes across 22 distinct growth conditions. Using this data, we identified genes that had substantial differential gene expression patterns across growth conditions, thus hinting at the presence of regulation and even how that regulation is elicited by environmental conditions (see further details in Supplemental Information Section A and Fig. 4.11A-C). On the basis of this survey, we chose to investigate the promoters of *purT*, *xylE*, and *dgoRKADT*. To apply Sort-Seq in a more exploratory manner, we considered three 60 bp mutagenized windows spanning the intergenic region of each gene. While it is certainly possible that regulatory features will lie outside of this window, a search of known regulatory binding sites suggest that this should be sufficient to capture just over 70% of regulatory features in *E. coli* and provide a useful starting point (Fig. 4.11D).

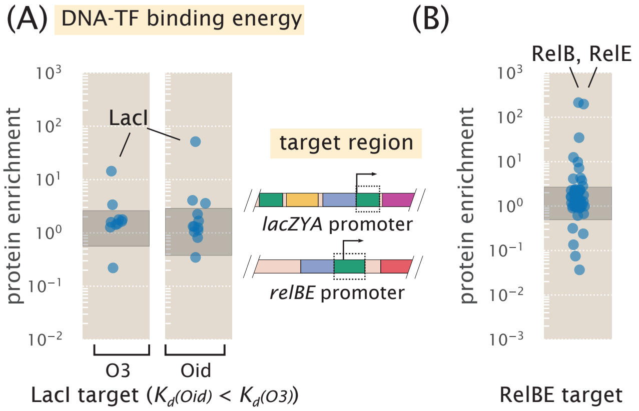


Figure 4.4: **DNA affinity purification and identification of LacI and RelBE by mass spectrometry using known target binding sites.** (A) Protein enrichment using the weak O3 binding site and strong synthetic Oid binding sites of LacI. LacI was the most significantly enriched protein in each purification. The target DNA region was based on the boxed area of the *lac* promoter schematic, but with the native O1 sequence replaced with either O3 or Oid. Data points represent average protein enrichment for each detected transcription factor, measured from a single purification experiment. (B) For purification using the RelBE binding site target, both RelB and its cognate binding partner RelE were significantly enriched. Data points show the average protein enrichment from two purification experiments. The target binding site is similarly shown by the boxed region of the *rel* promoter schematic. Data points in each purification show the protein enrichment for detected transcription factors. The gray shaded regions shows where 95% of all detected protein ratios were found.

The *purT* promoter contains a simple repression architecture and is repressed by PurR.

The first of our candidate promoters is associated with expression of *purT*, one of two genes found in *E. coli* that catalyze the third step in *de novo* purine biosynthesis (Rolfes, 2006; Cho et al., 2011). Due to a relatively short intergenic region, about 120 bp in length that is shared with a neighboring gene *yebG*, we also performed Sort-Seq on the *yebG* promoter (oriented in the opposite direction (Lomba et al., 1997); see schematic in Fig. 4.5A). To begin our exploration of the *purT* and *yebG* promoters, we performed Sort-Seq with cells grown in M9 minimal media with 0.5% glucose. The associated expression shift plots are shown in Fig. 4.5A. While we performed Sort-Seq on a larger region than shown for each promoter, we only plot the regions where regulation was apparent.

For the *yebG* promoter, the features were largely consistent with prior work, containing a binding sites for LexA and RNAP. However, we found that the RNAP binding site is shifted 9 bp downstream from what was identified previously through a computational search (Lomba et al., 1997), demonstrating the ability of our approach to identify and correct errors in the published record. We were also able to confirm that the *yebG* promoter was induced in response to DNA damage by repeating Sort-Seq in the presence of mitomycin C (a potent DNA cross-linker known to elicit the SOS response and proteolysis of LexA (Wade et al., 2005); see Fig. 4.12A, B, and D).

Given the role of *purT* in the synthesis of purines, and the tight control over purine concentrations within the cell (Rolfes, 2006), we performed Sort-Seq of the *purT* promoter in the presence or absence of the purine, adenine, in the growth media. In growth without adenine (Fig. 4.5A, right plot), we observed two negative regions in the expression shift plot. Through inference of an energy matrix, these two features were identified as the -10 and -35 regions of an RNAP binding site. While these two features were still present upon addition of adenine, as shown in Fig. 4.5B, this growth condition also revealed a putative repressor site between the -35 and -10 RNAP binding sites, indicated by a positive shift in expression (green annotation).

Following our strategy to find not only the regulatory sequences, but also their associated transcription factors, we next applied DNA affinity chromatography using this putative binding site sequence. In our initial attempt however, we were unable to identify any substantially enriched transcription factor (Fig. 4.12C). With repression observed only when cells were grown in the presence of adenine, we reasoned that the transcription factor may require a related ligand in order to bind the DNA, possibly through an allosteric mechanism. Importantly, we were able to infer an energy matrix to the putative repressor site whose sequence-specificity matched that of the well-characterized repressor, PurR (r=0.82; see Fig. 4.9). We also noted ChIP-chip data of PurR that suggests it might bind within this intergenic region (Cho et al., 2011). We therefore repeated the purification in the presence of hypoxanthine, which is a purine derivative that also binds PurR (Choi and Zalkin, 1992). As shown in Fig. 4.5C, we now observed a substantial enrichment of PurR with this putative binding site sequence. As further validation, we performed Sort-Seq once more in the adenine-rich growth condition, but in a Δ*purR* strain. In the absence of PurR, the putative repressor binding site disappeared (Fig. 4.5D), which is consistent with PurR binding at this location.

In Fig. 4.5E we summarize the regulatory features between the coding genes of *purT* and *yebG*, including the new features identified by Sort-Seq. With the appearance of a simple repression architecture (Bintu et al., 2005a) for the *purT* promoter, we extended our analysis by developing a thermodynamic model to describe repression by PurR. This enabled us to infer the binding energies of RNAP and PurR in absolute $k_B T$ energies (Atwal and Kinney, 2016), and we show the resulting model in Fig. 4.5E (see additional details in Appendix A).

The *xylE* operon is induced in the presence of xylose, mediated through binding of XylR and CRP.

The next unannotated promoter we considered was associated with expression of *xylE*, a xylose/proton symporter involved in uptake of xylose. From our analysis of the Schmidt *et al.* (Schmidt et al., 2016) data, we found that *xylE* was sensitive to xylose and proceeded by performing Sort-Seq in cells grown in this carbon source. Interestingly, the promoter exhibited essentially no expression in other media (Fig. 4.12E). We were able to locate the RNAP binding site between -80 bp and -40 bp relative to the *xylE* gene (Fig. 4.6A, annotated in blue). In addition, the entire region upstream of the RNAP appeared to be involved in activating gene expression (annotated in orange in Fig. 4.6A), suggesting the possibility of multiple transcription factor binding sites.

We applied DNA affinity chromatography using a DNA target containing this entire upstream region. Due to the stringent requirement for xylose to be present for any measurable expression, xylose was supplemented in the lysate during binding with the target DNA. In Fig. 4.6B we plot the enrichment ratios from this purification and find XylR to be most significantly enriched. From an energy matrix inferred for the entire region upstream of the RNAP site, we were able to identify two correlated 15 bp regions (dark yellow shaded regions in Fig. 4.6C). Mutations of the XylR protein have been found to diminish transport of xylose (Song and Park, 1997), which in light of our result, may be due in part to a loss of activation and expression of this xylose/proton symporter. These binding sites were also similar to those found on two other promoters known to be regulated by XylR (*xylA* and *xylF* promoters), whose promoters also exhibit tandem XylR binding sites and strong binding energy predictions with our energy matrix (Fig. 4.12F).

Within the upstream activator region in Fig. 4.6A there still appeared to be a binding site unaccounted for with these tandem XylR binding sites. From the energy matrix, we were further able to identify a binding site for CRP, which is noted upstream of the XylR binding sites in Fig. 4.6C. While we did not observe a significant enrichment of CRP in our protein purification, the most energetically favorable sequence predicted

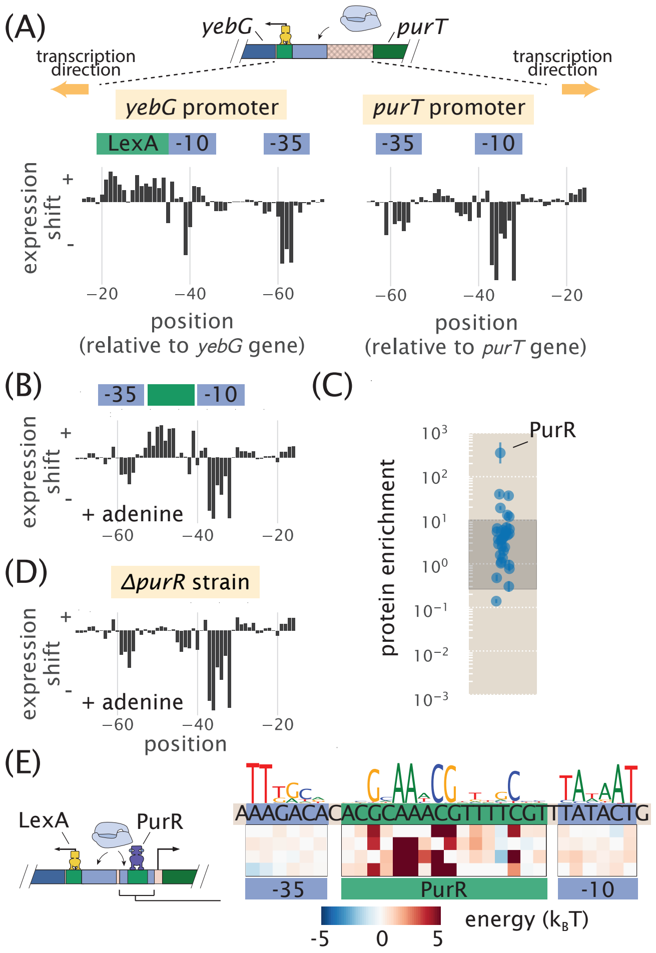


Figure 4.5: **Sort-Seq distinguishes directional regulatory features and uncovers the regulatory architecture of the *purT* promoter.** (A) A schematic is shown for the approximately 120 bp region between the *yebG* and *purT* genes, which code in opposite directions. Expression shifts are shown for 60 bp regions where regulation was observed for each promoter, with positions noted relative to the start codon of each native coding gene. Cells were grown in M9 minimal media with 0.5% glucose. The -10 and -35 RNAP binding sites of the *purT* promoter were determined through inference of an energy matrix and are identified in blue. (B) Expression shifts for the *purT* promoter, but in M9 minimal media with 0.5% glucose supplemented with adenine (100 μg/ml). A putative repressor site is annotated in green. (C) DNA affinity chromatography was performed using the identified repressor site and protein enrichment values for transcription factors are plotted. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose. Binding was performed in the presence of hypoxanthine (10 μg/ml). Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all protein detected. *(Caption continued on next page)*

Figure 4.5: *(continued from previous page)* (D) Identical to (B) but performed with cells containing a Δ*purR* genetic background. (E) Summary of regulatory binding sites and transcription factors that bind within the intergenic region between the genes of *yebG* and *purT*. Energy weight matrices and sequence logos are shown for the PurR repressor and RNAP binding sites. Data was fit to a thermodynamic of simple repression, yielding energies in units of $k_B T$.

by our model, TGCGACCNAGATCACA, closely matches the CRP consensus sequence of TGTGANNNNNNTCACA. In contrast to the *lac* promoter, binding by CRP here appears to depend more on the right half of the binding site sequence. CRP is known to activate promoters by multiple mechanisms (Browning and Busby, 2016), and CRP binding sites have been found adjacent to the activators XylR and AraC (Song and Park, 1997; Laikova et al., 2001), in line with our result. While further work will be needed to characterize the specific regulatory mechanism here, it appears that activation of RNAP is mediated by both CRP and XylR and we summarize this result in Fig. 4.6D (and considered further in Appendix A).

The *dgoRKADT* promoter is auto-repressed by DgoR, with transcription mediated by class II activation by CRP.

As a final illustration of the approach developed here, we considered the unannotated promoter of *dgoRKADT*. The operon codes for D-galactonate-catabolizing enzymes; D-galactonate is a sugar acid that has been found as a product of galactose metabolism (Cooper, 1978). We began by measuring expression from a non-mutagenized *dgoRKADT* promoter reporter to glucose, galactose, and D-galactonate. Cells grown in galactose exhibited higher expression than in glucose, as found by Schmidt *et al.* (Schmidt et al., 2016), and even higher expression when cells were grown in D-galactonate (Fig. 4.13A). This likely reflects the physiological role provided by the genes of this promoter, which appear necessary for metabolism of D-galactonate. We therefore proceeded by performing Sort-Seq with cells grown in either glucose or D-galactonate, since these appeared to represent distinct regulatory states, with expression low in glucose and high in D-galactonate. Expression shift plots from each growth conditions are shown in Fig. 4.7A.

We begin by considering the results from growth in glucose (Fig. 4.7A, top plot). Here we identified an RNAP binding site between -30 bp and -70 bp, relative to the native start codon for *dgoR* (Fig. 4.13B). Another distinct feature was a positive expression shift in the region between -140 bp and -110 bp, suggesting the presence

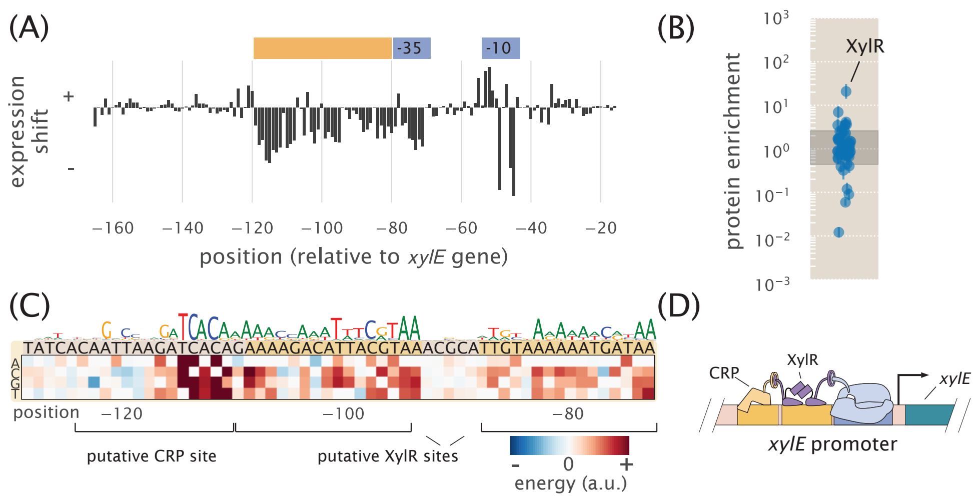


Figure 4.6: **Sort-Seq identifies a set of activator binding sites that drive expression of RNAP at the *xylE* promoter**. (A) Expression shifts are shown for the *xylE* promoter, with Sort-Seq performed on cells grown in M9 minimal media with 0.5% xylose. The -10 and -35 regions of an RNAP binding site (blue) and a putative activator region (orange) are annotated. (B) DNA affinity chromatography was performed using the putative activator region and protein enrichment values for transcription factors are plotted. Cell lysate was generated from cells grown in M9 minimal media with 0.5% xylose and binding was performed in the presence of xylose supplemented at the same concentration as during growth. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates. The gray shaded region represents 95% probability density region of all proteins detected. (C) An energy matrix was inferred for the region upstream of the RNAP binding site. The associated sequence logo is shown above the matrix. Two binding sites for XylR were identified (see also Fig. 4.9 and Fig. 4.12F) along with a CRP binding site. (D) Summary of regulatory features identified at *xylE* promoter, with the identification of an RNAP binding site and tandem binding sites for XylR and CRP.

of a repressor binding site. Applying DNA affinity chromatography using this target region we observed an enrichment of DgoR (Fig. 4.7B), suggesting that the promoter is indeed under repression, and regulated by the first coding gene of its transcript. As further validation of binding by DgoR, the positive shift in expression was no longer observed when Sort-Seq was repeated in a Δ*dgoR* strain (Fig. 4.7D and Fig. 4.13C). We also were able to identify additional RNAP binding sites that were not apparent due to binding by DgoR. While only one RNAP -10 motif is clearly visible in the sequence logo shown Fig. 4.7C (top sequence logo; TATAAT consensus sequence), we used simulations to demonstrate that the entire sequence logo shown can be explained by the convolution of three overlapping RNAP binding sites (See

Supplemental Information Section D and Fig. 4.13F).

Next we consider the D-galactonate growth condition (Fig. 4.7A, bottom plot). Like in the expression shift plot for the Δ*dgoR* strain grown in glucose, we no longer observe the positive expression shift between -140 bp and -110 bp. This suggests that DgoR may be induced by D-galactonate or a related metabolite. However, in comparison with the expression shifts in the Δ*dgoR* strain grown in glucose, there were some notable differences in the region between -160 bp and -140 bp. Here we find evidence for another CRP binding site. The sequence logo identifies the sequence TGTGA (Fig. 4.7C, bottom logo), which matches the left side of the CRP consensus sequence. In contrast to the *lac* and *xylE* promoters however, the right half of the binding site directly overlaps with where we would expect to find a -35 RNAP binding site. This type of interaction by CRP has been previously observed and is defined as class II CRP dependent activation (Browning and Busby, 2016), though this sequence-specificity has not been previously described.

In order to isolate and better identify this putative CRP binding site we repeated Sort-Seq in *E. coli* strain JK10, grown in 500 μM cAMP. Strain JK10 lacks adenlyate cyclase (*cyaA*) and phosphodiesterase (*cpdA*), which are needed for cAMP synthesis and degradation, respectively, and is thus unable to control intracellular cAMP levels necessary for activation by CRP (derivative of TK310 (Kuhlman et al., 2007)). Growth in the presence of 500 μM cAMP provided strong induction from the *dgoRKADT* promoter and resulted in a sequence logo at the putative CRP binding site that even more clearly resembled binding by CRP (Fig. 4.13E). This is likely because expression is now dominated by the CRP activated RNAP binding site. Importantly, this data allowed us to further infer the interaction energy between CRP and RNAP, which we estimate to be -7.3 k_BT (further detailed in Appendix A). We summarize the identified regulatory features in Fig. 4.7E.

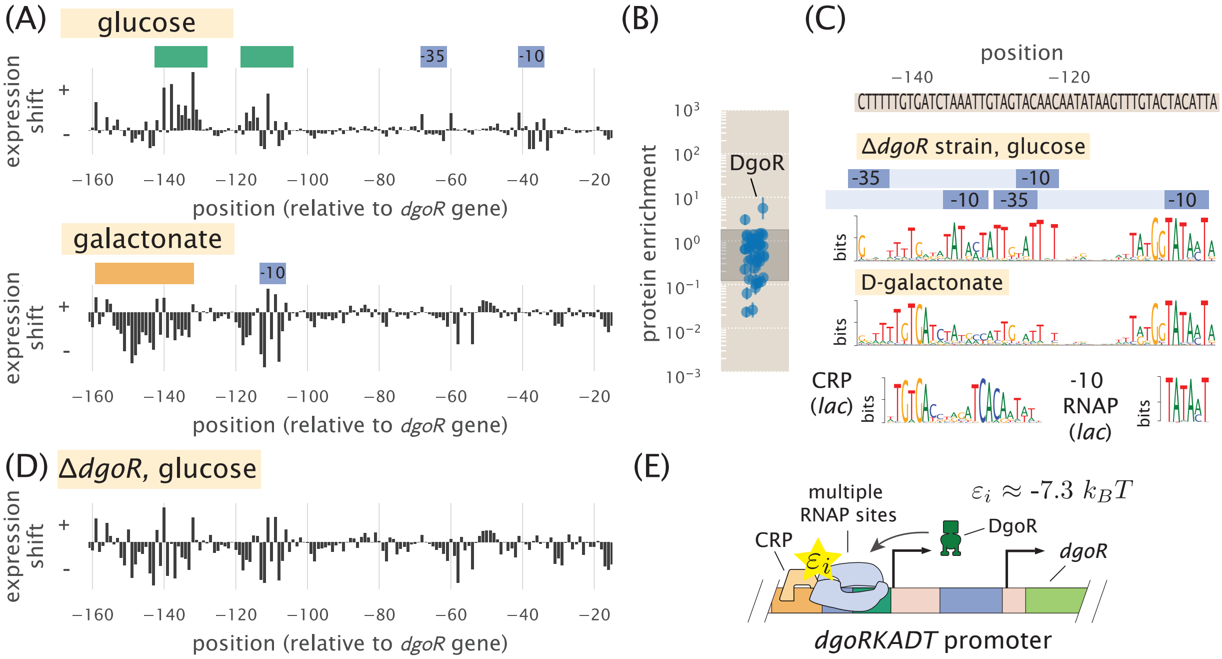


Figure 4.7: **The *dgoRKADT* promoter is induced in the presence of D-galactonate due to loss of repression by DgoR and activation by CRP.** (A) Expression shifts due to mutating the *dgoRKADT* promoter are shown for cells grown in M9 minimal media with either 0.5% glucose (top) or 0.23% D-galactonate (bottom). Regions identified as RNAP binding sites (-10 and -35) are shown in blue and putative activator and repressor binding sites are shown in yellow and green, respectively. (B) DNA affinity purification was performed targeting the region between -145 to -110 of the *dgoRKADT* promoter. The transcription factor DgoR was found most enriched among the transcription factors plotted. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all proteins detected. (C) Sequence logos were inferred for the most upstream 60 bp region associated with the upstream RNAP binding site annotated in (A). Multiple RNAP binding sites were identified using Sort-Seq data performed in a Δ*dgoR* strain, grown in M9 minimal media with 0.5% glucose. (further detailed in Fig. 4.13). Below this, a sequence logo was also inferred using data from Sort-Seq performed on wild-type cells, grown in D-galactonate, identifying a CRP binding site (class II activation; Browning and Busby, 2004). (D) Expression shifts are shown for the *dgoRKADT* promoter when performed in a Δ*dgoR* genetic background, grown in 0.5% glucose. This resembles growth in D-galactonate, suggesting D-galactonate may act as an inducer for DgoR. (E) Summary of regulatory features identified at *dgoRKADT* promoter, with the identification of multiple RNAP binding sites, and binding sites for DgoR and CRP. An initial estimate of -7.3 k_BT was determined for the interaction energy between CRP and RNAP, ε_i.

