Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction

Manuel Razo-Mejia[1,†], **Stephanie L. Barnes**[1,†], **Nathan M. Belliveau**[1,†], **Griffin Chure**[1,†], **Tal Einav**[2,†], **Mitchell Lewis**[3], **Rob Phillips**[1,4,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125, USA

[2]Department of Physics, California Institute of Technology, Pasadena, CA, 91125, USA

[3]Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA

[4]Department of Applied Physics, California Institute of Technology, Pasadena, CA, 91125, USA

[†]Contributed equally

*Lead contact: phillips@pboc.caltech.edu

Abstract

Allosteric regulation is found across all domains of life, yet we still lack simple, predictive theories that directly link the experimentally tunable parameters of a system to its input-output response. To that end, we present a general theory of allosteric transcriptional regulation using the Monod-Wyman-Changeux model. We rigorously test this model using the ubiquitous simple repression motif in bacteria by first predicting the behavior of strains that span a large range of repressor copy numbers and DNA binding strengths and then constructing and measuring their response. Our model not only accurately captures the induction profiles of these strains but also enables us to derive analytic expressions for key properties such as the dynamic range and $[EC_{50}]$. Finally, we derive an expression for the free energy of allosteric repressors which enables us to collapse our experimental data onto a single master curve that captures the diverse phenomenology of the induction profiles.

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. 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. 2014b). Building upon previous work (Garcia and Phillips 2011a; 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. 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. 1(B). By combining this MWC treatment of induction with a thermodynamic model of transcriptional regulation (Fig. 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. 2005b) 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 2011a; Garcia et al. 2011b; 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. 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.

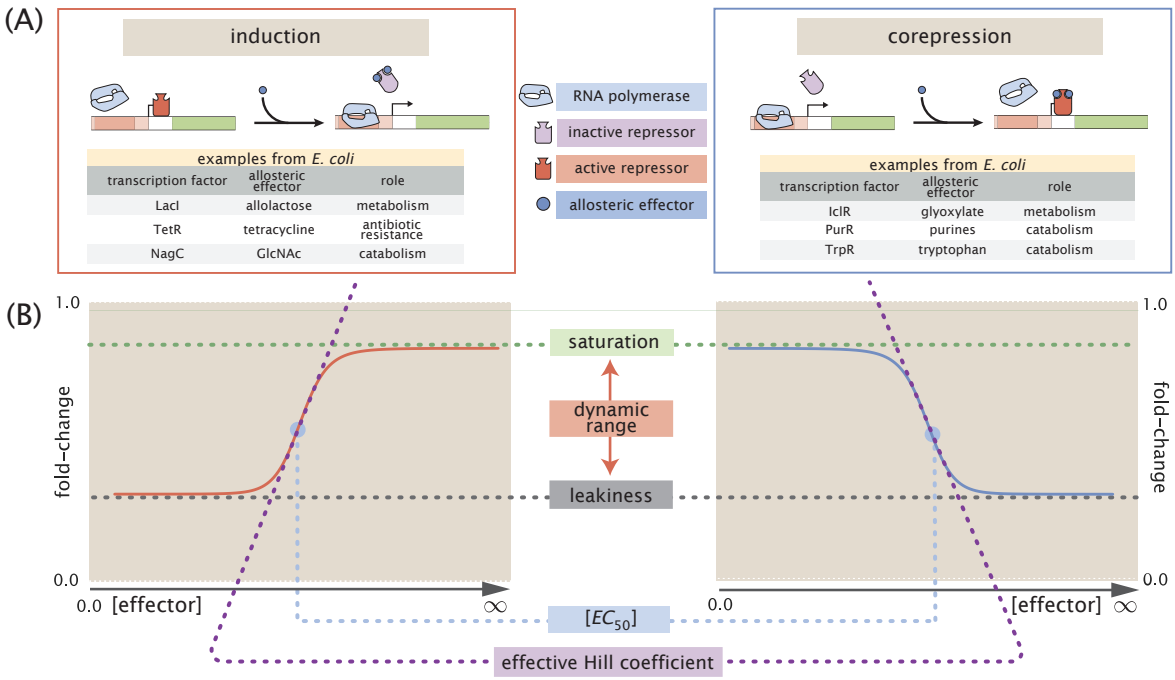

Figure 1. Transcription regulation architectures involving an allosteric repressor. (A) We consider a promoter regulated solely by an allosteric repressor. When bound, the repressor prevents RNAP from binding and initiating transcription. Induction is characterized by the addition of an effector which 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).