4.3 Discussion

We have established a systematic and scalable procedure for dissecting the functional mechanisms of previously uncharacterized regulatory sequences in bacteria. A massively parallel reporter assay, Sort-Seq (Kinney et al., 2010), is used to first elucidate the locations of functional transcription factor binding sites. DNA oligonucleotides containing these binding sites are then used to enrich the cognate transcription factors and identify them by mass spectrometry analysis. Information-based modeling and inference of energy matrices that describe the DNA binding specificity of regulatory factors provide additional insight into transcription factor identity and the growth condition dependent regulatory architectures.

To validate this approach we examined four previously annotated promoters. Our Sort-Seq results were in good agreement with established knowledge for *lacZYA*, *relBE*, *marRAB* (Oehler et al., 1990; Kinney et al., 2010; Garcia and Phillips, 2011; Bech et al., 1985; Gotfredsen and Gerdes, 1998; Overgaard et al., 2008; Seoane and Levy, 1995; Alekshun and Levy, 1997). For the *yebG* promoter, our approach corrected an error in a previous annotation. DNA affinity chromatography experiments on these promoters were found to be highly sensitive. In particular, LacI was unambiguously identified with the weak O3 binding site, even though LacI is present in only about 10 copies per cell (Garcia and Phillips, 2011).

Emboldened by this success, we then studied promoters having little or no prior regulatory annotation: *purT*, *xylE*, and *dgoR*. Through extensive modeling of the Sort-Seq data and DNA affinity chromatography of many identified binding sites, our analysis led to a collection of new regulatory hypotheses. For the *purT* promoter, we identified a simple repression architecture (Bintu et al., 2005a), with repression by PurR. The *xylE* promoter was found to undergo activation only when cells are grown in xylose, likely due to allosteric interaction between the activator XylR and xylose, and activation by CRP (Song and Park, 1997; Laikova et al., 2001). Finally, in the case of *dgoR*, the base pair resolution allowed us to tease apart multiple overlapping binding sites. In particular, we were able to identify multiple RNAP binding sites along the length of the promoter. Of these, one set of RNAP binding sites were repressed by DgoR when cells were grown in glucose, but activated through class II activation by CRP when D-galactonate was used as the sole carbon source. We view these results as a critical first step in the quantitative dissection of transcriptional regulation, which will ultimately be needed for a predictive understanding of how such regulation works. The regulatory cartoons shown in Fig. 4.5D, Fig. 4.6D, and

Fig. 4.7E will serve as a starting point for further mathematical dissection of these promoters and will lead to a series of quantitative predictions for how the different promoters work.

There are a number of ways to further increase the resolution and throughout of the methods we have described. Microarray-synthesized promoter libraries should allow multiple loci to be studied simultaneously. Landing pad technologies for chromosomal integration (Kuhlman and Cox, 2010; Zhang et al., 2016) should enable massively parallel reporter assays to be performed in chromosomes instead of on plasmids. Techniques that combine these assays with transcription start site readout (Vvedenskaya et al., 2015a; Vvedenskaya et al., 2015b) may further allow the molecular regulators of overlapping RNAP binding sites to be deconvolved, or the contributions from separate RNAP binding sites, like those observed on the *dgoR* promoter, to be better distinguished.

Although our work was directed toward regulatory regions of *E. coli*, there are no intrinsic limitations that restrict the analysis to this organism. Rather, it should be applicable to any bacterium that supports efficient transformation by plasmids. And although we have focused on bacteria, our general approach should be feasible in a number of eukaryotic systems – including human cell culture – using massively parallel reporter assays (Melnikov et al., 2012; Kheradpour et al., 2013; Patwardhan et al., 2012) and DNA-mediated protein pull-down methods (Mittler et al., 2009; Mirzaei et al., 2013) that have already been established.

4.4 Methods

Our intention was to construct a systematic and scalable experimental pipeline that would be applicable to the general objective of discovering regulatory architectures in generic bacteria. In this section we describe the work flow required by this pipeline with the aim of giving a sense of how each of the steps is carried out. We begin with a description of how we construct the mutated promoters used in the Sort-Seq experiment itself. Next we describe how the fluorescence levels are measured in a FACS machine and how the sorting and sequencing are performed. After that, the remainder of the Methods section focuses on the steps required to perform DNA affinity chromatography and mass spectrometry, which is necessary to identify the transcription factors that bind to the putative binding sites identified in the Sort-Seq procedure.

Sort-Seq libraries

Mutagenized single-stranded oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA), with a target mutation rate of 9%. In the case of the *lacZ* promoter, the associated Sort-Seq data was also used in the analysis in Razo-Mejia et al., 2014. The mutation rate for this library was approximately 3%. Each oligonucleotide was PCR amplified in order to produce double-stranded inserts, which were inserted into a PCR amplified plasmid backbone (i.e. vector) of pJK14 (Kinney et al., 2010) by Gibson Assembly (Gibson et al., 2009) (New England Biolabs, MA, USA). Note however that in the construction of the *lacZ* promoter, assembly was performed using restriction cloning as in Kinney *et al.* (Kinney et al., 2010).The template plasmid used for amplification of the backbone contained the toxic gene *ccdB* in place where the library was to be inserted. In this way any bacteria that took up any of the initial plasmid used in the PCR amplification would be removed from the population via negative selection due to toxicity by the *ccdB* gene (propagated in the immune strain DB3.1). This helped ensure that no template plasmid was propagated into the final plasmid library (see methods in reference (Kinney et al., 2010) for more detail). The plasmid is maintained at low copy numbers (about 5 copies per cell) by the SC101 origin of replication (Lutz and Bujard, 1997).

For each library construction, 40 ng of insert and 50 ng of backbone were combined in a 20 μL Gibson assembly reaction. To achieve high transformation efficiency, reaction buffer components from the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in 1 mL of SOC media, cells were diluted into 50 mL of LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies as assessed by plating 100 μL of cells diluted 1:10^4 onto an LB plate containing kanamycin.

Bacterial strains

All *E. coli* strains used in this work were derived from K12 MG1655, with deletion strains generated by the lambda red recombinase method (Datsenko and Wanner, 2000; Sawitzke et al., 2010). In the case of deletions for *lysA* ($\Delta lysA$::kan), *purR* ($\Delta purR$::kan), and *xylE* ($\Delta xylE$::kan), strains were obtained from the Coli Genetic Stock Center (CGSC, Yale University, CT, USA) and transferred into a fresh MG1655 strain using P1 transduction (Thomason et al., 2001). The others were generated in house and include the following deletion strains: $\Delta lacIZYA$, $\Delta relBE$::kan,

$\Delta marRAB$::kan, Δ *marR*::kan, $\Delta dgoR$::kan.

Here we describe the approach used to generate these deletion strains. Briefly, an overnight culture of MG1655 containing the plasmid pSIM6 was diluted 1:100 in 50 ml LB media and grown to an OD600 of ≈ 0.4 at 30°C. The culture was immediately placed in a water bath shaker at 43°C for 15 minutes and then cooled in an ice bath for 10 minutes. Cells were then spun down for 10 minutes (4,000 g, 4°C) and resuspended on ice in 50 ml of chilled water. This was repeated three times before resuspending in 200 μL of chilled water to generate competent cells. Homologous primer extension sequences for the appropriate gene were obtained from Baba *et al.* (Baba et al., 2006) and used to generate linear DNA containing a kanamycin resistance gene insert by PCR, which contained homology for the region on the chromosome to be deleted (Datsenko and Wanner, 2000). Electroporation of the competent cells was performed using 1 μL purified PCR product (about 100 ng DNA), mixed with 50 μL cells. Cells were immediately resuspended in 750 μL SOC media and placed on a shaker at 30°C for outgrowth, for 90-120 minutes. Cells were then plated on an LB-agar plate containing kanamycin (30 μg/ml) and grown overnight at 30°C. The deletions were confirmed by both colony PCR and sequencing. After confirmation, the deletion was transferred to a clean MG1655 strain through P1 transduction and selection on kanamycin. In the case of the lysine auxotrophic strain, we also confirmed deletion of *lysA* by checking that the cells were unable to grow in M9 minimal media unless lysine was supplemented (40 μg/ml).

To generate strains with different LacI tetramer copy numbers per cell (associated with data in Supplemental Fig. 4.10C), the LacI constructs from Garcia *et al.* (Garcia and Phillips, 2011) were P1 transduced into the $\Delta lacIZYA$ strain (integrated at the *ybcN* locus).

Sort-Seq fluorescence sorting

Cells were grown to saturation in LB and then diluted 1:10,000 into the appropriate growth media for the promoter under consideration. For cells grown in 0.23% D-galactonate in M9 minimal media, D-galactonate appeared to form precipitates, but cells otherwise appeared to grow normally. Upon reaching an OD600 of about 0.3, the cells were washed two times with chilled PBS by spinning down the cells at 4000 rpm for 10 minutes at 4°C. After washing with PBS, they were then diluted twofold with PBS to an OD of 0.1-0.15. This diluted cell solution was then passed through a 40 μm cell strainer to eliminate large clumps of cells.

A Beckman Coulter MoFlo XDP cell sorter was used to obtain fluorescence histograms of between 200,000 and 500,000 cell events per culture. For libraries, these histograms were used to set the four binning gates, which each covered ~ 15% of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were re-grown overnight in 10 ml of LB media, under kanamycin selection.

Sort-Seq sequencing

The contents of each bin were miniprepped following overnight growth (Qiagen, Germany). PCR was used to amplify the mutated region from each plasmid for Illumina sequencing. The primers contained Illumina adapter sequences as well as barcode sequences that enabled pooling of the samples. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) or NGX Bio (NextSeq sequencer; San Fransisco, CA). Single-end 100bp or paired-end 150bp flow cells were used, with a target read count of about 500,000 sequences per library bin. Joining of paired-end reads was performed with the FLASH tool (Magoc and Salzberg, 2011). For quality filtering, we collected sequences whose barcodes had a PHRED that score was greater than 20 at each position. Some libraries also contained non-mutagenized regions, and upon checking these, sequences that did not contain the expected sequence were excluded from our analysis. The total number of useful reads available to produce information footprints, fluorescence bin shift plots, energy weight matrices, and sequence logos from each Sort-Seq experiment generally ranged between 300,000 to 2,000,000 reads. Energy matrices were inferred using Bayesian parameter estimation with an error-model-averaged likelihood as previously described (Kinney et al., 2010; Kinney and Atwal, 2014), using the MPAthic software (Ireland and Kinney, 2016). A more detailed description of the data analysis procedures is available in Appendix A.

Lysate preparation and SILAC incorporation

SILAC labeling (Ong et al., 2002) was implemented by growing cells in either the stable isotopic form of lysine ($^{13}C_6H_{14}{}^{15}N_2O_2$), referred to as the heavy label, or natural lysine, referred to as the light label. By differentially labeling cell lysates we were able to simultaneously quantify the abundance of protein between two DNA affinity purification samples (i.e. one using a target binding site sequence and another as a reference control). This allows us to identify whether any protein shows a preference for the target binding site sequence.

To confirm heavy lysine was being incorporated, MG1655 Δ*lysA*::kan cells from an overnight M9 minimal media culture were diluted 1:200 and 1:1,000, and grown in 1 ml M9 minimal media supplemented with 40 μg/ml heavy lysine. Following approximately 7 and 10 cell divisions, cells were resuspended in lysis buffer (50 mM HEPES pH 7.5, 70 mM potassium acetate, 5 mM magnesium acetate, 0.2% (w/v) n-dodecyl-beta-D-maltoside, Roche protease inhibitor cOmplete tablet) and lysed by performing 10 freeze-thaw cycles with dry ice. Cellular debris was removed by centrifugation at 14000 g at 4°C on a tabletop centrifuge. Finally cellular lysates were prepared for mass spectrometry by in-solution digestion with endoproteinase Lys-C (Promega, Madison, WI). Digestion was performed as described elsewhere (Wisniewski et al., 2009) and labeling of the heavy isotope was confirmed by mass spectrometry measurement. In addition, we also characterized the SILAC enrichment ratio measurement by directly combining measurements from heavy and light lysates over a range from 0.1:1 to 1,000:1 heavy:light (see Supplemental Section 4.8).

To generate each lysate for DNA affinity purification experiments, an overnight starter culture of cells was grown in LB media supplemented with kanamycin (30 μg/ml). An aliquot was washed twice in M9 minimal media and resuspended to an OD600 of ≈1.0. For both heavy and light labeling, 500 ml M9 minimal media was then inoculated at 1:5,000 and grown to an OD600 of ≈0.6 (supplemented with the appropriate lysine; 40 μg/ml). Cultures were pelleted using an ultracentrifuge (8,000 g, 40 minutes) at 4°C and resuspended in chilled 20 ml lysis buffer containing 1% (w/v) n-dodecyl-beta-maltoside. The pellets could also be stored at -80°C for later use. Cells were then lysed with a Cell Disruptor (CF Range, Constant Systems Ltd., UK) and following removal of debris by centrifugation, concentrated to ~150 mg/ml using Amicon Ultra-15 centrifugation units (3kDa MWCO, Millipore). This provided about 600 μl of lysate, suitable for about six 80 μl DNA affinity purifications. Total protein concentration was assayed using the Bradford reagent (Sigma-Aldrich, St. Louis, MO). Following adjustment of protein concentration, sheared salmon sperm competitor DNA was added to the lysates (1 μg/ml; Life Technologies, Carlsbad, CA) and incubated for 10 minutes at 4°C. Finally, following centrifugation at 14,000 g to remove insoluble matter, lysates were either placed on ice or stored at 4°C prior to use.

Preparation of DNA-tethered magnetic beads

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dyanbeads MyOne T1, Life Technologies, Carlsbad) containing tethered

DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Note single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand. Briefly, DNA was suspended in annealing buffer (20 mM Tris-HCl, 10 mM MgCl2, 100 mM KCl) to 50 μM. Complementary strands were annealed by mixing 30 μL of the sense strand and 40 μL of the complement strand. Excess complement strand ensured all biotinylated-DNA would be in a double stranded form. Annealing was then performed using a thermocycler: 90°C for 5 minutes, gradient from 90°C to 65°C @ 0.1C /sec, incubated for 10 minutes at 65°C and allowed to return to room temperature on the thermocycler. Prior to attaching DNA, 150 μL beads were washed twice with 600 μL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and then twice with DW buffer (20 mM Tris-HC pH 8.0, 2 M NaCl, 0.5 mM EDTA (Mittler et al., 2009)). Approximately 640 pmol of DNA were then diluted to 600 μL in DW Buffer and incubated with the washed beads overnight at 4°C and on a rotatory wheel. Bound DNA was measured by determining the DNA concentration before and after incubation with beads using a NanoDrop (Thermo Scientific, Waltham, MA). Finally, beads were washed once with 600 μL TE buffer and three washes of 600 μL DW buffer, and resuspended in 150 μL DW buffer.

DNA affinity chromatography

Prior to DNA affinity purification the DNA tethered beads were incubated with blocking buffer (20 mM Hepes, pH 7.9, 0.05 mg/ml BSA, 0.05 mg/ml glycogen, 0.3 M KCl, 2.5 mM DTT, 5 mg/ml polyvinylpyrrolidone, 0.02% (w/v) n-dodeyl-β-D-maltoside; about 1.3 ml/mg beads (Mittler et al., 2009)) for one hour at 4°C for passivation. Excess blocking buffer was removed by washing the beads twice with 600 μL lysis buffer. The cell lysates were also incubated with washed magnetic beads that contained no tethered DNA. Following removal of these beads, cell lysates were incubated on a rotating wheel with the DNA tethered beads for approximately five hours at 4°C. Beads were then recovered with a magnet and washed three times using an equivalent volume of lysis buffer. The beads were then washed once more, but with NEB Buffer 3.1 (New England Biolabs, MA, USA). Both purifications (with the target DNA and reference control) were then combined by re-suspending in 50 μL NEB Buffer 3.1. To this suspension, 10 μl of the restriction enzyme PstI (100,000 units/ml, New England Biolabs) was added and incubated for 1.5 hours at 25°C. PstI cleaves the sequence CTGCAG, which was included between the biotin label and binding site sequence, allowing the DNA to be released from the magnetic beads. The beads were then removed and the samples diluted with 4x SDS-PAGE sample buffer. After incubation for five minutes at 95°C, the samples were then loaded on a SDS-PAGE gel (Any kD Mini-PROTEAN TGX Precast Protein Gels, 10-well , 50 µl; BioRad, CA, USA) and gel electrophoresis was performed for 45-55 minutes (200V) to separate proteins by size. The gel was stained using the Colloidal Blue Staining Kit (ThermoFisher Scientific, MA, USA) for visualization. Note that in general, we purified proteins from a heavy lysate using DNA containing the target binding site sequence, while devoting the light lysate to a control DNA sequence. However, for our LacI and RelBE, we also performed the alternative scenario (i.e. target binding site sequence purified with the light lysate). We did not observe major differences between either approach and therefore continued in our other experiments by purifying with the target binding site sequence in the heavy lysate.

In-gel digestion for mass spectrometry

After destaining, the gel was cut into four sections, each of which was cut into small pieces for in-gel digestion. The gel pieces were reduced, alkylated, and digested by endoproteinase Lys-C overnight at 37°C. This enzymatically cleaves proteins after lysine residues and is necessary for determining whether detected peptides are from the light or heavy lysine labeled purification. Digested peptides were extracted from gel and lyophilized. The peptide samples were further purified using StageTips to remove residual salts (Rappsilber et al., 2007). The extracts were re-suspended in 0.2% formic acid.

LC-MS/MS analysis and protein quantitation

Liquid chromatography-tandem-mass spectrometry (LC-MS/MS) experiments were carried out as previously described (Kalli and Hess, 2011). The LacI target purification experiments were performed on a nanoflow LC system, EASY-nLC II coupled to a hybrid linear ion trap Orbitrap Classic mass spectrometer equipped with a Nanospray Flex Ion Source (Thermo Fisher Scientific). The in- gel digested peptides were directly loaded at a flow rate of 500 nl/min onto a 16-cm analytical HPLC column (75 μm ID) packed in-house with ReproSil-Pur C18AQ 3 μm resin (120 Å pore size, Dr. Maisch, Ammerbuch, Germany). The column was enclosed in a column heater operating at 45°C. After 30 min of loading time, the peptides were separated in a solvent gradient at a flow rate of 350 nl/min. The gradient was as follows: 0–30% B (80 min), and 100% B (10 min). The solvent A consisted of 97.8% H2O, 2% ACN, and 0.2% formic acid and solvent B consisted of 19.8% H2O, 80% ACN, and 0.2% formic acid. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1600) in the Orbitrap (resolution 100,000) and subsequent 15 CID MS/MS scans (Top 15 method) in the linear ion trap. Collision induced dissociation (CID) was performed at normalized collision energy of 35% and 30 msec of activation time. All other measurements were performed on a hybrid ion trap-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific), which provided greater detection sensitivity and other fragmentation techniques as described. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1,800) in the Orbitrap (resolution 120,000) and subsequent 5 MS/MS scans also acquired in Orbitrap with 15,000 resolution. The MS/MS spectra were acquired for the top 5 ions alternating between higher collision dissociation (HCD) and electron transfer dissociation (ETD) fragmentations that are well suited for higher charge peptides. Higher collision dissociation was performed at a normalized collision energy of 30% and electron transfer dissociation reaction time was set to 100 msec. The analytical column for this instrument was a PicoFrit column (New Objective, Woburn, MA) packed in house with ReproSil-Pur C18AQ 1.9 μm resin (120Å pore size, Dr. Maisch, Ammerbuch, Germany) and the column was heated to 60°C. The peptides were separated either with a 90 or 60 min gradient (0-30% B in 90 min or 0-30% B in 60 min) at a flow rate of 220 nL/min.

Thermo RAW files were processed using MaxQuant (v. 1.5.3.30) (Cox and Mann, 2008; Cox et al., 2009). Spectra were searched against the UniProt E. coli K12 database (4318 sequences) as well as a contaminant database (256 sequences). Precursor ion mass tolerance was 4.5 ppm after recalibration by MaxQuant. Fragment ion mass tolerance was 20 ppm for high-resolution HCD and ETD spectra, and 0.5 Da for low-resolution CID spectra. Variable modifications included oxidation of methionine and protein N-terminal acetylation. Carboxyamidomethylation of cysteine was specified as a fixed modification. LysC was specified as the digestion enzyme and up to two missed cleavages were allowed. A decoy database was generated by MaxQuant and used to set a score threshold so that the false discovery rate was less than 1% at both the peptide and protein level. For all experiments match between runs and re- quantify were enabled. One evidence ratio per replicate per protein was required for quantitation. To calculate the overall protein ratio, the un-normalized protein replicate ratios were log transformed and then shifted so that the median protein log ratio within each replicate was zero (i.e., the median protein ratio was 1:1). The overall experimental log ratio was then calculated from the average of the replicate ratios.