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 2011a; 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(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. 2005b; Bintu et al. 2005a; Kuhlman et al. 2007; Daber et al. 2011; Garcia and Phillips 2011a; 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

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

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,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \tag{2}$$

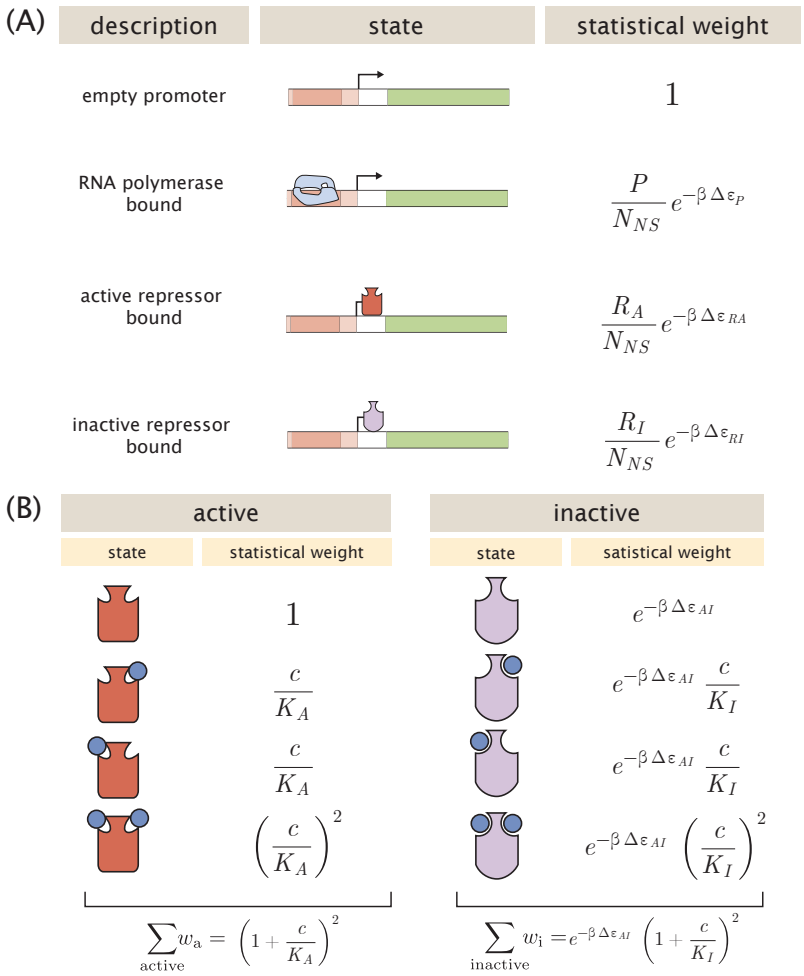

Figure 2. States and weights for the simple repression motif. (A) RNAP (light blue) and a repressor compete for binding to a promoter of interest. There are R_A repressors in the active state (red) and R_I repressors in the inactive state (purple). The difference in energy between a repressor bound to 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.

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_BT (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

$$\text{fold-change} \approx \left(1 + \frac{R_A}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1} \equiv \left(1 + p_A(c) \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}, \tag{3}$$

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(B)), $p_A(c)$ is given by the sum of the weights of the active states divided by the sum of the weights of all possible states, namely,

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}, \tag{4}$$

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(B).

Substituting $p_A(c)$ from Eq. (4) into Eq. (3) yields the general formula for induction of a simple repression regulatory architecture (Phillips 2015), namely,

$$\text{fold-change} = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}. \tag{5}$$

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. 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. 2005b; Brewster et al. 2014; Weinert et al. 2014).

This key formula presented in Eq. (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. 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 Appendix A 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 Appendix B).

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. 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 in the absence of inducer (Garcia and Phillips 2011a). Rydenfelt et al. (2014a) extended 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 Appendix A 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 Appendix, 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. S1). After fixing $\Delta\varepsilon_{AI}$ as described in the Appendix, 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 (2011a) and Brewster et al. (2014).

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 (2011a) 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 (2011a). 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 (2011a); 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 2011a). Additionally, we were able to obtain the value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ by fitting to previous data as discussed in Appendix A. 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

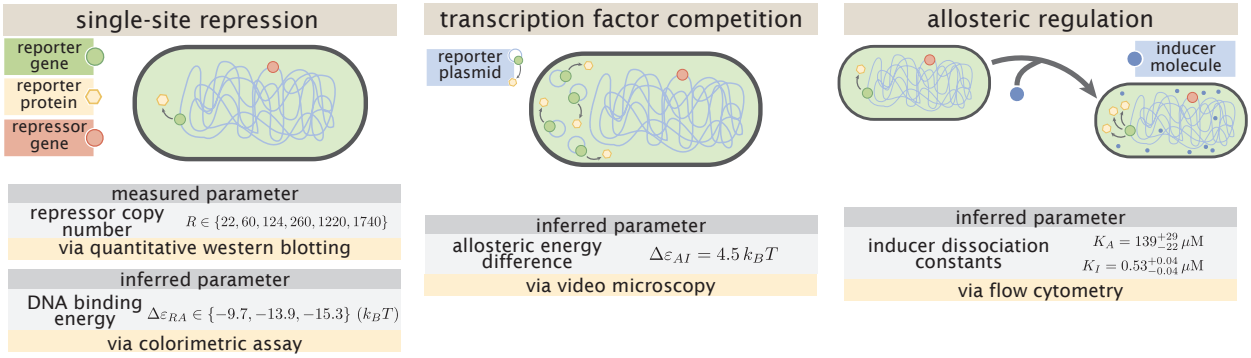

Figure 3. Understanding the modular components of induction. Over the past decade, we have refined both our experimental control over and theoretical understanding of the simple repression architectures. A first round of experiments used colorimetric assays and quantitative Western blots to investigate how single-site repression is modified by the repressor copy number and repressor-DNA binding energy (Garcia and Phillips 2011a). 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 Appendix A). 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.

measured parameters within Eq. (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. 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 Appendix C). 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 Appendix D). 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.

Determination of the *in vivo* MWC Parameters

The three parameters that we tune experimentally are shown in Fig. 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 Appendix A). 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 2011a) shown in Fig. 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 95$^{\text{th}}$ percentile of the parameter value distributions as depicted in Fig. 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 Appendix B. 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. 5(C)-(E)) together with the specific

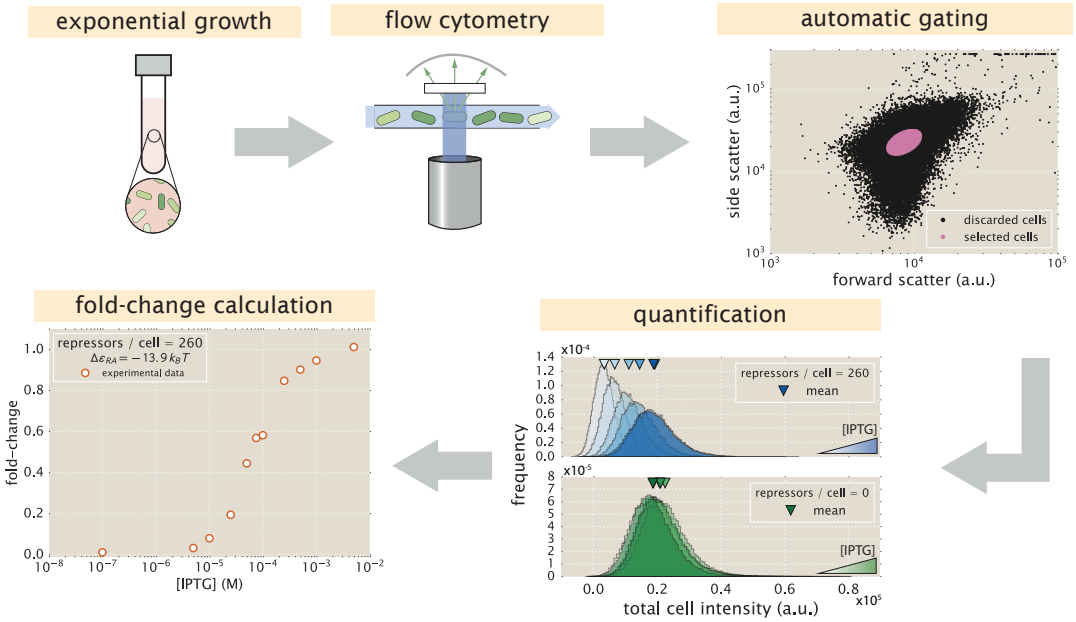

Figure 4. An experimental pipeline for high-throughput fold-change measurements. Cells are grown to exponential steady state and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single cells (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.

phenotypic properties described in Fig. 1 and discussed in detail below (see Fig. 5(F)-(J)). The shaded regions in Fig. 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. (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 Appendix E).

We stress that the entire suite of predictions in Fig. 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. 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 Appendix F). 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).