Data analysis, code, and data curation

Additional details about the data analysis and characterization of Sort-Seq and DNA affinity chromatography can be found in the Supplemental material. The identification of regulated operons shown in Fig. 4.1 was performed using the annotated operons listed on RegulonDB (Gama-Castro et al., 2016), which are based on manually curated experimental and computational data. An operon was considered to be regulated if it had at least one transcription factor binding site associated with it. All code used for processing data and plotting, as well as the final processed data can be found on our GitHub repository (https://github.com/RPGroup-PBoC/sortseq_belliveau). Thermo RAW files for mass spectrometry are available on the jPOSTrepo repository (Okuda et al., 2017) under accession code PXD007892. Sort-Seq sequencing files are available on the Sequence Read Archive (accession code SRP121362).

Acknowledgements

We thank David Tirrell, Bradley Silverman, and Seth Lieblich for access and training for use of their Beckman Coulter MoFlo XDP cell sorter. We thank Jost Vielmetter and Nina Budaeva for access and training for use on their Cell Disruptor. We also thank Hernan Garcia, Manuel Razo-Mejia, Griffin Chure, Suzannah Beeler, Heun Jin Lee, Justin Bois, and Soichi Hirokawa for useful advice and discussion. Hernan Garcia also helped generate Figure 1. This work was supported by La Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA), the Gordon and Betty Moore Foundation through GBMF227, the National Institutes of Health 1S10RR029594-01A1 and the Beckman Institute. NB is a Howard Hughes Medical Institute International Student Research fellow.

4.5 Supplemental Information: Characterization of library diversity and sorting sensitivity.

Sort-Seq of the *rel* promoter using different sorting conditions.

In the work of the main text, Sort-Seq was performed by sorting cell libraries into four bins based on their fluorescence, each containing about 15 percent of the population. The remaining population was not collected and was discarded to waste. Due to the variability in expression of a single clonal population (Fig. 4.8A), sorting into a larger number of narrower bins was not expected to provide additional resolution for the sequence-dependent fluorescence distribution. Given the success in identifying the known regulatory binding sites of the *lacZ*, *relB*, and *marR* promoters, and agreement between the inferred sequences logos and available sequence logos (see Supplemental Fig. 4.9), these conditions appeared to provide sufficient information to accurately analyze our libraries.

However, in order to further confirm that our results were not being influenced by the specific sorting scheme, we also tested several other sorting conditions using our *relB* promoter library. Here cells were sorted into either 4 or 8 bins, with a sorting gate containing between 10 and 22 percent of the population per bin. The associated expression shift plots and information footprints (defined in Supplemental Section A) are shown in Fig. 4.8B-D. In general we found little difference between each of these experiments. Energy matrices for the binding sites were similarly in agreement, with a Pearson correlation coefficient between matrix parameters generally greater than 0.9 across the different conditions tested.

Analysis of library diversity using data from the *mar* promoter.

Here we provide additional characterization of the mutagenized promoter libraries, using a library from the *marR* promoter as a representative example (70 bp region containing RNAP and MarR repressor sites). With the exception of the *lacZ* promoter, all library oligonucleotide pools were purchased from Integrated DNA Technologies (USA) with a target mutation rate of nine percent per nucleotide position. For the *lacZ* promoter library, we purchased an oligonucleotide pool using their Ultramer branded technology to allow for a longer mutagenized region that covered the known set of regulatory binding sites. While we intended to have a similar mutation rate, we found a mutation rate closer to three percent per nucleotide position. While unexpected, this allowed us to test two different mutation rates in our initial validation of the methodology using well-characterized promoters.

To get a better sense of how the mutation rate varies across the libraries, we plot a histogram of the number of mutations per base pair for the entire set of sequences found in the *marR* promoter library (Fig. 4.8E). While we obtained an average mutation rate of 10.4% in this library, close to our target rate of 9%, there is some variability in this mutation rate as might be expected given that the incorporation of mutations in the DNA synthesis procedure is a random process. Since we are using these sequence data sets to infer sequence-specific models of binding between DNA and transcription factors, it was also of interest to consider the mutational coverage found within the library. As shown in Fig. 4.8F, all single-point mutations and a large fraction of two-point mutations were present within the library. Due to the large number of possible three point mutants in a 60 bp region, only a small subset of the possible sequences will be found in the library.

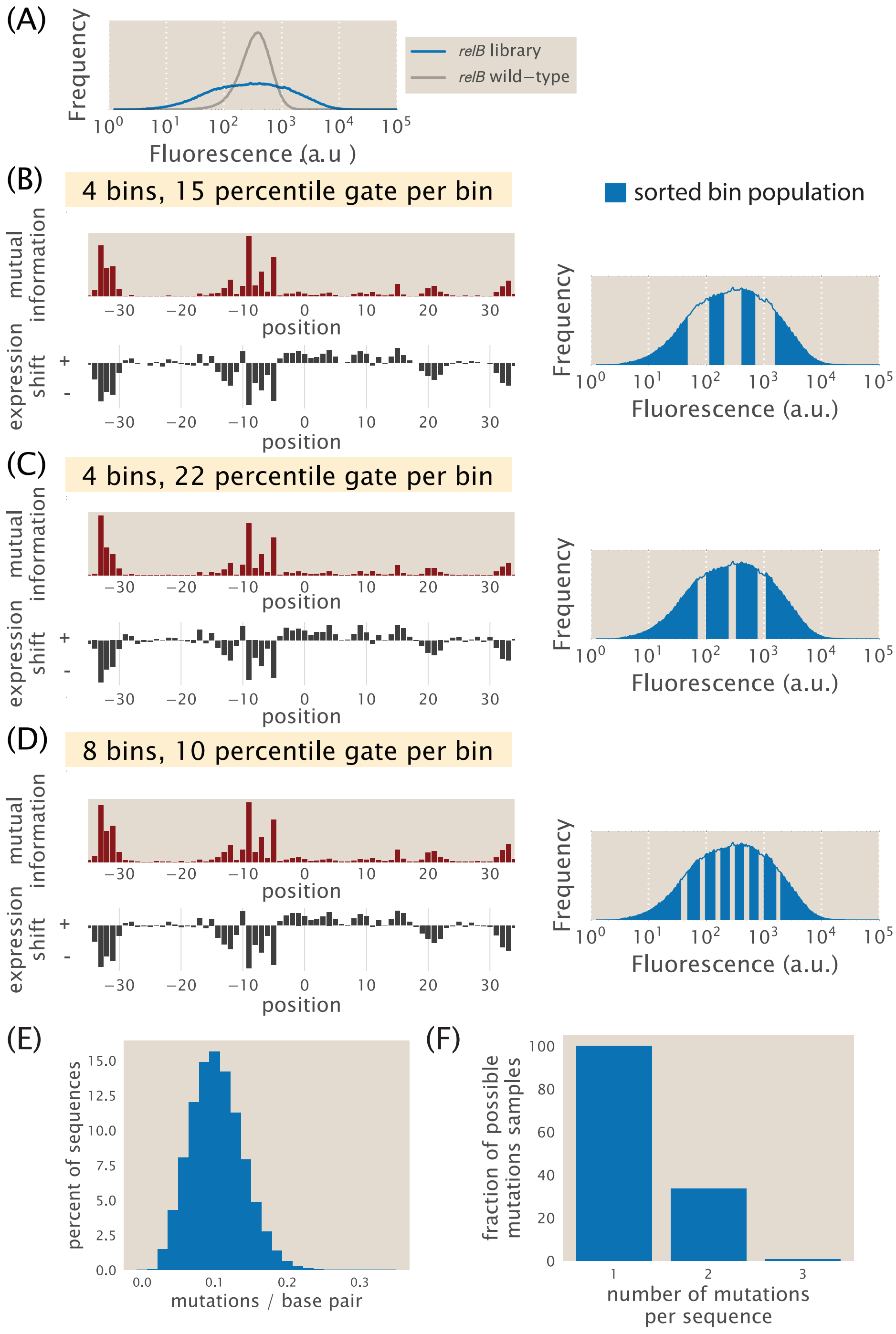


Figure 4.8: **Related to Fig. 4.2 and Fig. 4.3. Analysis of the library mutation spectrum and effect of Sort-Seq sorting conditions**. (A) Here we used our *relBE* promoter library to test whether the sorting procedure influenced our Sort-Seq data analysis. The fluorescence histogram of the wild-type promoter plasmid (single clonal population) and the mutated library for the *relB* promoter are shown. Expression shifts and information footprints are shown for cells sorted under three different scenarios in (B) -(D). In (B) cells were sorted using the approach of the main text where cells were sorted into 4 bins, each containing 15% of the population.

(continued on next page)

Figure 4.8: *(continued from previous page)* In (C) cells were similarly sorted into 4 bins, but where each bin contained about 22% of the population. In (D) cells were sorted into 8 bins, each containing about 10% of the population. The histograms beside each information footprint identify the approximate gating windows used to sort each fluorescence bin population. Histograms were based on between 400,000-500,000 cell counts. The same cell culture was used for each of the three Sort-Seq experiments performed here, sorted during the same sorting session. Cells were grown in M9 minimal media with 0.5% glucose like in the main text. (E) Histogram showing the mutation rate across all sequences found in the 60 bp *marRAB* library containing the RNAP and MarR repressor binding sites. Analysis was based on sequences from all fluorescence sorted bins. (F) The fraction of all possible unique sequences with one, two, or three mutations is shown for the *marRAB* library of (E). The coverage quickly drops for possible three-point mutations due to the large sequence space at this mutation frequency.

4.6 Supplemental Information: Generation of sequence logos.

Sequence logos provide a simple way to visualize the sequence specificity of a transcription factor to DNA, as well as the amount of information present at each position (Schneider and Stephens, 1990). Here we describe how we generate them using either known genomic binding sites or the energy matrices that were determined from our Sort-Seq data. In each case we need to calculate a 4xL position weight matrix for a binding site of length L, which is used to estimate the position-dependent information content needed to construct a sequence logo. We construct a position weight matrices (Supplemental Section 4.6 for genomic binding sites, and 4.6 for the energy matrix), and use these to construct sequence logos (Supplemental Section 4.6).

Generating position weight matrices from known genomic binding sites.

From RegulonDB, we find there are $N_g = 260$ known binding sites for CRP on the *E. coli* genome (Gama-Castro et al., 2016). To construct a position weight matrix using these genomic binding sites, we must first align all the sequences and determine the nucleotide statistics at each position. Specifically, we count the number of each nucleotide, N_{ij}, at each position along the binding site. Here the subscript i refers to the position, while j refers to the nucleotide, A, C, G, or T. We can then calculate a position probability matrix (also 4xL) where each entry is found by dividing these counts by the total number of sequences in our alignment,

$$p_{ij} = \frac{N_{ij}}{N_g}. \tag{4.1}$$

Note that in situations where the number of aligned sequences is small (e.g., less than five), pseudocounts (Nishida et al., 2009) are often added to regularize the probabilities of the counts in the calculation of position probabilities,

$$p_{ij} = \frac{N_{i,j} + B_p}{N_g + 4 \cdot B_p}, \tag{4.2}$$

where B_p is the value of the pseudocount. The argument for their use is that when selecting from a small number of binding site sequences, just by chance infrequent nucleotides will be absent, and assigning them a probability (p_{ij}, noted above) of zero may be too stringent of a penalty (Xia, 2012; Nishida et al., 2009). We let $B_p = 0.1$. In the limit of zero binding site sequences (i.e., with no sequences observed), this will result in probabilities p_{ij} approximately equal to the background probability used in calculating the position weight matrix below (and a non-informative sequence logo).

Finally, the values of the position weight matrix are found by calculating the log probabilities relative to a background model (Stormo, 2000),

$$PWM_{ij} = \log_2 \frac{p_{ij}}{b_j}. \tag{4.3}$$

The background model reflects assumptions about the genomic background of the system under investigation. For instance, in many cases it may be reasonable to assume each base is equally likely to occur. Given that we know the base frequencies for *E. coli*, we choose a background model that reflects these frequencies (b_j: A = 0.246, C = 0.254, G = 0.254, and T = 0.246 for strain MG1655; BioNumbers ID 100528, http://bionumbers.hms.harvard.edu). From Equation 4.3, we can see that the value at the i, j^{th} position will be zero if the probability, p_{ij}, matches that of the background model, but non-zero otherwise. This reflects the fact that base frequencies matching the background model tell us nothing about the binding preferences of the transcription factor, while deviation from this background frequency indicates sequence specificity.

Generating position weight matrices from Sort-Seq data.

Next we construct a position weight matrix using the CRP energy matrix from our Sort-Seq data. Here we appeal to the result from Berg and von Hippel, that the logarithms of the base frequencies above should be proportional to their binding energy contributions (Berg and Hippel, 1987; Stormo, 2000). Berg and von Hippel considered a statistical mechanical system containing L independent binding site positions, with the choice of nucleotide b_j at each position corresponding to a change in the energy level by ε_{ij} relative to the lowest energy state at that position. This ε_{ij} corresponds to the energy entry in our energy matrix, scaled to absolute units, $A \cdot \theta_{ij} + B$ (where θ_{ij} is the i, j^{th} entry as noted in Supplemental Section A). An important assumption is that all nucleotide sequences that provide an equivalent binding energy must have equal probability of being present as a binding site. In this way, we can relate the binding energies considered here to the statistical distribution of binding sites in the previous section. The probability p_{ij} of choosing nucleotide b_j at position i for protein binding will then be proportional to probability that position i has energy ε_{ij}. Specifically, the probabilities will be given by their Boltzmann factors normalized by the sum of states for all nucleotides,

$$p_{ij} = \frac{b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}{\sum_{j=A}^{T} b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}, \tag{4.4}$$

where $\beta = 1/k_B T$, with k_B is Boltzmann's constant and T the absolute temperature. Note that the energy scaling factor B drops out of this equation since it is shared across each term. As above, b_j refers to the background probabilities of each nucleotide.

One difficulty that arises when we use energy matrices that are not in absolute energy units is that we are left with an unknown scale factor A, preventing calculation of p_{ij}. We appeal to the expectation that mismatches usually involve an energy cost of 1-3 $k_B T$ (Lässig, 2007). In other work within our group, we have found this to be a reasonable assumption for LacI. Therefore, we approximate it such that the average cost of a mutation $\langle A \times \theta_{i,j} \rangle = 2 k_B T$. We can then calculate a position weight matrix from Equation 4.3.

Construction of sequence logo

With our position weight matrices in hand we can now construct sequence logos by calculating the average information content at each position along the binding site. With our four letter alphabet there is a maximum amount of information of 2 bits ($\log_2 4 = 2 bits$) at each position i. The information content will be zero at a position when the nucleotide frequencies match the genomic background, and will have a maximum of 2 bits only if a specific nucleotide is completely conserved. The total information content at position i is determined through calculation of the Shannon entropy, and is given by

$$I_i = \sum_{j=A}^{T} p_{ij} \cdot log_2 \frac{p_{ij}}{b_i} = \sum_{j=A}^{T} p_{ij} \cdot \text{PWM}_{ij}. \tag{4.5}$$

Here, PWM_{ij} refers to the i, j^{th} entry in the position weight matrix (Schneider et al., 1986; Stormo, 2000). The total information content contained in the position weight matrix is then the sum of information content across the length of the binding site.

To construct a sequence logo, the height of each letter at each position i is determined by

$$\text{Seqlogo}_{ij} = p_{ij} \cdot I_i, \tag{4.6}$$

which is in units of bits. This causes each nucleotide in the sequence logo to be displayed as the proportion of the nucleotide expected at that position scaled by the amount of information contained at that position (Schneider and Stephens, 1990). To construct sequence logos we use custom Python code written by Justin Kinney and available on our GitHub repository for this work (https://github.com/RPGroup-PBoC/sortseq_belliveau).

Comparison of Sort-Seq sequence logos.

For the various annotated binding sites identified in this work we used our Sort-Seq data to generate energy matrices. While these energy matrices provide a concrete way to understand the sequence-dependent DNA-protein interaction, it was also useful to generate sequence logos from energy matrices to visually compare with sequence logos more conventionally generated using known genomic binding site sequences. In Fig. 4.9 we show this comparison for transcription factors with three or more known genomic binding sites, with agreement more apparent when genomic binding site logos are based on a larger number of known sequences.

We also report the Pearson correlation coefficient between the position weight matrices from the Sort-Seq inference and the genomic alignment. To compare the two position weight matrices we first apply gauge fixing to each matrix in a similar manner as our energy matrix (see Supplemental Section A). Each column is set to have a mean energy of zero and the matrix norm (or inner product) is normalized to have value one. Under this constraint, the Pearson correlation coefficient is simply given by the summed product of matrix entries,

$$r = \frac{COV(\text{PWM'}_X, \text{PWM'}_Y)}{\sigma_X \cdot \sigma_Y} = \sum_{i=1}^{L} \sum_{j=A}^{T} \text{PWM'}_{X,i,j} \cdot \text{PWM'}_{Y,i,j}. \qquad (4.7)$$

Here, COV refers to the covariance between PWM'$_X$ and PWM'$_Y$, where the superscript prime indicates that the matrices have been gauge fixed (mean energy in each column of zero and the matrix norm of 1). The subscript X, for example, would correspond to the Sort-Seq matrix, and Y, to the genomic matrix. σ_X and σ_Y refer to the standard deviation of the matrix entries for PWM'$_X$ and PWM'$_Y$.

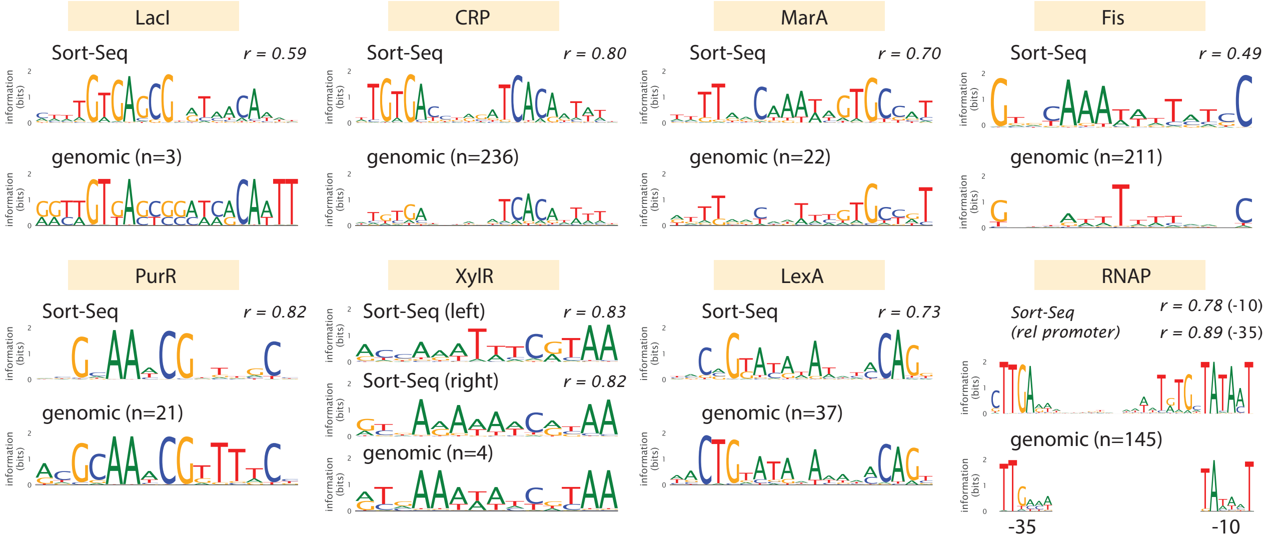


Figure 4.9: **Comparison between Sort-Seq and genomic-based sequence logos.** Comparisons are shown for LacI, CRP, MarA, Fis, PurR, XylR, LexA, and RNAP. Binding site sequences were obtained from RegulonDB, where *n* identifies the number of genomic binding sites that were used to construct the sequence logo. The Sort-Seq RNAP logo is based on data from the *rel* promoter. For the genomic RNAP logo, sequences were taken from computationally predicted RNAP binding sites on RegulonDB (top 3.3 % scored sequences using their reported metric) for the 6 bp regions of the -10 and -35 binding sites. Pearson correlation coefficients are calculated with Equation 4.7 using the position weight matrices from the Sort-Seq and genomic matrices. For LexA, the first four bp were not used in the calculation due to overlap with the -10 RNAP binding site of the *yebG* promoter.

4.7 Supplemental Information: Statistical mechanical model of the DNA affinity chromatography approach.

In order to better understand the factors that govern the success of the DNA affinity chromatography method, we took a statistical-mechanical approach to help identify the key parameters that will influence the fold enrichment of transcription factors that we measure. We are interested in calculating the probability that the transcription factor of interest binds to the target DNA sequence used for purification. We will ignore possible binding by proteins to the magnetic beads, to which the DNA oligonucleotides are tethered.

To calculate the probability that the transcription factor of interest is bound, we will simplify our problem by assuming that all other proteins in the lysate will bind the DNA with some average nonspecific binding energy. This must be included since these proteins will act as potential competitors for the tethered DNA. We must first enumerate the possible states of our DNA. For each DNA affinity purification, this will include the following three states: 1) no protein bound to the DNA, 2) the target transcription factor bound, and 3) a nonspecific protein is bound. These are shown in Supplemental Fig. 4.10D for each of the DNA oligonucleotides used for the two different purifications performed.

The non-normalized probability of each state occurring is simply given by $e^{-\beta(\varepsilon_i - \mu_i)}$. Here, ε_i is the protein-DNA binding energy and μ_i, the chemical potential, for species i (Weinert et al., 2014). $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. The chemical potential contains information about concentration, and it is possible to alternatively write the non-normalized probability in terms of these, which is given by $C_i/C_o e^{-\beta\Delta\varepsilon_i}$. Here, C_i is the concentration of protein species i, and C_o, is the standard concentration, which is taken as 1 M. $\Delta\varepsilon_i$ is the binding energy for species i, relative to the unbound state.