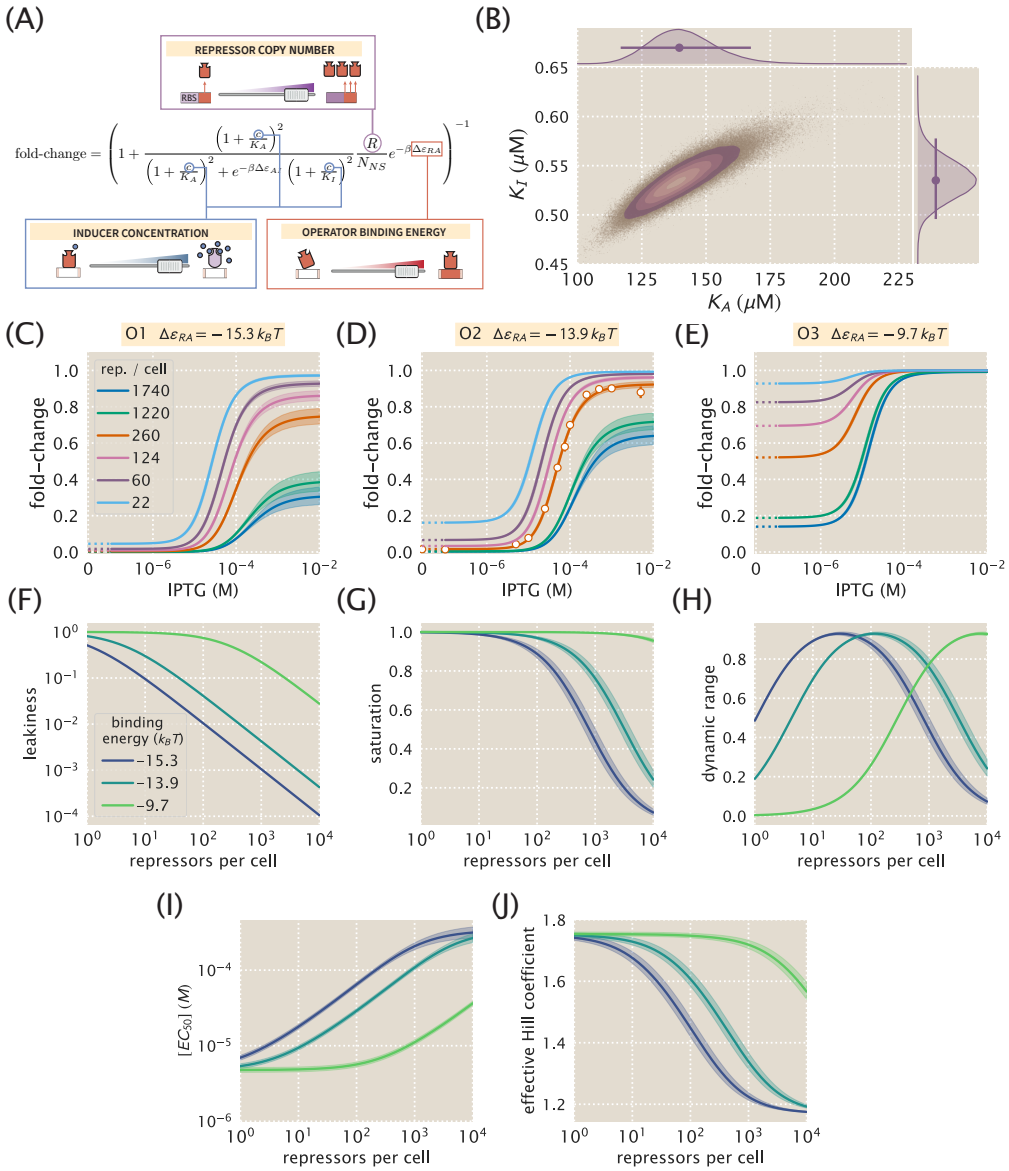

Figure 5. Predicting induction profiles for different biological control parameters. (A) We can quantitatively tune R via ribosomal binding site (RBS) modifications, $\Delta\varepsilon_{RA}$ by mutating the operator sequence, and c by adding different amounts of IPTG to the growth medium. (B) Previous experiments have characterized the R, N_{NS}, $\Delta\varepsilon_{RA}$, and $\Delta\varepsilon_{AI}$ parameters (see Fig. 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). The remaining solid lines predict the fold-change Eq. (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. 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.