We can now write the statistical weight for each state, which is summarized in Supplemental Fig. 4.10D. We allow the unbound state to act as our reference state with an energy equal to zero, and a corresponding statistical weight of 1. The probability of our target protein being bound to a certain DNA target, $P_{bound,DNA}$, will then be given by the statistical weight for the state where the target protein is bound, divided by the sum of statistical weights for each state. This is given by

$$P_{bound,DNA} = \frac{\frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,DNA}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,DNA}}}, \tag{4.8}$$

where the subscript 'TF, DNA' identifies the target transcription factor and its binding to a specific DNA target. In regard to our two purifications shown in Supplemental Fig. 4.10D, $\Delta\varepsilon_{TF,s}$ refers to the binding energy of the transcription factor to its target binding site, while $\Delta\varepsilon_{TF,ns}$ refers to the nonspecific binding energy to non-target reference DNA. In addition, $\Delta\varepsilon_{ns}$ refers to the binding energy of other proteins present in the lysate, which may bind the DNA nonspecifically (and assumed to be similar im magnitude to $\Delta\varepsilon_{TF,ns}$).

We can now calculate the fraction of bound transcription factor, $P_{bound,DNA}$, using some reasonable values for *E. coli* (Bintu et al., 2005a; Moran et al., 2010). Here we use $C_{TF} = 10^{-8}M$ (about 10 copies per cell), $C_o = 1M$, $\Delta\varepsilon_{TF,s}$ = - $15k_BT$, and $\Delta\varepsilon_{ns}$ = - $5k_BT$. $C_{ns} = 3 \cdot 10^{-3}M$, which is the approximate number of proteins in *E. coli*. The specific numbers will depend on the DNA target sequence used, the concentration of target protein, as well as the lysate preparation itself. Here we find $P_{bound} \approx 0.02$. In contrast, for the nonspecifically bound fraction we calculate about a ten fold higher fraction of protein bound to the DNA. Even though the binding energy for a target transcription factor is significantly stronger than the competitor proteins that bind nonspecifically, the target transcription factor is generally several orders of magnitude lower in abundance. This result in particular highlights our rationale for using a additional reference purification to distinguish the target transcription factor from non-specifically bound proteins (Mittler et al., 2009). We consider the consequences of this next.

In this second reference purification, the DNA no longer has the target binding site, and thus the value of $P_{bound,DNA}$ for the transcription factor should be significantly smaller. We can use Equation 4.8 to calculate expected ratio of transcription factor bound to target DNA versus reference DNA, given by

$$\frac{P_{bound,target}}{P_{bound,reference}} = \frac{\frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,s}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,s}}} \cdot \frac{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,ns}}}{\frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,ns}}} \tag{4.9}$$

$$= \frac{e^{-\beta\Delta\varepsilon_{TF,s}}}{e^{-\beta\Delta\varepsilon_{TF,ns}}} \frac{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,ns}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta\Delta\varepsilon_{TF,s}}}. \tag{4.10}$$

Again, the subscript $\Delta\varepsilon_{TF,ns}$ refers to the binding energy of the transcription factor to the non-target (i.e., non-specific) reference DNA. Using the example values from our calculation of P_{bound} above, we find that $1 + \frac{C_{ns}}{C_o} e^{-\beta\Delta\varepsilon_{ns}} \gg e^{-\beta\Delta\varepsilon_{TF,s}} \gg e^{-\beta\Delta\varepsilon_{TF,ns}}$, with Equation 4.10 simplifying to

$$\frac{P_{bound,target}}{P_{bound,reference}} \approx \frac{e^{-\beta\Delta\varepsilon_{TF,s}}}{e^{-\beta\Delta\varepsilon_{TF,ns}}} = e^{-\beta(\Delta\varepsilon_{TF,s} - \Delta\varepsilon_{TF,ns})}. \tag{4.11}$$

This result suggests that the enrichment ratio should mainly depend on the difference in binding energy between the DNA sequences used in the two purifications. Our results from purifying LacI with strains containing different LacI copy number per cell and with different DNA target sequences (see Supplemental Fig. 4.10C) appear to agree with this result in general, where we see greater enrichment when using the strong Oid target LacI binding site sequence than the weaker O3 binding site sequence. This appears to influence the enrichment ratio more significantly than protein concentration, although further work will be needed to fully characterize this relationship.

4.8 Supplemental Information: DNA affinity chromatography and mass spectrometry experimentation and analysis.

In this section we provide additional details on the use of DNA affinity chromatography and mass spectrometry to identify the transcription factors that bind to our putative binding sites. In particular, we provide additional data to demonstrate protein labeling and characterize the dynamic range expected from our enrichment measurements (see Methods Section for more details about the approach). We also provide data from an affinity chromatography experiment in which the same DNA oligonucleotide sequence was used for both target and control purifications. The ideal result from such an experiment is that each protein detected is found in equal abundance between the two purifications performed, yielding an enrichment ratio equal to one. However, there is some inherent variability in such a measurement and we provide some characterization of that uncertainty here. Lastly, we provide additional data showing that we can purify and identify transcription factors at concentrations ranging from about 10 to 1,000 copies per cell.

Characterization of SILAC labeling and measurement of protein enrichment ratios.

To ensure *E. coli* cells incorporated the heavy isotope of lysine ($^{13}\mathrm{C}^{15}_{6}\mathrm{N}_2$-L-lysine, heavy lysine), we first generated an auxotropic strain which was unable to synthesize its own lysine through deletion of the *lysA* gene (Ong and Mann, 2007). LysA is an enzyme that catalyzes the last step in lysine biosynthesis. Furthermore, to ensure proteins would be sufficiently labeled when growing cultures for lysate preparation we inoculated our cultures with a large dilution of 1:5,000. This large dilution is important since the inoculate represents an unlabeled fraction of the cell population. We checked the effective labeling efficiency by combining lysates from cells grown with heavy and light (natural) lysine over a range of ratios between 0.1/1 to 1,000/1 (heavy / light). The measured ratio in abundance for each of the proteins detected among the two lysates are plotted in Supplemental Fig. 4.10A. In calculating these values, we found that the median average was measured to be 0.71 (heavy / light). We do not expect a discrepancy between measured heavy and light protein of similar abundance, and this suggested there may have been some inaccuracy in the Bradford assay used to measure protein concentration prior to mixing our lysates. We therefore renormalized the ratios according to this measured ratio. The data suggests a labeling efficiency of at least 99% (red dashed line, in comparison to perfect labeling shown by the gray dashed line). One important aspect highlighted by this data is that the highest enrichment ratio we should expect to measure in our DNA affinity experiments is several hundred fold.

Characterization of protein enrichment variability from identical DNA targets. For each DNA affinity chromatography experiment, we are trying to identify a DNA-binding protein that shows up in higher abundance when we use the target binding site sequence identified by Sort-Seq (i.e., a transcription factor binding site), relative to a purification where that target sequence has been mutated away. To ensure that our measured enrichment ratios were not an artifact of noise in the measurement, it was important to also check the measurement variability when both lysate purifications used an identical DNA sequence. In this way, we could characterize the inherent variability in such a measurement. To proceed, we performed experiments using the control DNA sequence that was used in our purification of the *purT* promoter target (Fig. 4.5C, though any DNA oligonucleotide could have been used). We performed this in triplicate and consider the average enrichment ratios for each protein measured across the three experiments. In the left panel of Supplemental Fig. 4.10B we show the average enrichment values that were measured for each of the detected proteins. Since many of the data points fall on top of one another, we also provide a histogram of the associated data (Supplemental Fig. 4.10B, right plot). Here we have taken the logarithm of the enrichment ratios so that the bins are equally spaced. The shaded region in both plots identifies the range between the 2.5^{th} and 97.5^{th} percentiles, highlighting that the majority of proteins were found between an enrichment ratio of 0.2 and 3.3 (or log enrichment ratio of between -1.5 and 1.2). The ideal enrichment expected would be a value of 1.0 or log ratio of 0. In the main text, the enrichment values for transcription factors found using targets associated with the *lacZ*, *relB*, *purT*, *xylE*, and *dgoR* promoters fall well outside of the range of variability established here.

Identification of LacI by mass spectrometry using strains with a variable LacI copy number.

Finally, one experiment that we performed in addition to purifying LacI with different strength binding site targets (Fig. 4.3F) was to consider the copy number per cell of the LacI target, as copy number should influence the fraction of bound LacI (see details in Supplemental Section 4.7). Here we used strains whose protein concentration has been measured during growth in M9 minimal media with 0.5% glucose and whose average LacI number had previously been measured to range from the native expression of 11 ± 2 tetramers per cell, to a maximum concentration of 870 ± 170 tetramers per cell. In Supplemental Fig. 4.10C we show the enrichment ratios measured for LacI from individual experiments (*n* = 1-2 per strain). Here we were able to purify LacI using either the weak O3 or strong Oid binding site sequence for each of the different strains, though we also see that the O3 target sequence provides an enrichment that is much closer to the tail of the control experiment in Fig. 4.10B. Additionally, while the copy number of LacI appears to affect the enrichment ratio in some experiments, it does not have a consistently significant effect.

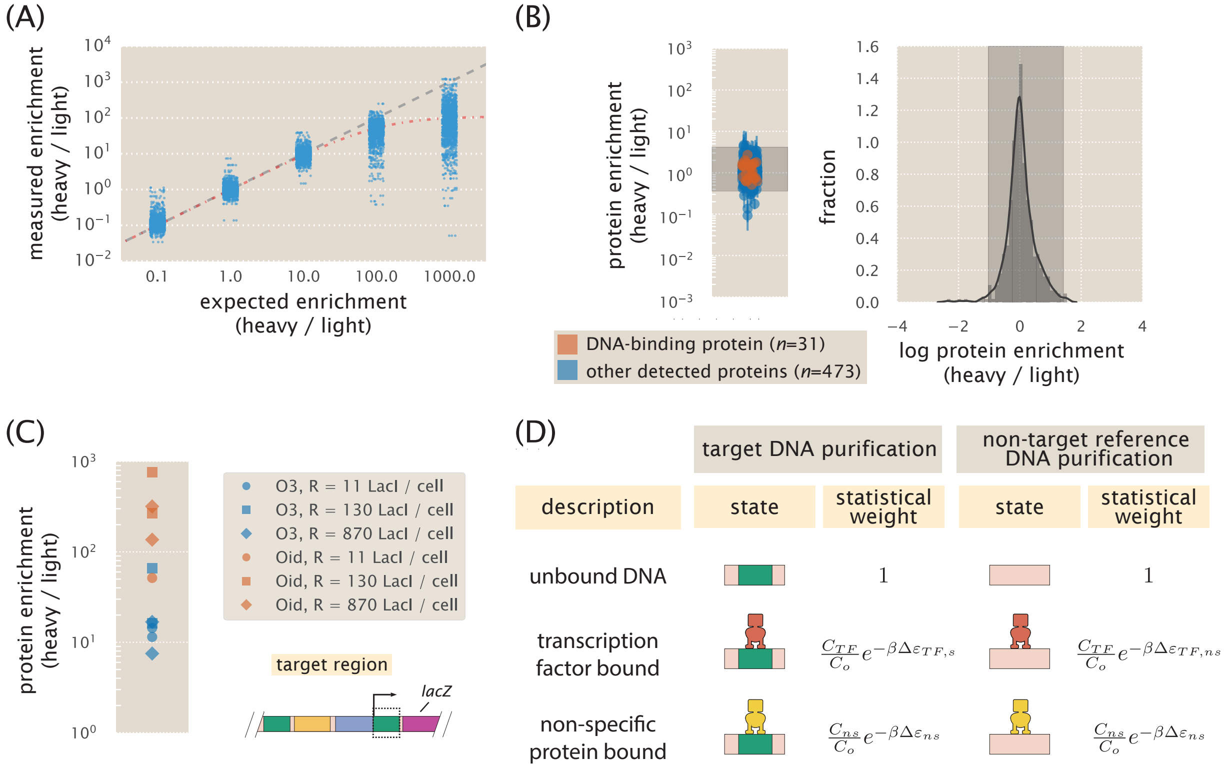


Figure 4.10: **Identification of transcription factors using DNA-affinity chromatography and mass spectrometry.** (A) Characterization of stable isotopic lysine labeling and mass spectrometry measurement sensitivity. Lysates from cell cultures grown in either heavy ($^{13}C^{15}_{6}N_2$-L-lysine) or normal L-lysine were combined at ratios between 0.1:1 to 1000:1 heavy:light and the measured ratios in abundance are plotted for each protein. Note that for the 1:1 ratio we found a median ratio of 0.71. We therefore renormalized the ratio values using this as a correction factor. Data points represent the average values from n = 3 replicates. The gray line represents the expected measurement under perfect labeling, while the red line represents a 99.1 % labeling efficiency (assuming that some fraction of heavy lysate is unlabeled). (B) DNA-affinity purification using the same DNA oligonucleotide to purify protein for both heavy and light cell lysates (n = 3). The scatter plot shows the average enrichment values for each protein detected. Proteins with DNA binding motifs (Keseler et al., 2013) are shown in red (n = 41), while other detected proteins are in blue (n = 581). Error bars represent the standard deviation, calculated from log protein enrichment values. The histogram shows the distribution of the measured ratios for all detected proteins, with 95% of the measurements contained between a log enrichment of -1.5 and 1.2, as indicated by the shaded region. Lysates were prepared from cells grown in M9 minimal media with 0.5% glucose. *(Caption continued on next page)*

Figure 4.10: *(continued from previous page)* (C) DNA-affinity purification of LacI using three different *E. coli* strains with repressor copy numbers per cell of 11 ± 2, 130 ± 20, and 870 ± 170 (tetramers per cell) (Garcia and Phillips, 2011). Operator strength was varied by purifying LacI with either the weak O3 or strong Oid operators. LacI was detected as the most significantly enriched protein among all proteins detected. Each data point represents the enrichment from a single purification experiment (n = 1-2 for each strain). (D) States and weights are shown for an oligonucleotide in which a target transcription factor and other cellular proteins compete for a DNA binding site. Within the cell lysate, the target protein is present at a concentration C_{TF}, while all other proteins, which may bind the DNA nonspecifically are present at a concentration C_{ns}. C_o is the standard concentration. The difference in energy between a repressor bound to the target DNA binding site and an unbound DNA is $\Delta\varepsilon_{TF,s}$ when the binding site is present. Otherwise, the binding energy is given by $\Delta\varepsilon_{TF,ns}$. Other proteins that bind nonspecifically, irrespective of the DNA sequence, have a binding energy of $\Delta\varepsilon_{ns}$.

4.9 Supplemental Information: Identification of unannotated promoters in *E. coli* whose expression appears to be regulated.

Here we briefly describe how the unannotated promoters of the main text (*purT*, *xylE*, and *dgoR*) were chosen. In attempting to identify candidate promoters to which to apply Sort-Seq, we made use of a variety of genome-wide datasets (Schmidt et al., 2016; Marbach et al., 2012; Li et al., 2014; Cho et al., 2011). Specifically, in the case of the *purT* promoter, network inference approaches (Marbach et al., 2012) and ChIP-chip data on the PurR repressor (Cho et al., 2011) led us to a variety of purine sensitive promoters that lacked regulatory annotation (others included *yieH* and *adeP*). Since the *purT* promoter lacked any experimental characterization, it appeared to be a good starting point with which to apply our approach.

The promoters of *xylE* and *dgoR*, were identified from a recent study by Schmidt *et al.* (Schmidt et al., 2016). They measured the copy number per cell of more than 2,300 proteins (about 55% of the *E. coli* proteome) across 22 growth conditions. These conditions included different carbon sources, temperature and pH, growth phase, media, and growth in chemostats. This provided us with a rich set of measurements with which to identify unannotated promoters where a particular growth condition influenced expression and may be under transcriptional regulation. The rest of this section describes how that data was used to identify candidate promoters.

In order to identify candidate genes using the mass spectrometry data, we ranked each protein based on its copy number in a particular growth condition, divided by the average copy number across the 22 conditions. Regulated proteins should be among those that exhibit a large change in copy number in one or a few growth conditions. As a confirmation of this, among the proteins with known regulation, we came across the GalE protein which was found to have significantly higher expression when cells were grown in galactose (Fig. 4.11A). GalE is involved in galactose catabolism, and its expression is known to increase due to loss of repression of the *galE* promoter when cells were grown in galactose (Irani et al., 1983; Semsey et al., 2007). Among promoters without any known regulation, we show the expression of DgoD in Fig. 4.11B for several different carbon sources. Cells grown in galactose showed much higher expression of the DgoD gene, with about 675 copies per cell, compared to at most 15 copies per cell across the other growth conditions. This is only one of many examples where a protein showed a large differential expression level across growth conditions and suggests many of these unannotated promoters may possibly be under regulation.

Another way to view this data is to calculate the coefficient of variation (the ratio of the standard deviation to the mean protein copy number) for each gene across the 22 growth conditions. In Fig. 4.11C, the coefficient of variation is plotted for each of the proteins measured in this study, separated by whether their promoter contains any known transcription factor binding sites (identified from RegulonDB; Gama-Castro et al., 2016). For GalE, whose expression was perturbed by growth in galactose, we find a calculated coefficient of variation of 1.18. Using this as our reference for a regulated gene that was perturbed in the study, there appear to be many unannotated genes that may in fact be under regulation. Among these, DgoD for example has a coefficient of variation of 3.64. Among the other proteins we investigated, XylE also has a high coefficient of variation, equal to 2.73, and shows almost no expression unless cells are grown in the presence of xylose as the carbon source. While we only pursued the promoters associated with expression of DgoR, DgoD, DgoK, DgoA, and XylE, there are many other unannotated promoters that will be of interest in future work.

4.10 Supplemental Information: Selection of the mutagenesis window for promoter dissection by Sort-Seq.

In designing our mutagenized promoter libraries, we found it useful to consider what was known regarding both the genes of interest and general patterns of transcriptional regulation in *E. coli* and bacteria more broadly. Two useful resources were RegulonDB (Gama-Castro et al., 2016) and EcoCyc (Keseler et al., 2013), which summarize much of what is known about transcriptional regulation in *E. coli*. RegulonDB, in particular, aims to compile all available data regarding gene regulation in *E. coli* into a single database and is the most complete record available for *E. coli* (Rydenfelt et al., 2014).

While Sort-Seq enables us to identify all proteins involved at a promoter, one potential limitation is that a transcription factor binding site will only be identified if it was contained within our mutagenized region. Using the known transcription factor binding sites in *E. coli* as a guide in our design, we made an educated guess regarding where we should search for transcription factor binding sites. Fig. 4.11D shows a histogram of all of the transcription factor binding site positions from RegulonDB. By staggering a set of 60bp windows to cover a 150 bp region, we found we would expect to capture 73 percent of the known transcription factor binding sites. We chose 60 bp-70 bp windows for most libraries since they could be readily synthesized by Integrated DNA Technologies (USA) and were more economical than longer oligonucleotides. We also included about 15 bp of overlap between staggered regions to provide some replicates of the mutated base pairs on the different libraries.

It is also useful to note that our approach does not require that this specific strategy be used to create mutagenized promoter constructs. The methodology only requires compatibility between the length of the mutagenized region probed and the sequencing platform used. Microarray synthesized oligonucleotides provide another approach for targeted oligonucleotide design (Bonde et al., 2015), and error-prone PCR can enable longer mutagenized windows within a single library (Rohlhill et al., 2017; Zhang et al., 2016). In addition, advances in sequencing, either through longer reads or alternative sequencing platforms such as PacBio (Pacific Bioscience, USA) and MinION (Oxford Nanopore Technologies, UK) are making it possible to sequence longer mutagenized regions, and CRISPR technologies could make it possible to identify longer range interactions such as DNA looping in bacteria (e.g., the 1 megabase region considered in (Fulco et al., 2016)).

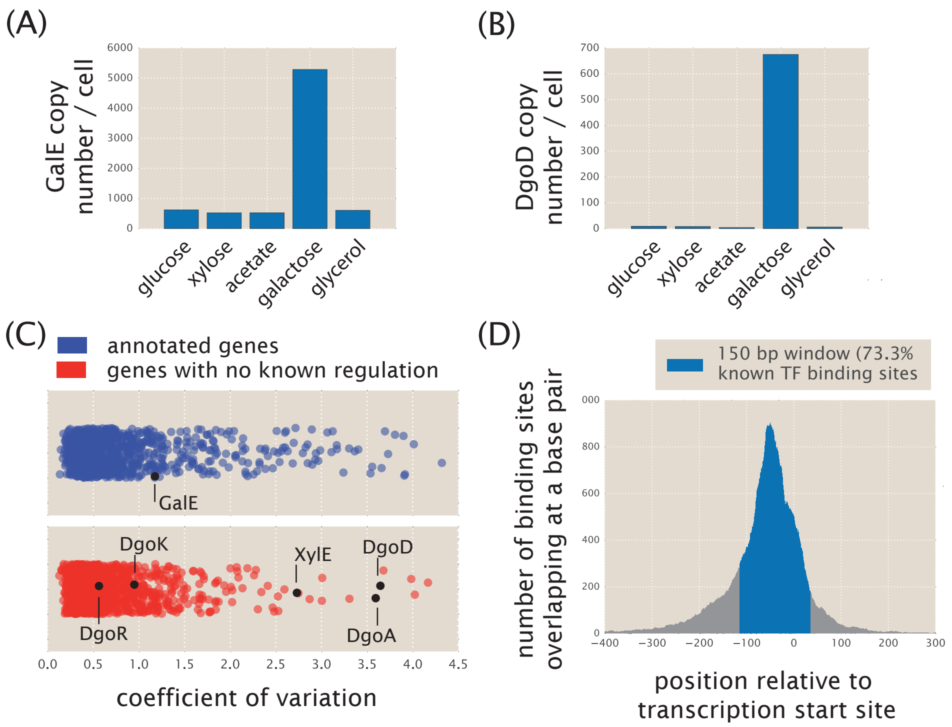


Figure 4.11: **Identification of unannotated genes with potential regulation and distribution of known transcription factor binding sites in *E. coli*.** (A) Here we show the protein copy numbers per cell for GalE across several carbon sources. Expression was sensitive to the presence of galactose which is consistent with its known regulation (with about 5000 copies per cell, versus about 500 for most other growth conditions). (B) DgoD was also found to be sensitive to the presence of galactose as the carbon source. The copy number was measured to be 675 copies per cell when cells were grown in galactose, and 15 copies per cell or less in all other conditions considered. For both (A) and (B), values are shown for growth in M9 minimal media, with glucose, xylose, acetate, galactose, and glycerol as carbon sources and obtained from (Schmidt et al., 2016). (C) Coefficient of variation (standard deviation divided by mean copy number) across the 22 growth conditions for each protein measured in (Schmidt et al., 2016). Proteins are identified as either having regulatory annotation (blue) or not (red) using the annotations in RegulonDB (Gama-Castro et al., 2016). GalE is noted among the annotated genes and provides a reference as a gene that is known to be regulated and be perturbed in this study, as shown in (A). (D). The histogram shows the genome-wide distribution of transcription factor binding sites relative to their respective transcription start sites. Binding sites were compiled from RegulonDB and used to calculate the number of overlapping binding sites at each position using the length and position of each binding site sequence. The location of the 150 bp mutation window used in this study is shown in blue, expected to capture upwards of 70% of known transcription factor binding site position.

4.11 Supplemental Information: Additional data from Sort-Seq experiments on the *yebG*, *purT*, *xylE*, and *dgoR* promoters.

Here we provide additional data and analysis on the promoters of *yebG*, *purT*, *xylE*, and *dgoR* to provide additional support for the results and conclusions made in the main text.

The *yebG* promoter

The *yebG* promoter is among a variety of genes known to increase expression when cells are under DNA damage stress (Wade et al., 2005), and shared the intergenic region with the *purT* promoter. In the main text we considered the *yebG* promoter in cells grown in standard M9 minimal media with 0.5% glucose (Fig. 4.5A). While the expression shifts appeared to align with annotated binding sites for LexA (positive shift), and the RNAP binding site (negative shift), we did not show evidence for the identity of each binding protein in the main text. Here we present results from our inference of energy matrices using our Sort-Seq data, which confirm the identity of the binding proteins. We also explore the regulation of *yebG* by perturbing the regulatory state through induction of the SOS response (Lomba et al., 1997; Wade et al., 2005).

We begin by considering the Sort-Seq data from cells grown in M9 minimal media with 0.5% glucose. In Fig. 4.12A we show the inferred energy matrices associated with the annotated site for LexA. This was in excellent agreement with the known sequence specificity of LexA (see Fig. 4.9 for a direct comparison with the genomic sequence logos). We note, however, that the RNAP binding site was shifted by 9 bp from the annotated binding site (Lomba et al., 1997), with an overlap between the -10 RNAP site and 4 bp of the LexA binding site.

We were also interested in confirming that the *yebG* promoter responds DNA stress and is induced as part of the SOS response. By repeating Sort-Seq in cells grown in non-lethal concentrations of mitomycin C (1 μg/ml) (Lomba et al., 1997) we observed a dramatic increase in expression relative to growth without mitomycin C. Fluorescence histograms showing expression from our plasmid reporter in non-mutagenized promoter constructs are shown in Fig. 4.12B. From the expression shift plots and information footprint (which are defined in Supplemental Section A and used in Kinney et al., 2010) in Fig. 4.12D we find that this is due to loss of repression at the LexA binding site. This is consistent with the expectation that LexA undergoes proteolysis as part of the SOS response (Wade et al., 2005).

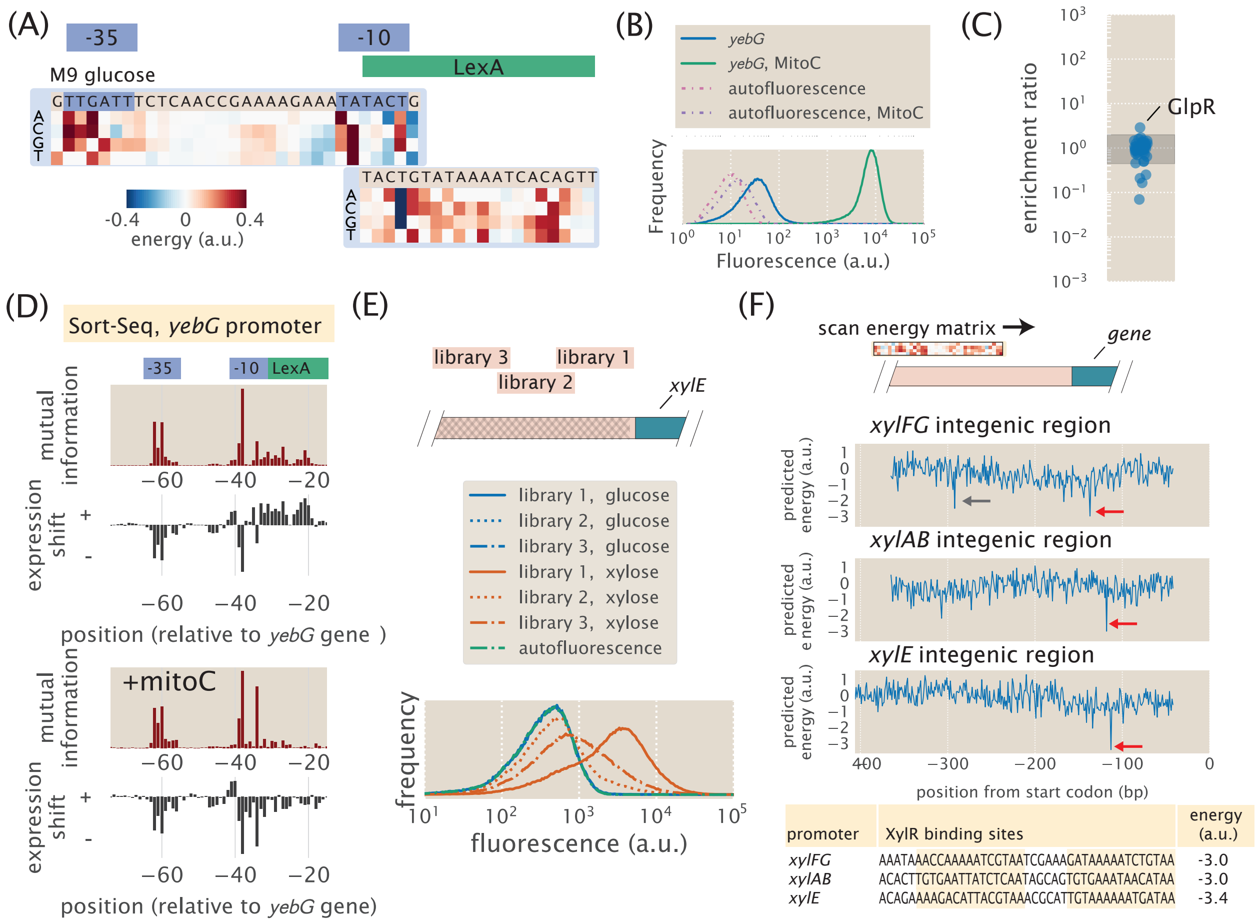
The *purT* promoter

When cells were grown in the presence of adenine, we identified a putative repressor site between the -10 and -35 regions of the RNAP binding site of the *purT* promoter. In our initial attempt to identify the associated transcription factor we performed a DNA affinity purification using conditions that matched the growth conditions where repression was observed. However, as shown in Fig. 4.12C, the most significantly enriched protein (GlpR) only showed an enrichment of about 2.9, which was near the shaded region associated with most other non-specific proteins detected. Only upon repeating our purification in the presence of hypoxanthine (10 μg/ml) (Fig. 4.5C) did we find enrichment of PurR (approximately 350 fold relative to our reference purification).

The *xylE* promoter

In the main text it was noted that we could not perform Sort-Seq on the *xylE* promoter unless cells were grown in xylose. In Supplemental Fig. 4.12E, we show the associated fluorescence histograms from libraries grown in either glucose or xylose. Interestingly, each mutated window was essentially identical to autofluorescence when cells were grown in glucose. In contrast, growth in xylose showed differential expression for each of the mutated regions. While the promoter was expected to be sensitive to the presence of xylose (causing an increase in expression; Schmidt et al., 2016), this was still a non-obvious result without prior knowledge of whether repressors or activators were involved.

In our analysis we also noted that the identified set of activator binding sites conformed well with the two other promoters regulated by XylR and CRP, namely *xylFG* and *xylAB*. Here we scanned our inferred energy weight matrix across the intergenic regions of *xylFG* and *xylAB*, in order gain further confidence that the identified feature matched the known binding specificity of these transcription factors. These are shown in Fig. 4.12F. At each position in these plots, we use the energy matrix to calculate the binding energy of the putative transcription factors. For each we identify a strong peak that does indeed align well with the annotated binding sites of XylR and CRP. While our predicted binding energies are not in absolute k_BT units, they are much more negative than the promoter background and predict a similar binding energy (in arbitrary units) to the binding site region of the *xylE* promoter.



promoter	XylR binding sites	energy (a.u.)
xylFG	AAATAAACCAAAAATCGTAATCGAAAGATAAAAATCTGTAA	-3.0
xylAB	ACACTTGTGAATTATCTCAATAGCAGTGTGAAATAACATAA	-3.0
xylE	ACAGAAAAGACATTACGTAAACGCATTGTAAAAAATGATAA	-3.4

Figure 4.12: **Extended analysis of the *yebG*, *purT*, and *xylE* promoters.** (A) Energy matrices were inferred for the binding sites of LexA and RNAP. Data are from cells grown in M9 minimal media with 0.5% glucose. (B) Fluorescence histograms for a wild-type *yebG* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, and with or without mitomycin C (1 μg/ml). Mitomycin C induces the SOS response (Lomba et al., 1997) and dramatically increases expression from the *yebG* promoter. Autofluorescence histograms refer to cells that did not contain the GFP promoter plasmid. (C) DNA affinity chromatography performed using the identified repressor site on the *purT* promoter. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose and binding was performed in the presence of adenine (100 μg/ml) to match the growth conditions where repression was observed. (D) Information footprints and expression shift plots are shown for the *yebG* promoter in the presence or absence of mitomycin C (1 μg/ml). Cells were grown in M9 minimal media 0.5% glucose. (E) Fluorescence histograms are shown for the three *xylE* libraries (different mutated regions), with cells grown in M9 minimal media with either 0.5% glucose or 0.5% xylose. While xylose led to differential expression for the different libraries, cells grown in glucose were identical to autofluorescence. *(Caption continued on next page)*

Figure 4.12: *(continued from previous page)* (F) The energy matrix associated with two tandem putative binding sites for xylR and CRP (Fig. 4.6C) was scanned across the intergenic regions of *xylAB*, *xylFG*, and *xylE*. The predicted energy is plotted for each position, and a strong binding site was identified in each promoter (red arrow). For *xylAB*, and *xylFG*, this matched the known binding sites for XylR and CRP on these promoters and their sequences and binding energy predictions are noted below the plots. The promoters of *xylAB* and *xylFG* share the same intergenic regions, but are in opposite coding directions. The reverse complement of the binding site identified in the *xylAB* promoter also showed a strong binding energy prediction (gray arrow in *xylFG* scan).

The *dgoR* promoter

The last promoter we considered was associated with the expression of the *dgoRKADT* operon. Due to the complexity observed, we were unable to show all data in the main text that supported our identification of the regulatory architecture. In particular, here we show the sensitivity to the different carbon sources considered and additional analysis of the identified regulatory binding sites for DgoR, RNAP, and CRP.

The *dgoR* promoter is induced when cells are grown in galactose and D-galatonate.

Prior to performing Sort-Seq on this promoter, we confirmed prior observations that expression was sensitive to the presence of galactose and D-galactonate (Cooper, 1978; Schmidt et al., 2016). Using a wild-type promoter plasmid for the *dgoR* promoter, cells were grown in M9 minimal media with either 0.5% glucose, 0.23% D-galactose, or 0.23% D-galactonate. Fluorescence histograms are shown in Fig. 4.13A, where we observed higher expression in galactose over glucose, and even higher expression when cells were grown in D-galactonate.

An RNAP binding site is apparent in the downstream region of the *dgoR* promoter when cells were grown in glucose.

In Fig. 4.7A we showed plots comparing the expression shifts upon mutation when cells were grown in either glucose or D-galactonate. In Fig. 4.13B we reproduce the expression shift plots along with an energy matrix for the region from approximately -70 to -30, which helped us to identify the RNAP binding site in this region. While the -10 TATAAT motif is quite apparent, the -35 site is less clear. Interestingly, while the -35 region shows a most energetically favorable sequence of TTTACA (close to the consensus of TTGACA), the wild-type sequence is CCCCCC and suggests this is a weak RNAP binding site.

Deletion of the *dgoR* gene recovers the induced phenotype.

Comparing the expression shift values at each position in cells grown in either glucose or D-galactonate, we find that they are poorly correlated (Fig. 4.13C, left plot). However, upon identifying DgoR as a putative regulator in the upstream region of the promoter, we then performed Sort-Seq in a $\Delta dgoR$ strain. This was shown in Fig. 4.7C with cells grown in glucose. Interestingly, the expression shifts were much more similar to the wild-type cells grown in D-galactonate, suggesting that deletion of *dgoR* has switched regulation to the induced state (Fig. 4.13C, right plot).

While it is unclear what causes the noisy profiles in the expression shift plots, one hypothesis was that the different RNAP binding sites were producing at least two distinct mRNA transcriptions, whose 5' untranslated might influence transcript stability and GFP expression. In particular, the upstream RNAP binding site will generate a much longer 5' untranslated region and mutations that influence mRNA structure and stability might show up as an effect on expression within the region we considered by Sort-Seq. Using the Salis lab ribosomal binding site calculator (Salis et al., 2009) and RNA structure predictions with NUPACK (Zadeh et al., 2011), we predicted the secondary structure of the two expected mRNAs transcripts (Fig. 4.13D). We find that the longer transcript (expected when cells are grown with D-galactonate) does indeed predict a strong secondary structure that alter translation from this transcript.

Simulations of upstream promoter region identify multiple overlapping RNAP binding sites.

Next we consider additional analysis to support the presence of overlapping RNAP sites that was noted in Fig. 4.7C. Since Sort-Seq does not differentiate between multiple transcription start sites, the sorted data will represent a mixture of all transcripts generated from the promoter. Using our RNAP energy matrix from the *relBE* promoter (with an additional 1 bp spacer included to increase the distance between -10 and -35 to 18 bp), we were able to identify multiple overlapping sequences that each predicted a similar binding energy by RNAP. The sequence logo in Fig. 4.7D therefore likely represents the convolution of these multiple binding sites and would explain why we do not see the conventional -35 RNAP motif in the sequence logo.

To convince ourselves that this was a reasonable hypothesis, we performed several Sort-Seq simulations of the *dgoR* promoter to estimate what we may have expected if 1-3 of these identified RNAP binding sites were functional. These simulations use energy matrices and a thermodynamic model of regulation to predict gene expression as a function of regulatory sequence in an attempt to mimic a real Sort-Seq experiment. The code used is available on our GitHub repository (https://github.com/RPGroup-PBoC/sortseq_belliveau) and we briefly describe the approach here. We began by first generating a library of five million mutated *dgoR* promoter sequences (10% mutation rate). We then assumed that transcription from each RNAP is proportional to $P/N_{NS} \cdot e^{-\beta E}$, where P is the RNAP copy number per cell, $N_{NS} = 4.6 \times 10^6$ refers to the number of non-specific binding sites on the genome, and $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. We introduced noise into our simulation by assuming that the RNAP copy number P was normally distributed across our library with a mean value of $3,000$ and standard deviation of 750 copies per cell (Schmidt et al., 2016; Jones et al., 2014). As defined in Supplemental Section A, E represents the binding energy as determined from the energy matrix.

Using these calculations to predict expression from each mutated sequence, the sequences were then computationally sorted in the same manner as that performed experimentally. We did this assuming the presence of one, two, or three active RNAP binding sites based on those identified. As shown in Fig. 4.13F, the presence of three RNAP binding sites produces a result that conforms much better with experimental results than the presence of only one RNAP binding site. Note that binding sites were successively included into the model based on their predicted binding energies (wild-type RNAP 1: -1.99 a.u., wild-type RNAP 2: -1.74 a.u., wild-type RNAP 3: -1.60 a.u.; versus an average of -0.14 a.u. and standard deviation of 0.56 a.u. when the energy matrix is scanned across the promoter).

The presence of the class II CRP activator binding site is enhanced using strain JK10, grown with cAMP.

Lastly, we show additional evidence to support the claim of a putative binding site for CRP. Since CRP binds to DNA by co-activation through binding with cAMP, we used the strain JK10 (based on TK310 Kinney et al., 2010; MG1655 $\Delta cyaA \Delta cpdA$), where we could control binding of CRP to DNA by direct supplement of cAMP to the growth media. Here we grew cells in EZrich MOPS media (Teknova, CA, USA) with D-galactonate as the carbon source and supplemented with 500 μM cAMP.

While the sequence logos in Fig. 4.7E showed a good match with the left site of the CRP binding site, our hypothesis here was that addition of a high concentration of cAMP might enhance the CRP motif in our data. This appeared to be the case, and the right side of the binding site (which overlaps the -35 RNAP binding site) shows a stronger preference for the sequence CAC than present with the wild-type *E. coli* strain (important for binding by CRP in both the *lac* and *xylE* promoters).

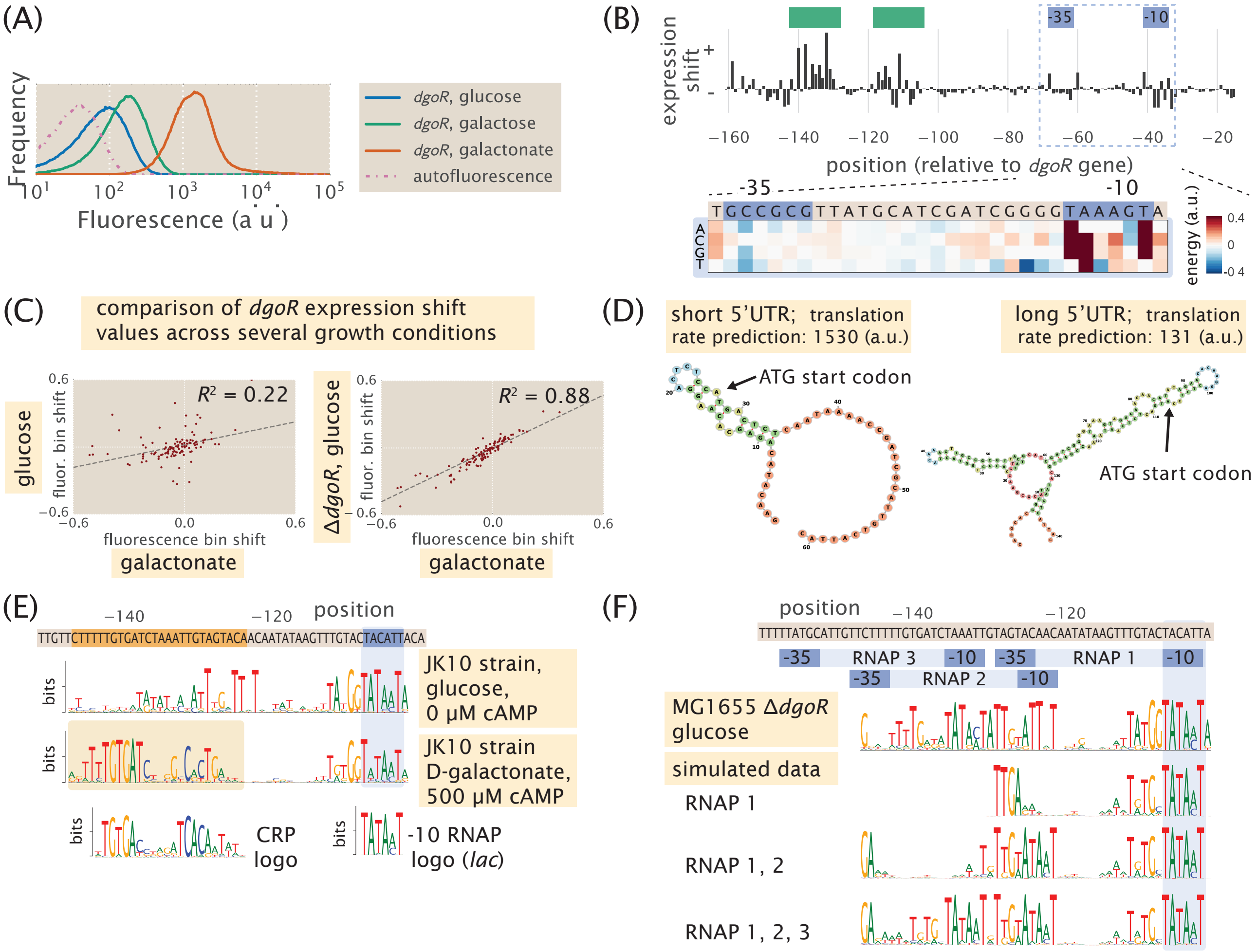


Figure 4.13: **Extended analysis of the *dgoR* promoter.** (A) Flow cytometry histograms of cells containing a wild-type *dgoR* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, 0.23% galactose, or 0.23% D-galactonate. (B) Identification of an RNAP binding site that appears active when cells are grown in M9 minimal media with 0.5% glucose. The inferred energy matrix exhibits a clear -10 RNAP binding site (consensus sequence is TATAAT) and a poor -35 binding site (CCCCCC). (C) Expression shift values are plotted against each other (glucose vs. D-galactonate, and Δ*dgoR* glucose vs. D-galactonate) for positions -120 bp to -14 bp relative to the *dgoR* coding gene. Note that these are the same values used to generate the bar plot in Fig. 4.7A, just plotted against each other for each position. Δ*dgoR* cells appear to have the same regulatory phenotype as cells grown in D-galactonate, with a line of best fit showing much higher correlation between these data sets. (D) Predicted RNA transcript structure based on the two distinct RNAP binding sites. Growth in D-galactonate leads to the long 5' untranslated region and is found to produce strong secondary structure which predicts significantly lower translation rates of the *dgoR* gene than with the short transcript. The ATG start codon is identified. *(Caption continued on next page)*

Figure 4.13: *(continued from previous page)* (E) Sequence logos were generated for the most upstream 60bp region containing the putative RNAP and CRP binding sites. Data is from Sort-Seq in strain JK10 (derivative of TK310 (Kinney et al., 2010)) and binding of CRP was induced through addition of 500 μM cAMP. Cells were grown in EZrich MOPS media (Teknova, CA, USA) with D-Galactonate as the carbon source. In comparison to the sequence logos shown in Fig. 4.7E, the right side of the CRP binding site has become more apparent. (F) Sequence logos are shown for simulated data for the upstream region of the *dgoR* promoter assuming one, two, or three RNAP binding sites. The top sequence logo shows the experimental result for Sort-Seq performed in a Δ*dgoR* genetic background, with cells grown in glucose.

References

Alekshun, M. N. and Levy, S. B. (1997). Regulation of chromosomally mediated multiple antibiotic resistance: the *mar* regulon. *Journal of Molecular Biology* 41.10, pp. 2067–2075.

Aono, R., Tsukagoshi, N., and Yamamoto, M. (1998). Involvement of Outer Membrane Protein TolC, a Possible Member of the mar-sox Regulon, in Maintenance and Improvement of Organic Solvent Tolerance of *Escherichia coli* K-12. *Journal of Bacteriology* 180.4, 938–944.

Arnold, C. D., Gerlach, D, Stelzer, C, Boryn, L. M., Rath, M, and Stark, A (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* 339.6123, pp. 1074–1077.

Atwal, G. S. and Kinney, J. B. (2016). Learning Quantitative Sequence-Function Relationships from Massively Parallel Experiments. *Journal of Statistical Physics* 162.5, pp. 1203–1243.

Baba, T. et al. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology* 2.1, p. 2006.0008.

Bech, F. W., Jørgensen, S. T., Diderichsen, B, and Karlström, O. H. (1985). Sequence of the *relB* transcription unit from *Escherichia coli* and identification of the *relB* gene. *The EMBO journal* 4.4, pp. 1059–1066.

Berg, O. G. and Hippel, P. H. von (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* 193.4, pp. 723–743.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005a). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics and Development* 15.2, pp. 116–124.

Bonde, M. T., Kosuri, S., Genee, H. J., Sarup-Lytzen, K., Church, G. M., Sommer, M. O. A., and Wang, H. H. (2015). Direct Mutagenesis of Thousands of Genomic Targets Using Microarray-Derived Oligonucleotides. *ACS Synthetic Biology* 4.1, pp. 17–22.

Bonocora, R. P. and Wade, J. T. (2015). ChIP-Seq for genome-scale analysis of bacterial DNA-binding proteins. (New York, Humana Press), pp. 327–340.

Browning, D. F. and Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology* 2.1, pp. 57–65.

Browning, D. F. and Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology* 14.10, pp. 638–650.

Busby, S. and Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology* 293.2, pp. 973–979.

Cataudella, I., Trusina, A., Sneppen, K., Gerdes, K., and Mitarai, N. (2012). Conditional cooperativity in toxin-antitoxin regulation prevents random toxin activation and promotes fast translational recovery. *Nucleic Acids Research* 40.14, 6424–6434.

Cataudella, I., Sneppen, K., Gerdes, K., and Mitarai, N. (2013). Conditional cooperativity of toxin - antitoxin regulation can mediate bistability between growth and dormancy. *Nucleic Acids Research* 9.8, e1003174.

Cho, B.-K., Federowicz, S. A., Embree, M., Park, Y.-S., Kim, D., and Palsson, B. (2011). The PurR regulon in *Escherichia coli* K-12 MG1655. *Nucleic Acids Research* 39.15, pp. 6456–6464.

Choi, K. Y. and Zalkin, H (1992). Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *Journal of Bacteriology* 174.19, pp. 6207–6214.

Cipriano, M. J., Novichkov, P. N., Kazakov, A. E., Rodionov, D. A., Arkin, A. P., Gelfand, M. S., and Dubchak, I. (2013). RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* 14.1, pp. 213–221.

Cooper, R. (1978). The utilisation of D-galactonate and D-2-oxo-3-deoxygalactonate by *Escherichia coli* K-12. Biochemical and genetical studies. *Archives of Microbiology* 1.118, pp. 199–206.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26.12, pp. 1367–1372.

Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J. V., and Mann, M. (2009). A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nature Protocols* 4.5, pp. 698–705.

Datsenko, K. A. and Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences* 97.12, pp. 6640–6645.

Fulco, C. P. et al. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* 354.6313, pp. 769–773.

Gama-Castro, S. et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* 44.D1, pp. D133–D143.

Garcia, H. G. and Phillips, R. (2011). Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–8.

Gerdes, K., Christensen, S. K., and Løbner-Oleson, A. (2005). Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology* 2.5, pp. 371–382.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6.5, pp. 343–345.

Gotfredsen, M. and Gerdes, K. (1998). The *Escherichia coli relBE* genes belong to a new toxin-antitoxin gene family. *Molecular Microbiology* 29.4, 539–548.

Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J., and Busby, S. J. W. (2005). Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proceedings of the National Academy of Sciences* 102.49, pp. 17693–17698.

Griffin, M. A., Davis, J. H., and Strobel, S. A. (2013). Bacterial Toxin RelE: A Highly Efficient Ribonuclease with Exquisite Substrate Specificity Using Atypical Catalytic Residues. *Biochemistry* 52.48, pp. 8633–8642.

Irani, M. H., Orosz, L., and Adhya, S. (1983). A control element within a structural gene: The *gal* operon of *Escherichia coli*. *Cell* 32.3, pp. 783–788.

Ireland, W. T. and Kinney, J. B. (2016). MPAthic: quantitative modeling of sequence-function relationships for massively parallel assays. *bioRxiv*, p. 054676.

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346.6216, pp. 1533–1536.

Kalli, A. and Hess, S. (2011). Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *Proteomics* 12.1, pp. 21–31.

Keseler, I. M. et al. (2013). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research* 41.D1, pp. D605–D612.

Kheradpour, P, Ernst, J, Melnikov, A, Rogov, P, Wang, L, Zhang, X, Alston, J, Mikkelsen, T. S., and Kellis, M (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research* 23.5, pp. 800–811.

Kılıç, S., White, E. R., Sagitova, D. M., Cornish, J. P., and Erill, I. (2013). CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Research* 42.D1, pp. D156–D160.

Kinney, J. B. and Atwal, G. S. (2014). Parametric Inference in the Large Data Limit Using Maximally Informative Models. *Neural Computation* 26.4, pp. 637–653.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* 107.20, 9158–9163.

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences* 110.34, pp. 14024–14029.

Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences* 104.14, pp. 6043–6048.

Kuhlman, T. E. and Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research* 38.6, e92–e92.

Laikova, O. N., Mironov, A. A., and Gelfand, M. S. (2001). Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS microbiology letters* 205.2, pp. 315–322.

Lässig, M. (2007). From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* 8.Suppl 6, S7–21.

Latif, H., Federowicz, S., Ebrahim, A., Tarasova, J., Szubin, R., Utrilla, J., Zengler, K., and Palsson, B. (2016). ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions. *bioRxiv*, p. 069021.

Lee, J. O., Cho, K.-S., and Kim, O. B. (2014). Overproduction of AcrR increases organic solvent tolerance mediated by modulation of SoxS regulon in *Escherichia coli*. *Applied Microbiology and Biotechnology* 98.20, pp. 8763–8773.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 157.3, pp. 624–635.

Li, G.-Y., Zhang, Y., Inouye, M., and Ikura, M. (2008). Structural Mechanism of Transcriptional Autorepression of the *Escherichia coli* RelB/RelE Antitoxin/Toxin Module. *Journal of Molecular Biology* 380.1, pp. 107–119.

Lomba, M. R., Vasconcelos, A. T., Pacheco, A. B. F., and Almeida, D. F. (1997). Identification of *yebG* as a DNA damage-inducible *Escherichia coli* gene. *FEMS Microbiology Letters* 156.1, 119–122.

Loomis, F. W., and Magasanik, B. (1967). Glucose-Lactose Diauxie in *Escherichia coli*. *Journal of Bacteriology* 93.4, pp. 1397–1401.

Lutz, R. and Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research* 25.6, pp. 1203–10.

Magoc, T and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27.21, pp. 2957–2963.

Maisonneuve, E. and Gerdes, K. (2014). Molecular Mechanisms Underlying Bacterial Persisters. *Cell* 157.3, 539–548.

Marbach, D. et al. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods* 9.8, pp. 796–804.

Maricque, B. B., Dougherty, J. D., and Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Research* 45, e16–e16.

Martin, R. G. and Rosner, J. L. (1997). Fis, an accessorial factor for transcriptional activation of the *mar* (multiple antibiotic resistance) promoter of *Escherichia coli* in the presence of the activator MarA, SoxS, or Rob. *Journal of Bacteriology* 179.23, pp. 7410–7419.

Melnikov, A. et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* 30.3, pp. 271–277.

Minchin, S. D. and Busby, S. J. W. (2009). Analysis of mechanisms of activation and repression at bacterial promoters. *Methods* 47.1, pp. 6–12.

Mirzaei, H. et al. (2013). Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. *Proceedings of the National Academy of Sciences* 110.9, pp. 3645–3650.

Mittler, G., Butter, F., and Mann, M. (2009). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research* 19.2, pp. 284–293.

Moran, U., Phillips, R., and Milo, R. (2010). SnapShot: Key Numbers in Biology. *Cell* 141.7, 1262–1262.e1.

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Research* 31.1, pp. 266–269.

Mustonen, V., Kinney, J., Callan, C. G., and Lassig, M. (2008). Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences* 105.34, 12376–12381.

Nishida, K., Frith, M. C., and Nakai, K. (2009). Pseudocounts for transcription factor binding sites. *Nucleic Acids Research* 37.3, 939–944.

Oehler, S., Eismann, E. R., Krämer, H., and Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal* 9.4, pp. 973–979.

Oehler, S., Alberti, S., and Müller-Hill, B. (2006). Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction. *Nucleic Acids Research* 34.2, pp. 606–612.

Okuda, S. et al. (2017). jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research* 45.D1, pp. D1107–D1111.

Ong, S. E., Blagoev, B, Kratchmarova, I, Kristensen, D. B., Steen, H, Pandey, A, and Mann, M (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics* 1.5, pp. 376–386.

Ong, S.-E. and Mann, M. (2007). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols* 1.6, pp. 2650–2660.

Overgaard, M., Borch, J., Jørgensen, M. G., and Gerdes, K. (2008). Messenger RNA interferase RelE controls *relBE* transcription by conditional cooperativity. *Molecular Microbiology* 69.4, pp. 841–857.

Overgaard, M., Borch, J., and Gerdes, K. (2009). RelB and RelE of *Escherichia coli* Form a Tight Complex That Represses Transcription via the Ribbon–Helix–Helix Motif in RelB. *Journal of Molecular Biology* 394.2, pp. 183–196.

Patwardhan, R. P. et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* 30.3, pp. 265–270.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols* 2.8, pp. 1896–1906.

Razo-Mejia, M, Boedicker, J. Q., Jones, D, DeLuna, A, Kinney, J. B., and Phillips, R (2014). Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. *Physical Biology* 11.2, p. 026005.

Rohlhill, J., Sandoval, N. R., and Papoutsakis, E. T. (2017). Sort-Seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* Growth on methanol. *ACS Synthetic Biology* 6 (8), pp. 1584–1595.

Rolfes, R. J. (2006). Regulation of purine nucleotide biosynthesis: in yeast and beyond. *Biochemical Society transactions* 34.Pt 5, pp. 786–790.

Ruiz, C. and Levy, S. B. (2010). Many chromosomal genes modulate MarA-mediated multidrug resistance in *Escherichia coli*. *Antimicrobial Agents and Chemotherapy* 54.5, pp. 2125–2134.

Rydenfelt, M., Garcia, H. G., Cox, R. S., and Phillips, R. (2014). The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS ONE* 9.12, pp. 1–31.

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27.10, pp. 946–950.

Sawitzke, J. A., Thomason, L. C., Costantino, N., Bubunenko, M., Datta, S., and Court, D. L. (2010). Recombineering: In Vivo Genetic Engineering in *E. coli*, *S. enterica*, and Beyond. *Science Direct* 421, pp. 171–199.

Schmidt, A. et al. (2016). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology* 34 (1), pp. 104–111.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18.20, 6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* 188.3, 415–431.

Semsey, S., Krishna, S., Sneppen, K., and Adhya, S. (2007). Signal integration in the galactose network of *Escherichia coli*. *Molecular Microbiology* 65.2, pp. 465–476.

Seoane, A. S. and Levy, S. B. (1995). Characterization of MarR, the repressor of the multiple antibiotic resistance (mar) operon in *Escherichia coli*. *Journal of Bacteriology* 177.12, pp. 3414–3419.

Sharon, E. et al. (2012). inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology* 30.6, pp. 521–530.

Shimada, T, Yamamoto, K, and Ishihama, A (2011). Novel Members of the Cra Regulon Involved in Carbon Metabolism in *Escherichia coli*. *Journal of Bacteriology* 193.3, pp. 649–659.

Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., and Grainger, D. C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. *Genes & Development* 28.3, pp. 214–219.

Song, S and Park, C (1997). Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *Journal of Bacteriology* 179.22, pp. 7025–7032.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16.1, pp. 16–23.

Thomason, L. C., Costantino, N., and Court, D. L. (2001). *E. coli* Genome manipulation by P1 transduction. Vol. 2. (Hoboken, John Wiley & Sons, Inc.)

Vvedenskaya, I. O., Zhang, Y., Goldman, S. R., Valenti, A., Visone, V., Taylor, D. M., Ebright, R. H., and Nickels, B. E. (2015a). Massively Systematic Transcript End Readout, "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Molecular Cell* 60.6, pp. 953–965.

Vvedenskaya, I. O., Goldman, S. R., and Nickels, B. E. (2015b). Preparation of cDNA Libraries for High-Throughput RNA Sequencing Analysis of RNA 5´ Ends. *Methods in Molecular Biology* 1276, pp. 211–228.

Wade, J. T. (2015). ChIP-Seq for Genomic-Scale Analysis of Bacterial DNA-Binding Proteins. *Prokaryotic Systems Biology* 883.Chapter 7, pp. 119–134.

Wade, J. T., Reppas, N. B., Church, G. M., and Struhl, K. (2005). Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes and Development* 19.21, 2619–2630.

Weatherspoon-Griffin, N., Yang, D., Kong, W., Hua, Z., and Shi, Y. (2014). The CpxR/CpxA Two-component Regulatory System Up-regulates the Multidrug Resistance Cascade to Facilitate *Escherichia coli* Resistance to a Model Antimicrobial Peptide. *The Journal of Biological Chemistry* 289.47, pp. 32571–32582.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., and Kegel, W. K. (2014). Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters* 113.25, pp. 1–5.

Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nature Methods* 6.5, pp. 359–362.

Xia, X. (2012). Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. *Scientifica* 2012.1, pp. 1–15.

Yamaguichi, Y. and Inouye, M. (2011). Regulation of growth and death in *Escherichia coli* by toxin–antitoxin systems. *Nature Reviews Microbiology* 9.11, 779–790.

Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011). NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry* 32.1, pp. 170–173.

Zhang, H., Susanto, T. T., Wan, Y., and Chen, S. L. (2016). Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences* 113.15, pp. 4182–4187.

Zheng, D (2004). Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research* 32.19, pp. 5874–5893.

Zheng, D., Constantinidou, C., Hobman, J. L., and Minchin, S. D. (2004). Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research* 32.19, pp. 5874–5893.

Appendix A

EXTENDED DETAILS ON SORT-SEQ DATA ANALYSIS.

This appendix contains additional details on the analysis of Sort-Seq data and model inference that is associated with Chapters 3 and 4.

Calculation of fluorescence shift upon mutation

One of the first ways we analyze the sequence data from our Sort-Seq experiment is to look at the consequence of mutations at each position on the overall fluorescence. Specifically, at each position we calculate the average fluorescence bin of mutated nucleotides and compare this to the average bin for all the sequences in the data set. Since we find that most mutations are deleterious to the binding of transcription factors or RNAP, we can use the change in fluorescence to identify regions associated with binding by repressors or activators and RNAP.

First we calculate the average bin for all the sequences in the data set. We let N_f be the total number of sequences in each bin, where f refers to the bin number (f = 1, 2, 3, and 4, for four bins). The average fluorescence bin is then given by the arithmetic average across all bins,

$$\langle f \rangle = \sum_{f=1}^{4} f \cdot p(f) = \sum_{f=1}^{4} f \cdot \frac{N_f}{\sum_{f=1}^{4} N_f}, \tag{A.1}$$

where $p(f)$ is the fraction of sequences in bin f. Note that the denominator is just the total number of sequences, $N = \sum_{f=1}^{4} N_f$, and that this average will be independent of position.

Next we need to determine the average fluorescence bin of a mutated nucleotide at each position i. Since the number of mutated nucleotides may differ at each position, we define the number of mutated nucleotides in each bin and position as $M_{f,i}$. The subscript 'f, i' is used to identify which bin f and position i are being considered. The average fluorescence bin of a mutated nucleotide can then similarly be found,

$$\langle f_{mut,i} \rangle = \sum_{f=1}^{4} f \cdot p_{mut,i}(f) = \sum_{f=1}^{4} f \cdot \frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}}, \tag{A.2}$$

where in this case, $p_{mut,i}(f)$ refers to the fraction of mutated nucleotides in bin f, and at position i.

Finally, we can now calculate the average fluorescence bin shift upon mutation, which is given by the differences in Equation A.2 and Equation A.1,

$$\langle \Delta f_{mut,i} \rangle = \langle f_{mut,i} \rangle - \langle f \rangle == \sum_{f=1}^{4} f \cdot \left(\frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}} - \frac{N_f}{\sum_{f=1}^{4} N_f} \right). \tag{A.3}$$

Note that when we plot the fluorescence bin shift for a region where we have multiple data points (i.e., from different mutated, but overlapping regions of the DNA), we plot the average calculated value of $\langle \Delta f_{mut,i} \rangle$ from the different experiments. We also note that it is possible to re-weight each bin by its mean fluorescence, $\widetilde{f}$ (i.e., instead of $f = 1, 2, 3, 4$, use the average fluorescence shift in arbitrary fluorescence units). Here we replace f with $\widetilde{f}$ in Equation A.3. For example, under situations where different sort conditions were used across experiments, this re-normalization should allow better comparison of values across experiments. The fluorescence values for $\widetilde{f}$ can be determined by regrowing the sorted cells and measuring the mean fluorescence of each sorted cell population.

Calculation of information footprints

Another way that we analyze the data from our Sort-Seq experiments is to calculate an information footprint (Kinney et al., 2010). This allows us to identify whether there are any positions along the mutagenesis window that are informative in relating sequence S and fluorescence bin f. Said differently, an informative region would be one that if given some knowledge about the sequence, we should be able to predict which fluorescence bin the promoter sequence might be found in. The mathematical way of implementing this intuition is to use the quantity known as the mutual information.

We can calculate the mutual information between sequence and fluorescence bin, $I(b_j, f)$, at each position i along the mutagenesis window by calculating the fraction of each nucleotide b_j (= A, C, G, T) found within each bin f. This allows us to estimate the joint probability distribution $p_i(b_j, f)$ at each position i. For example, $p_{10}(A, 2)$ would denote the probability that we observe an A in the second fluorescence bin at position i=10 along our promoter. The mutual information at each position is then defined by,

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log \left(\frac{p_i(b_j, f)}{p_i(b_j) p_i(f)} \right), \tag{A.4}$$

where we have summed over all nucleotides and the N_f fluorescent bins that the sequences were found in. There is also a finite sample correction that can be applied, (Treves and Panzeri, 1995), since Equation A.4 tends to overestimate the true mutual information. This is given by

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log \left(\frac{p_i(b_j, f)}{p_i(b_j) p_i(f)} \right) - \frac{(n_{b_j} - 1) \cdot (n_f - 1) \cdot log_2 e}{2 \cdot N} + O(N^{-2}), \tag{A.5}$$

where $n_{b_j} = 4$ is the number of nucleotides, and n_f is the number of bins that cells have been sorted into.

Inference of energy matrix models with Sort-Seq data.

In order to predict the influence of DNA sequence on binding of regulatory proteins, we use the Sort-Seq data to generate quantitative models of the sequence-dependent binding energy. Through a relationship between likelihood and mutual information, Kinney *et al.* (Kinney et al., 2007; Kinney et al., 2010) showed that in the large data limit it is possible to infer biophysical parameters such as the binding energies that relate the interaction between proteins and DNA sequence. In this section we describe in detail the approach used to infer energy matrices from our Sort-Seq data using Markov Chain Monte Carlo (MCMC). A full discussion of MCMC is beyond the scope of this work, but we point the interested reader to further details regarding inference using mutual information in work from Kinney *et al.* (Kinney et al., 2010; Jones et al., 2014; Atwal and Kinney, 2016). We also stress that while we make extensive use of linear energy matrix models, the inference procedure is in no way limited to such models and can be extended to allow, for example, epistatic effects through the addition of other parameters. The simple linear models, however, provide us with a useful starting point to gain insight and describe the protein-DNA interaction.

We begin with a summary of the procedure used to infer an energy matrix model using MCMC, and use the RNAP binding site of the *relB* promoter as an example. The inference was performed using the MPAthic software (Ireland and Kinney, 2016). A general schematic of the procedure is shown in Supplemental Fig. A.1. More specific details are then discussed in the following subsections. First we must initialize a 4xL set of energy parameters, $\Theta = \{\theta_{i,j}\}$, for a binding site of length L and four base pairs (see Supplemental Fig. A.1, part 1). We begin by randomly selecting parameter values for our energy matrix with which to initialize the MCMC. Here we select values from a normal distribution centered at zero with variance equal to 1, although this choice does not appear to be too critical and rather, just provides us with a starting point for our MCMC chain. Using this energy matrix we then estimate the mutual information between the binned sequences and the associated set of energy model predictions. As shown in Supplemental Fig. A.1, part 2, initially the energy matrix will be of little value in describing the observed sequence data since it was randomly chosen. This is shown by the almost uniform joint probability distribution and low mutual information in Supplemental Fig. A.1A, and Fig. A.1B.

We now begin the MCMC by perturbing the energy matrix parameters using the Metropolis-Hastings algorithm with the PyMC package in Python (Patil et al., 2010) (within the MPAthic software (Ireland and Kinney, 2016)). After each step of the chain, we re-calculate the mutual information between the data and new model predictions, which allows us to calculate how well this new set of energy matrix parameters describe the data. Dependent on whether the new energy matrix parameters lead to an improvement in mutual information, these new parameters are either retained or discarded and the process is repeated (again, according to the Metropolis-Hastings algorithm (Patil et al., 2010)). As will be discussed later, we also renormalize the matrix entries to constrain certain gauge freedoms after each iteration.

After a sufficient number of steps, and assuming that a model exists that can describe the Sort-Seq data, we will arrive at a model whose joint probability distribution between model predictions and binned sequences show a clear correlation. This is shown by the joint probability distribution in Supplemental Fig. A.1C, as well as the plateau in the mutual information trace in Supplemental Fig. A.1A, since changes to the energy matrix parameters are unable to increase the mutual information any further. In this first portion of MCMC we have performed many samplings to reach a high probability region where the energy matrix will be more representative of the distribution we are sampling from. This first step is usually referred to as the 'burn-in' period (Patil et al., 2010) and allows us to begin sampling from the distribution, $p(\Theta|data)$ (defined below), that describes the distribution of energy matrix model parameters.

Finally, now that we are sampling from the desired distribution, we can estimate energy matrix parameters just by sampling this distribution many times. This brings us to part 3 of Supplemental Fig. A.1. While the mutual information no longer shows a substantial change, the parameters of the energy matrix are continuing to be perturbed following the Metropolis-Hastings algorithm, and according to the distribution $p(\Theta|data)$. We can now estimate each entry in the energy matrix by taking the arithmetic mean of the matrix parameters across all the sampling steps. This is shown by a set of contour plots and marginalized distributions for the binding energy parameters from column five of the RNAP energy matrix (Fig. A.1D). To ensure that multiple energy minima were not present in this energy landscape, we repeated the inference procedure 20 times and used the average across all appropriate MCMC chains to estimate the energy matrix parameters. The calculated mutual information will be indifferent the particular sign of the energy matrix and adjust the energy matrices such that the wild-type sequence has a negative predicted binding energy and check that energy predictions from the energy matrices from each MCMC are correlated (keeping energy matrices that provide a Pearson correlation coefficient of 0.85 or greater across model predictions). Note that for inference of parameters using thermodynamic models, separate from these energy weight matrices, we did find the presence of multiple minima and apply a parallel tempering MCMC procedure to properly sample these distributions (described in further detail at the end of this appendix).

Using the schematic in Supplemental Fig. A.1 as our guide, the sub-sections that follow expand on the details introduced here to perform this inference procedure. In particular, we begin by describing the linear energy matrix model. We then outline the Bayesian approach taken to formally write the posterior distribution, $p(\Theta|data)$, that provides us with a relationship between the energy matrix parameters and observed sequence data. When sampling this distribution we need to estimate mutual information at each iteration of the MCMC sampling procedure, and describe how to calculate it.

Linear energy matrix models are used to describe DNA-protein interaction.

We begin by outlining the linear energy matrix model shown in Fig. A.1A that describes the binding interaction between the DNA and a DNA-binding protein. We treat each base pair position j along a binding site as contributing a certain amount to the binding energy, where the total binding energy is then the sum of the contributions from all base pairs. Mathematically the energy matrix model is described by a 4xL matrix, Θ, consisting of energy parameters $\{\theta_{ij}\}$. Here each column j of matrix parameters will represent the energies for each nucleotide $i = A, C, G$, or T $(= 1, 2, 3,$

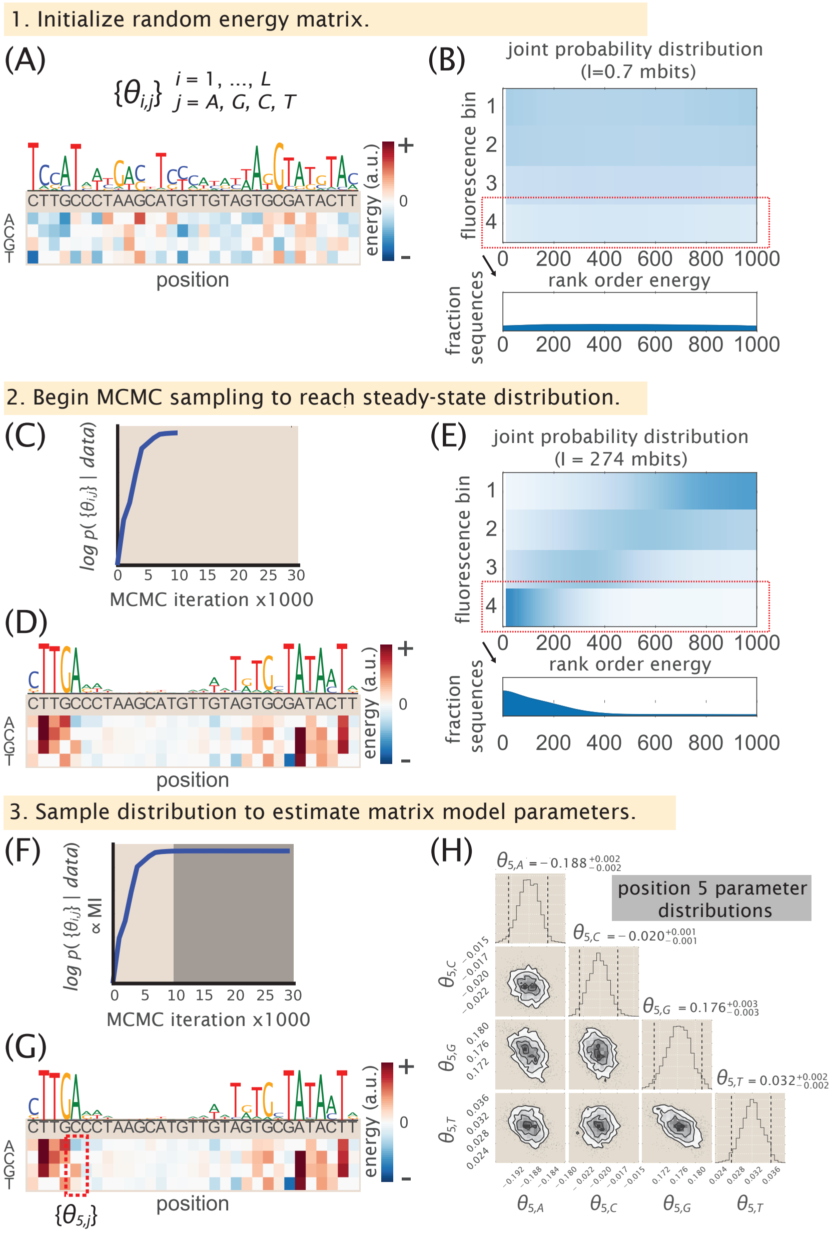


Figure A.1: **Schematic of the inference procedure used to determine energy matrices from Sort-Seq data using Markov Chain Monte Carlo.** 1. To begin the inference of a set of 4xL model parameters, $\{\theta_{ij}\}$, are chosen from a normal distribution. (A) Example set of parameters used to initialize the MCMC sampling. Matrix entries are first normalized such that energy predictions have mean of zero and standard deviation of one. For plotting energy matrices, each column has been shifted such that the wild-type sequence has zero energy. The associated sequence logo is shown above the energy matrix. *(Caption continued on next page)*

Figure A.1: *(continued from previous page)* (B) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (A), using all sequences in the *rel* promoter data set. The bottom plot shows, the histogram of rank ordered predictions of only bin four, corresponding to the red boxed region, which is nearly uniform due to the randomly chosen matrix entries used to predict energies from each sequence. Since the matrix parameters were randomly chosen, the nearly uniform distribution results in low mutual information (0.7 mbits, where 1 mbit = 10^{-3} bits) between fluorescence bin and rank order energy predictions. 2. MCMC sampling of the energy matrix model is performed using the Sort-Seq data associated with the *rel* RNAP binding site. (C) The log posterior, Eq. (A.9), is plotted for the first 1000 iterations and corresponds to the 'burn-in' period. The log posterior is proportional to the mutual information between fluorescent bin and rank order energy predictions (see Appendix A). During each sampling iteration, the parameters will be retained or discarded with some probability given by the the Metropolis-Hasting algorithm (Patil et al., 2010). (D) The energy matrix and sequence logos are shown using the set of parameters at the 1000^{th} iteration. (E) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (D). The energy matrix provides energy predictions for each sequence that clearly distributes across the sorted bins and results in much higher mutual information (274 mbits). 3. Finally, matrix parameters are estimated by continuing to sample the posterior distribution many more times and determined from a weighted average of these samples. (F) The log posterior is plotted for the entire set of MCMC iterations. The sampled model parameters during the shaded region are used to estimation each matrix entry. (G) The mean energy matrix entries from these samples are plotted. (H) Contour plots and marginalized distributions summarize the sampled values for each of the four parameters at position five of the RNAP energy matrix. Note that entries in (G) have been shifted such that the wild-type nucleotide has zero energy.

or 4) associated with position j of the binding site. For example, $\theta_{2,3}$ represents the energy parameter for nucleotide C at position 3. To make our computation of binding energies more convenient, we also represent our DNA sequence as another matrix, S, having identical dimensions, 4xL. This matrix consists of parameters $\{s_{ij}\}$, where the ij^{th} entry again corresponds to the the nucleotide identity i and sequence position j. Each parameter will have a value of 1 if it corresponds to the sequence's nucleotide identity at position j, and a value of 0 otherwise. For example, for a sequence with a C at position $j = 4$, the entry $s_{2,4} = 1$ and $s_{i=1,3,4,j=4} = 0$. The binding energy, E (also defined by $\varepsilon_{i,\text{mat}}$ in Chapter 3), of any sequence, S, will then be given by

$$E = \sum_{i=A}^{T} \sum_{j=1}^{L} \theta_{ij} \cdot s_{ij}. \tag{A.6}$$

One aspect we have not considered thus far is the scale of the energy parameter. When considering binding between between DNA and a DNA-binding protein, a statistical mechanical approach would suggest that the probability of such an event occurring will be given by the Boltzmann factor, $e^{-\varepsilon_s/(k_B T)}$ (Bintu et al., 2005a). Here ε_s is the binding energy that describes this interaction in absolute energy units (e.g., units of $k_B T$; 1 kcal/mol = 1.62 $k_B T$ at 37°C), k_B is the Boltzmann constant, and T is temperature. In relation to the binding energy, E, described by our Equation A.6 above, $\varepsilon_s = A \cdot E + B$, where the constant A scales the energy matrix into absolute energy units, while B provides an additive shift that depends on the choice of reference energy. Here, the matrix entries that are used to calculate E are 'gauge fixed' such that the mean energy in each column is set to zero and the matrix norm (or inner product) has a value of 1. Note however that when plotting each energy matrix we find it useful to shift the energy in each column such that the wild-type sequence has zero energy.

When fitting the data to a model of the form $e^{-\varepsilon_s/(k_B T)}$, the fitting procedure is unable to determine the scale factors A and B noted above. For example, in most instances we report energy values in arbitrary units. This is consequence of the fitting procedure, where in the absence of a specific thermodynamic model, there remain some scale parameters that cannot be determined (Kinney et al., 2010). This parameter insensitivity has been termed 'diffeomorphic modes' and is discussed at length in other work (Atwal and Kinney, 2016). One especially interesting aspect of this is that when considering biophysical models of regulation, diffeomorphic modes often disappear and make it possible to infer parameters that were not accessible by fitting simpler models. For the cases of repression by PurR at the *purT* promoter, or activation by CRP at the *dgoR* promoter, this allowed us to estimate binding energy in absolute energy. We discuss this further in the section on inference of thermodynamic model paramters below.

Probability distribution relating energy matrix model parameters to the Sort-Seq data.

Given our FACS-sorted sequence data, we want to find the set of energy matrix parameters that best describe the distribution of sequences across our fluorescence bins (i.e., parameters that provide binding energy predictions that describe the data as shown in Supplemental Fig. A.1C). To perform this inference we take a Bayesian approach in our analysis, and as mentioned earlier, rely on MCMC to sample from the complex distribution relating our energy matrix parameters to the sequence data. While a full discussion of Bayesian analysis is outside the scope of this section, the book, Data Analysis by Sivia and Skilling (Sivia and Skilling, 2006), and online material available from the Caltech course, *BE/Bi 103: Data analysis in the biological sciences*, taught by Justin Bois (http://bois.caltech.edu/teaching.html), are excellent resources.

Formally, we want to find the set of energy matrix parameters that maximize the probability distribution of our energy predictions (through our energy matrix model) given our Sort-Seq sequence data, $p(E|\{S, f\})$, where $\{S, f\}$ refers to our array of N sequences S and the bin f where they were found (referred to as the 'data' in the initial summary of the inference procedure). x_S is the binding energy as defined in Equation A.6. From Bayes' theorem, we can re-write this distribution as,

$$p(E|\{S, f\}) = \frac{p(\{S, f\}|E)p(E)}{p(\{S, f\})} \propto p(\{S, f\}|E)p(E), \tag{A.7}$$

where the term $p(\{S, f\}|E)$ is called the likelihood, and $p(E)$ is known as the prior and encompasses our prior knowledge on the energy matrix parameters. The denominator $p(\{S, f\})$ is known as the marginalized likelihood and acts as a normalization factor, but is unimportant for our inference.

To proceed we follow the approach of Kinney *et al.* (Kinney et al., 2007; Kinney et al., 2010). We assume a uniform prior over the energy matrix model parameters. In addition, we also assume our sequence measurements are independent. The second assumption allows us to write $p(\{S, f\}|E)$ as the product of probabilities across all sequences contained within our data set, $p(\{S, f\}|E) = \prod_{s=1}^{N} p((S_i, f_i)|E)$. This is also referred to as the error model, since by relating the binned sequence data to binding energy it must also encompass the additional noise sources from our experiment that actually led to our array of sequence data. Noise sources that might influence this include the sensitivity of the FACS GFP measurements, and the rate of mis-sorting events. Expression variability due to stochastic gene expression, differences in cell size, and plasmid copy number fluctuations are also likely to contribute. However, since these are not known exactly, Kinney *et al.* computed the likelihood by averaging over an ensemble of all possible error models. Using a uniform prior over the possible error models they found,

$$p(\{S,f\}|E) = \left\langle \prod_{s=1}^{N} p((S_i, f_i)|E) \right\rangle_{\text{all possible } p(S_i, f_i|E)} = C \cdot 2^{N\cdot(I(f,E)+\Delta)}, \tag{A.8}$$

where N is the total number of sequences considered, $I(f, E)$ is the mutual information between the observed fluorescence bins and binding energies predicted by the energy matrix for all the sequences, and C is a constant of integration that will be unimportant to us. Here, Δ is a small correction that goes to zero as N goes to infinity (Kinney et al., 2007). Inserting Equation A.8 into Equation A.7, we can write

$$p(E|\{S,f\}) \propto 2^{N\cdot I(f,E)}. \tag{A.9}$$

Here we have assumed that N is sufficiently large so that the prior (which does not scale with N), as well as the Δ term in Equation A.8 can be ignored. To reiterate in reference to our MCMC procedure (shown in Supplemental Fig. A.1), this is the probability distribution that we are sampling from to find the set of energy matrix parameters that describe our sorted sequence data set. The mutual information values shown in the plots of Fig. A.1C, F (mutual information traces in part 2 and 3) are reflected by our choice of energy matrix parameters. MCMC enables us to sample from the distribution and essentially find the set of matrix parameters that maximize this mutual information. In the next section we continue by describing how we estimate mutual information.

Estimating mutual information using the energy model predictions.

In the last section we found that the energy matrix parameters should be related to the data through Equation A.9. By performing many samples from this distribution using MCMC, it is possible to estimate the most probable energy matrix parameters, $\theta_{i,j}$, that make up our energy matrix. Here we consider how to estimate the mutual information term in Equation A.9 needed for our calculation. While a non-trivial problem in general, the following approach appears to work well in practice. In this case the fluorescence bins, f, are discrete variables while our binding energies, E, are continuous, with the mutual information given by

$$I(f,E) = \int_{E=-\infty}^{E=\infty} dE \sum_{f} p(f,E) \log_2 \frac{p(f,E)}{p(E)\cdot p(f)}. \tag{A.10}$$

In our sequence data set, we can easily estimate $p(f)$ by counting the number of sequences in each fluorescence bin. However, we do not have direct access to the probability distribution $p(E)$ *a priori*.

To proceed, we further bin our N sequences into 1000 bins, by rank ordering them by their associated binding energy predictions (using the energy matrix of the current MCMC step). This provides us with an estimate of the probability distribution in binding energy across our sequences. Specifically, this is shown for fluorescence bin 4 in Supplemental Fig. A.1B and E. While this is not a direct estimate of $p(E)$, we invoke the fact that the mutual information will be invariant under monotonic transformations ($I(f, E) = I(f, z_s)$) (Kinney et al., 2010). Therefore, instead of calculating $I(f, E)$, we instead calculate $I(f, z_s)$, where z_s is instead the ranked ordering of the N sequences.

In order to calculate the mutual information we now construct a 2-d histogram (joint distribution) by binning the rank ordered energy predictions into $z_s = 1$ to 1000 bins across each of the different fluorescence bins. We define this by the frequency matrix $F(f, z_s)$, and from our finite data set, use kernel density estimation with a kernel width equal to 4% to estimate the joint distribution. This is what is plotted in Supplemental Fig. A.1B, and E, where the mutual information is then calculated as

$$I(f, z_s)_{smooth} = \sum_{z_s=1}^{1000} \sum_{f} F(f, z_s) \log_2 \frac{F(f, z_s)}{F(z_s) \cdot F(f)}. \qquad \text{(A.11)}$$

Inference of thermodynamic model parameters using parallel tempering Markov chain Monte Carlo (MCMC).

So far, we have applied MCMC using an error-model-averaged likelihood to infer the parameters of an energy matrix. One limit initially observed by Kinney *et al.* (Kinney et al., 2010) was an inability of the fitting procedure to constrain certain parameters (due to free diffeomorphic modes, noted earlier). Interestingly however, it was found that certain diffeomorphic modes often disappear when fitting the Sort-Seq data to non-linear models. For a thorough discussion of diffeomorphic modes refer to the work of Kinney *et al.* (Kinney and Atwal, 2014). We applied this strategy in several of our data sets from the *purT*, *dgoR*, and *xylE*, where specific thermodynamic models appeared appropriate. Here we briefly outline the models used and the main results from our MCMC analysis.

We begin with the *purT* promoter. Here we identified an RNAP binding site that is repressed by PurR, which binds between the -10 and -35 RNAP sites. Given the presence of only these two binding sites, we modeled the promoter as having a simple repression architecture (Bintu et al., 2005a). Some additional complexity arises due to the presence of other PurR binding sites on the genome, and the allosteric dependence of a purine metabolite for co-repression. Following the approach of Weinert *et al.* Weinert et al., 2014, this can be quantitatively described by,

$$P_{bound} = \frac{\lambda_p e^{-\beta \varepsilon_p}}{1 + \lambda_p e^{-\beta \varepsilon_p} + \lambda_r e^{-\beta \varepsilon_r}}. \tag{A.12}$$

Here λ_p and λ_r represent the fugacity, which describes the relative availability of RNAP and PurR, respectively, to bind their binding sites. These parameters depend on the concentration of each protein (through their chemical potentials), and for PurR, will also depend on its allosteric state. ε_p and ε_r represent the binding energies of RNAP and PurR to their binding sites, respectively.

As noted in our definition of an energy matrix, we can also describe each binding energy through the gauge-fixed energy matrix prediction, which is multiplied by a scale factor and additive shift (e.g., $\varepsilon_r = A_r \cdot x_r + B_r$, where A_r is the scale factor, x_r is the energy matrix prediction, and B_r is the additive shift). To being fitting to the model described by Equation A.12, we first inferred the energy matrices for RNAP and PurR following the MCMC procedure noted above. We then performed a second MCMC to fit the remaining thermodynamic parameters. In this second MCMC we sampled using error-model-averaged likelihood against the posterior $p(P_{bound}|\{S, f\})$. This allowed us to infer the following parameters: $A_r = -11.55 k_B T$, $\lambda_r e^{-\beta B_r} = e^{0.64}$, and $A_p = 2.4 k_B T$, where A_p is the RNAP scale factor. Note that in this second MCMC, we performed parallel tempering MCMC (using the PTSampler in package emcee (Foreman-Mackey et al., 2013)) to better sample the posterior distributions of our thermodynamic parameters (see supplemental of Kinney *et al*, 2010).

Next we consider the *dgoR* promoter. While we found the promoter to be quite complex, here we use data from the JK10 strain (see Supplemental Section 4.11) where activation by CRP appeared to dominate transcription. Here we apply the model used by Kinney *et al.* (Kinney et al., 2010), which consists of a binding site for RNAP and CRP, but also includes an interaction energy between these two proteins. Again using fugacity terms to describe the availability of each protein, this will be given by,

$$P_{bound} = \frac{\lambda_p e^{-\beta\varepsilon_p} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p+\varepsilon_a+\varepsilon_i)}}{1 + \lambda_p e^{-\beta\varepsilon_p} + \lambda_a e^{-\beta\varepsilon_a} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p+\varepsilon_a+\varepsilon_i)}}. \tag{A.13}$$

In this architecture we have the fugacity λ_a for the activator CRP and its binding energy to the binding site, ε_a. In addition, there is an additional energy term ε_i that describes the interaction between RNAP and CRP. Again, we can write $\varepsilon_p = A_p \cdot x_p + B_p$. We can also write the CRP binding energy as $\varepsilon_a = A_a \cdot x_a + B_a$, where similarly, A_a is the scale factor, x_a is the gauge-fixed energy prediction, and B_a is an additive shift. Using parallel tempering MCMC to sample $p(P_{bound}|\{S, f\})$, we obtained the following values: $\varepsilon_i = -7.3k_BT$, $A_a = -13.6k_BT$, $\lambda_a e^{-\beta B_a} = e^{-1.89}$, and $A_p = -12.7k_BT$.

Lastly we consider the *xylE* promoter. This promoter contains two XylR sites which are likely bound as a dimer (Song and Park, 1997). There is also a CRP site directly upstream of the xylR sites. The binding signature of CRP is only observed for the right half of the binding site, implying the left half of the protein does not make as significant DNA contact. Since CRP still has a powerful impact on gene expression, it suggests that there is a cooperative interaction between xylR and the weak CRP site. The short distance between the xylR sites and the RNAP also suggests that there is a direct interaction between the xylR sites and the RNAP. In addition, there is also a spacing between the RNAP polymerase and the CRP site of 35 bp (approximately three helical turns of the DNA). For this spacer length in the *lac* promoter there is a expected to be a significant interaction energy even in the absence of XylR (Ushida and Aiba, 1990; Gaston et al., 1990). A thermodynamic model of RNAP polymerase binding probability for this architecture will be

$$P_{bound} = \frac{f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}{g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}, \tag{A.14}$$

where

$$\begin{aligned} f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= \lambda_p e^{-\beta\varepsilon_p} + \lambda_p \lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_{x_i})} \\ &\quad +\lambda_p \lambda_c e^{-\beta(\varepsilon_p+\varepsilon_c+\varepsilon_{c_i})} \\ &\quad +\lambda_p \lambda_c \lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_c+\varepsilon_{c_i}+\varepsilon_{x_i}+\varepsilon_{cx_i})} \qquad \text{(A.15)} \\ g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= 1 + \lambda_x e^{-\beta\varepsilon_x} + \lambda_c e^{-\beta\varepsilon_c} + \lambda_x \lambda_c e^{-\beta(\varepsilon_x+\varepsilon_c+\varepsilon_{cx_i})} \\ &\quad +\lambda_p e^{-\beta\varepsilon_p} + \lambda_p \lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_{x_i})} \\ &\quad +\lambda_p \lambda_c e^{-\beta(\varepsilon_p+\varepsilon_c+\varepsilon_{c_i})} \\ &\quad +\lambda_p \lambda_c \lambda_x e^{-\beta(\varepsilon_p+\varepsilon_x+\varepsilon_c+\varepsilon_{c_i}+\varepsilon_{x_i}+\varepsilon_{cx_i})}. \qquad \text{(A.16)} \end{aligned}$$

Here, the λ_x and ε_x terms mark the fugacity and binding energy of XylR respectively. The λ_c and ε_c represent the fugacity and binding energy of CRP, and λ_p and ε_p do the same for RNAP. The terms ε_{x_i}, ε_{c_i}, and ε_{cx_i} are interaction terms between XylR and RNAP, CRP and RNAP, and CRP and XylR, respectively.

Due to the position of the library windows (with a 60 bp window containing the two XylR binding sites, but only partial binding sites for CRP and RNAP), we were unable to fit this model to the data. The fitting procedure requires sequences with mutations throughout the multiple binding sites and further experimentation will be needed to fit and characterize the proposed model further.

References

Atwal, G. S. and Kinney, J. B. (2016). Learning Quantitative Sequence-Function Relationships from Massively Parallel Experiments. *Journal of Statistical Physics* 162.5, pp. 1203–1243.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005a). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics and Development* 15.2, pp. 116–124.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific* 125.925, pp. 306–312.

Gaston, K., Bell, A., Kolb, A., Buc, H., and Busby, S. (1990). Stringent spacing requirements for transcription activation by CRP. *Cell* 62.4, pp. 733–743.

Ireland, W. T. and Kinney, J. B. (2016). MPAthic: quantitative modeling of sequence-function relationships for massively parallel assays. *bioRxiv*, p. 054676.

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346.6216, pp. 1533–1536.

Kinney, J. B. and Atwal, G. S. (2014). Parametric Inference in the Large Data Limit Using Maximally Informative Models. *Neural Computation* 26.4, pp. 637–653.

Kinney, J. B., Tkačik, G., and Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences* 104.2, 501–506.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* 107.20, 9158–9163.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software* 35.4, pp. 1–811.

Sivia, D. and Skilling, J. (2006). Data analysis: a Bayesian tutorial. OUP Oxford.

Song, S and Park, C (1997). Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *Journal of Bacteriology* 179.22, pp. 7025–7032.

Treves, A. and Panzeri, S. (1995). The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Computation* 7.2, pp. 399–407.

Ushida, C. and Aiba, H. (1990). Helical phase dependent action of CRP: effect of the distance between the CRP site and the -35 region on promoter activity. *Nucleic Acids Research* 18.21, pp. 6325–6330.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., and Kegel, W. K. (2014). Scaling of gene expression with transcription-factor fugacity. *Physical Review Letters* 113.25, pp. 1–5.

Appendix B

USING SEQUENCE PERTURBATIONS TO PROBE SEQUENCE-FITNESS RELATIONSHIPS.

The work of this thesis has been concerned with developing quantitative descriptions of transcription in bacteria and enabling such descriptions more broadly across the bacterial genome. An important consideration in motivating such descriptions is the expectation that a cell ‘cares’ about how much of a given protein is produced. Obviously, proteins that form larger complexes will need to be produced in a regulated manner such that the proteins are available in the correct proportions (Li et al., 2014). Bacteria also do not live in a world of isolation. Rather, they are constantly competing for resources among themselves and with other organisms. Those with the greatest fitness will in general be the ones who can succeed and propagate in a population (Lässig, 2007; Poelwijk et al., 2011; Vos et al., 2013).

One of the next steps that then follows from the quantitative work on regulatory sequence is to try relating such sequences to cellular fitness. Here we briefly consider this by modifying the scheme employed by Sort-Seq. It is worth highlighting that although the data analysis associated with Sort-Seq is quite involved, the premise behind such experiments is quite straightforward: we have some region of interest on the DNA and our library provides a large set of small perturbations; we then sort that library in a way that should reflect the changes to regulation that underly the region under study. With Sort-Seq we expect changes to the regulatory DNA to influence gene expression and therefore sort the library of cells by their expression level (using a GFP fluorescence reporter).

In order to probe the effect of regulatory DNA on cellular fitness, here we instead place our mutated promoter library on a plasmid that drives expression of a gene that will influence cellular fitness (see schematic in Fig. B.1). Here the gene could intentionally be chosen to enforce some level of selection on the cells (e.g. *sacB* in the presence of sucrose (Poelwijk et al., 2011)), or be a native gene whose effect on cellular fitness is under investigation. Instead of sorting our library of cells by fluorescence, we instead grow the population of cells over many generations and sequence at different time intervals. The change in the *distribution* of our sequence library should reflect whether or not the particular sequences had any influence on

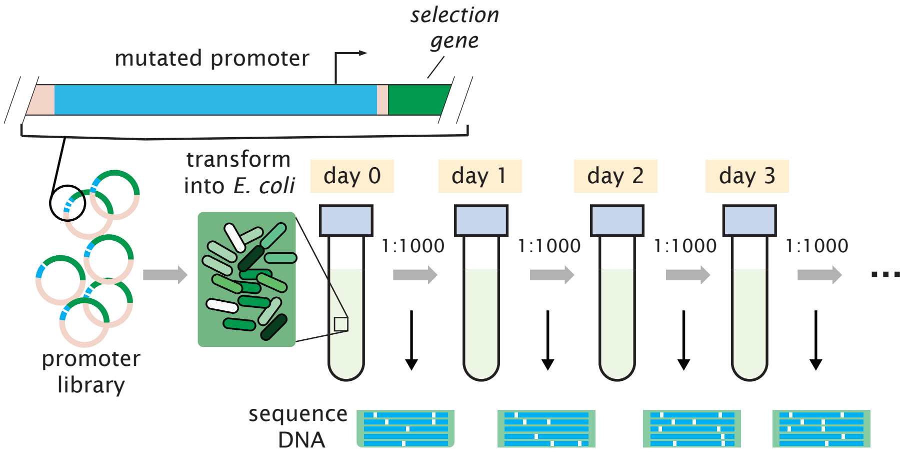


Figure B.1: **Schematic of fitness assay (fit-seq).** A promoter library drives expression of a selection gene, placed on a low copy-number plasmid (5-10 copies per cell). The regulatory sequence under investigation is randomly mutated at a rate of approximately 10%. The library is transformed into *E. coli* cells such that each cell contains a different mutated sequence. Under the current protocol, repeated 1:1000 dilutions are performed each day, with cells allowed to reach saturation. After cells have reached saturation, the plasmid library is collected by miniprepping the plasmid DNA and prepared for sequencing.

cellular fitness within the population. We can use the change in distribution over time to probe the relationship between regulatory sequence and fitness. For simplicity we will define this approach as fit-seq.

Fitness effects of the *rel* promoter and toxin-antitoxin genes, *relB* and *relE*.

We explore this approach with the *rel* promoter, which we also investigated by Sort-Seq in Chapter 4. The *rel* promoter is of interest because it drives expression of a toxin-antitoxin system (RelE-RelB) that is among about 30 such systems found on the *E. coli* chromosome, and whose physiological role is not totally understood. These systems are capable of causing major growth inhibition or even cell death when the toxin is in excess, and they have been implicated in the formation of persister cells (Li et al., 2008; Yamaguichi and Inouye, 2011). The promoter itself is repressed by the antitoxin RelB, which binds to the promoter while in complex with RelE. The ratio of antitoxin to toxin is roughly 10:1 in the cell, but this can also be modulated through degradation of RelB by proteases such as Lon (Gotfredsen and Gerdes, 1998; Overgaard et al., 2008; Gerdes and Maisonneuve, 2012).

Here we constructed two different promoter libraries and placed them into a strain of *E. coli* where the entire *relBE* operon was deleted from the chromosome. The first library consisted of the *rel* promoter driving the expression of the native *relB* and *relE* genes, exactly as they are found on the chromosome. In the second library, the *rel* promoter instead drove expression of *gfp* (matching the library used in Sort-Seq).

The libraries were transformed into *E. coli* and grown to saturation following a 1:1000 dilution in M9 minimal media with 0.5% glucose. In this particular experiment, the 1:1000 dilution was repeated over three days, after cells had reached saturation, and the libraries were miniprepped at day 0 (before transformation), day 1, and day 3 for sequencing.

The first way we can analyze this data is to plot the effect of mutations relative to the wild-type promoter sequence over time. Analogous to the expression shift plots from Sort-Seq, we calculate the average time point that we find a mutated sequence relative to the wild-type sequence. A negative value would suggest that mutating that position is harmful to the cell's fitness, while a positive value would suggest that the cell's fitness is enhanced by the mutation. In Fig. B.2 we perform such an analysis of the two libraries. Note that the axes have been inverted relative to how an expression shift plot is usually shown. Interestingly, we find very different behavior dependent on the downstream gene(s) being expressed, though in both cases the dominant features align with the known regulatory binding sites quite well.

When expressing the native *relBE* genes, we find that mutating the promoter appears generally detrimental to the cell's growth (Fig. B.2(A)). One aspect that provides some validation of the data is that when we mutate the start codon of the antitoxin gene, *relB*, we see a negative effect on those cells in the population (see positions 1-3 in Fig. B.2(A)). In the absence of translation of the *relB* gene we would expect reduced or no growth due to an increased amount of free toxin, RelE. The *rel* promoter therefore appears to be stabilized through expression of its toxin-antitoxin genes, *relB* and *relE*. Further work will be needed to properly understand what is happening as the promoter is mutated that provides this stability.

For the library driving expression of the *gfp* gene, shown in Fig. B.2(B), we find that in contrast, most mutations actually appear beneficial to whether such sequences are observed at a later time. This is particularly the case in the region where RNAP binds the promoter, suggesting that there may actually be some fitness disadvantage associated with expression of GFP over the course of this experiment.

While the above data is quite intriguing, it is only qualitative and it is difficult to extract any clear meaning in the absence of a specific model relating the regulatory

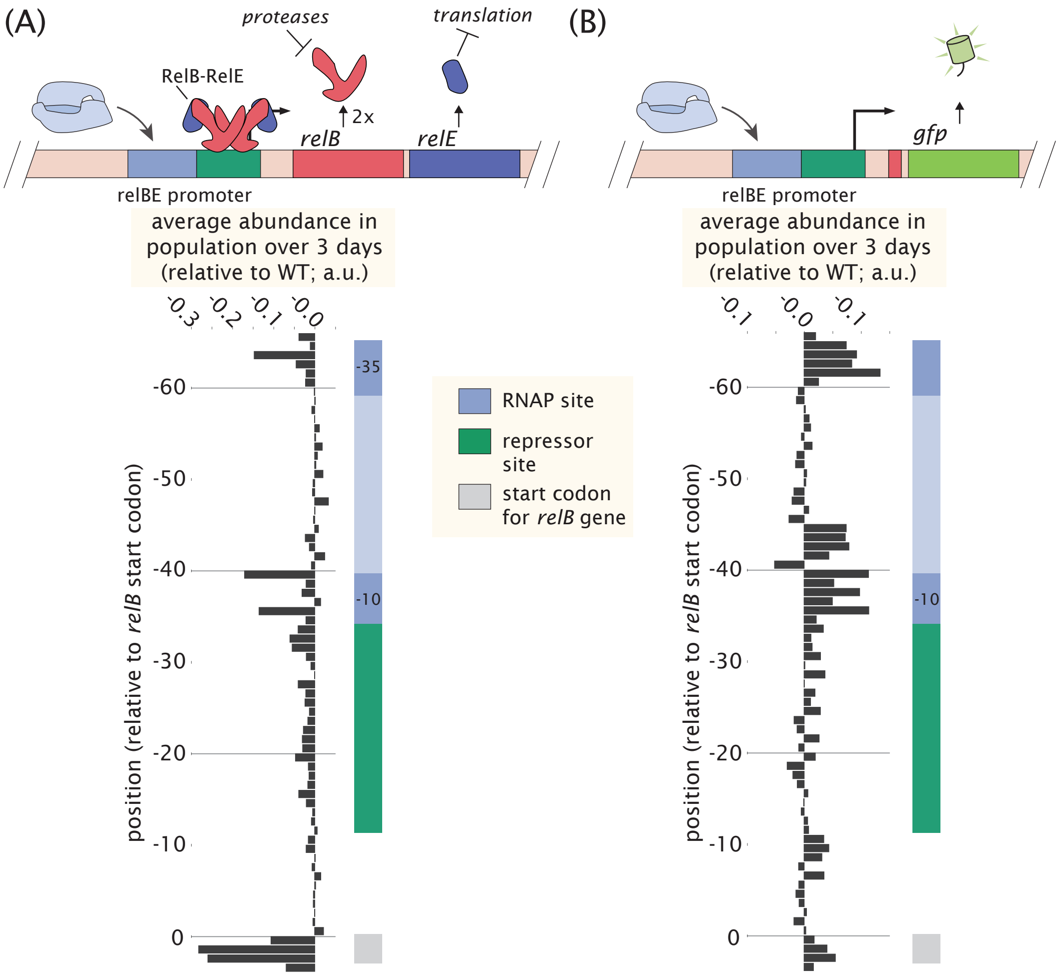


Figure B.2: **Fit-seq of the *rel* promoter.** (A) A *rel* promoter library was placed upstream of the native *relBE* genes and placed on a low copy-number plasmid. (B) A *rel* promoter library was placed upstream of *gfp* and placed on a low copy-number plasmid. A portion of the *relB* gene was still present on this second library and matched the library used for Sort-Seq. Both libraries were sequenced before transformation (day 0), after one day of growth (day 1), and after three days of growth (day 3). Plots were calculated identically to expression shift plots, but instead of calculating the average bin, the average time point was calculated for mutated sequences relative to the wild-type sequence. Both experiments were performed in MG1655 cells with the *relBE* operon deleted.

DNA to fitness or growth rate. One question that does arise is whether the observed changes in sequence distribution over time are due to disruption of binding of regulatory proteins, or whether something completely different is going on that we are unaware of. If it is due to the perturbation to the regulatory architecture, it should be possible to extract models that reflect the sequence specificity of the specific proteins involved. Here we considered this by performing an identical model

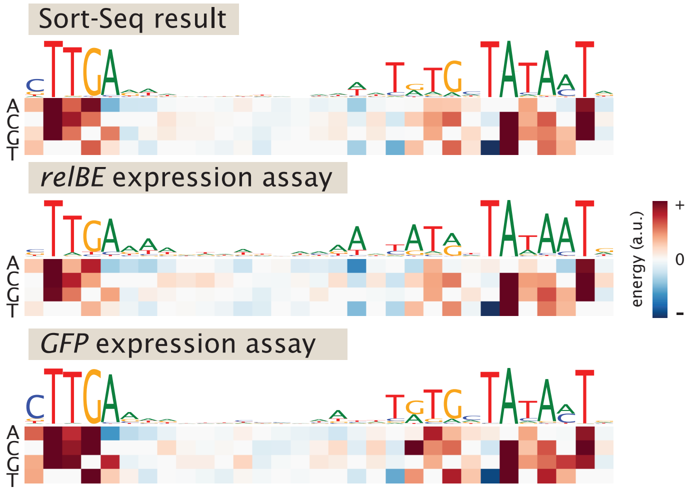


Figure B.3: **Inference of energy matrices for RNAP at the *rel* promoter**. Energy matrices for RNAP were inferred using the fit-seq data for the *rel* promoter, driving expression of either *relBE* or *gfp*. Data analysis was performed identically to Sort-Seq, except that instead of different bins, sequences represented different time points. The top matrix is what was obtained by Sort-Seq. The second row is the matrix obtained using data from the library driving expression of *relBE*, while the third matrix is from the library driving expression of *gfp*. Note that for the library driving expression of *gfp*, the sign of the matrix had to be multiplied by -1 to match the other two matrices.

inference procedure that was performed with Sort-Seq data, and inferred energy matrices across the region where RNAP binds the promoter (using time points as our 'bins'). Indeed, as shown in Fig. B.3, it was possible to infer energy matrices for binding by RNAP that are nearly identical to those obtained by Sort-Seq. It is especially impressive that while expression of either *relBE* and *gfp* appeared to have very different effects on the distribution of promoter libraries (one whose mutations appeared detrimental, while the other appeared to improve fitness), both could be used to infer a model for binding by RNAP that are almost identical.

References

Gerdes, K. and Maisonneuve, E. (2012). Bacterial Persistence and Toxin-Antitoxin Loci. *Annu. Rev. Microbiol.* 66.1, pp. 103–123.

Gotfredsen, M. and Gerdes, K. (1998). The *Escherichia coli relBE* genes belong to a new toxin-antitoxin gene family. *Molecular Microbiology* 29.4, 539–548.

Lässig, M. (2007). From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* 8.Suppl 6, S7–21.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 157.3, pp. 624–635.

Li, G.-Y., Zhang, Y., Inouye, M., and Ikura, M. (2008). Structural Mechanism of Transcriptional Autorepression of the *Escherichia coli* RelB/RelE Antitoxin/Toxin Module. *Journal of Molecular Biology* 380.1, pp. 107–119.

Overgaard, M., Borch, J., Jørgensen, M. G., and Gerdes, K. (2008). Messenger RNA interferase RelE controls *relBE* transcription by conditional cooperativity. *Molecular Microbiology* 69.4, pp. 841–857.

Poelwijk, F. J., Vos, M. G. J. de, and Tans, S. J. (2011). Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell* 146.3, pp. 462–470.

Vos, M. G. de, Poelwijk, F. J., and Tans, S. J. (2013). Optimality in evolution: new insights from synthetic biology. *Current Opinion in Biotechnology* 24.4, pp. 797–802.

Yamaguichi, Y. and Inouye, M. (2011). Regulation of growth and death in *Escherichia coli* by toxin–antitoxin systems. *Nature Reviews Microbiology* 9.11, 779–790.