Comparison of Experimental Measurements with Theoretical Predictions

We tested the predictions shown in Fig. 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 (2011a), and inducer concentrations spanning several orders of magnitude. With a few provocative exceptions, the results shown in Fig. 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. 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 Appendix G). 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 2011a), 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 Appendix H 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 Appendix I and Fig. 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, the repressor copy numbers R, and the repressor DNA binding energy $\Delta\varepsilon_{RA}$ (see Appendix G), 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.

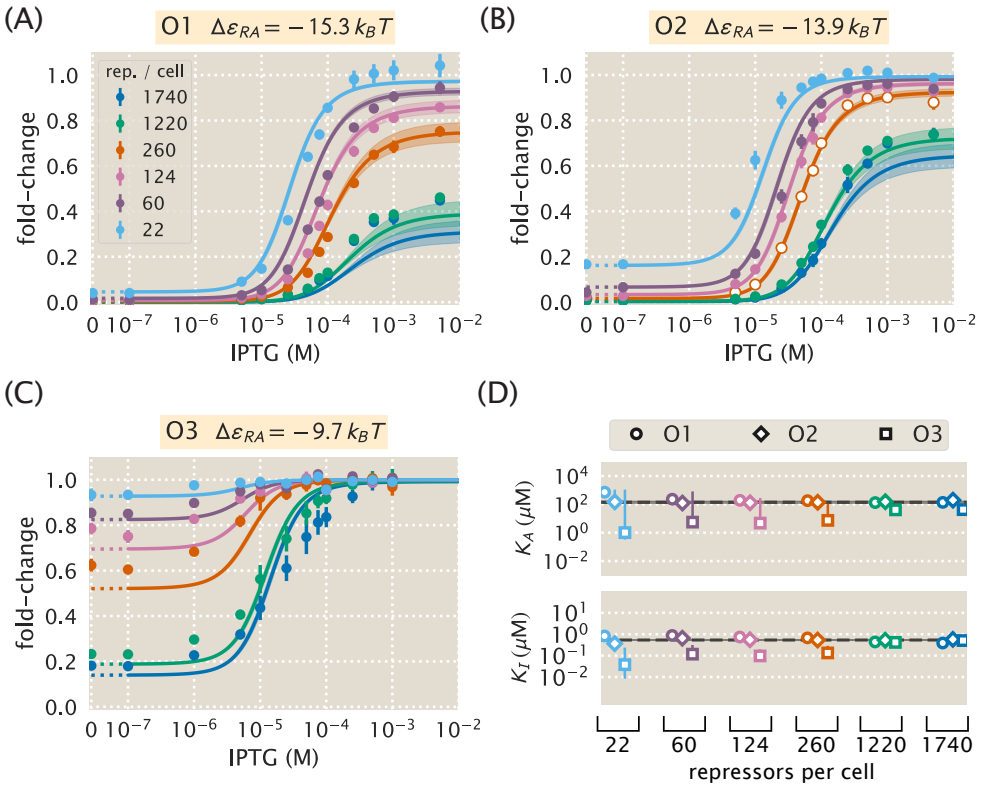

Figure 6. Comparison of predictions against measured and inferred data. Flow cytometry measurements of fold-change over a range of IPTG concentrations for (A) O1, (B) O2, and (C) O3 strains at varying repressor copy numbers, overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Fig. 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.

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. 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. 5(F)-(J).

We define these five phenotypic traits using expressions derived from the model, Eq. (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

$$\begin{aligned}\text{leakiness} &= \text{fold-change}(c = 0) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1},\end{aligned} \tag{6}$$

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

$$\begin{aligned}\text{saturation} &= \text{fold-change}(c \to \infty) \\ &= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}.\end{aligned} \tag{7}$$

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

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \tag{8}$$

These three properties are shown in Fig. 5(F)-(H). We discuss these properties in greater detail in Appendix J. 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. 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.

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. (5) halfway between its minimum and maximum value,

$$\text{fold-change}(c = [EC_{50}]) = \frac{\text{leakiness} + \text{saturation}}{2}. \tag{9}$$

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

$$h = \left(2\frac{d}{d\log c}\left[\log\left(\frac{\text{fold-change}(c) - \text{leakiness}}{\text{dynamic range}}\right)\right]\right)_{c=[EC_{50}]}. \tag{10}$$

Fig. 5(I)-(J) shows how the $[EC_{50}]$ and effective Hill coefficient depend on the repressor copy number. In Appendix J, we discuss the analytic forms of these two properties as well as their dependence on the repressor-DNA binding energy.

Fig. 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. (5) to each individual titration curve and computing the $[EC_{50}]$ and effective Hill coefficient using Eq. (9) and Eq. (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 Appendix F, 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 G and I). 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.

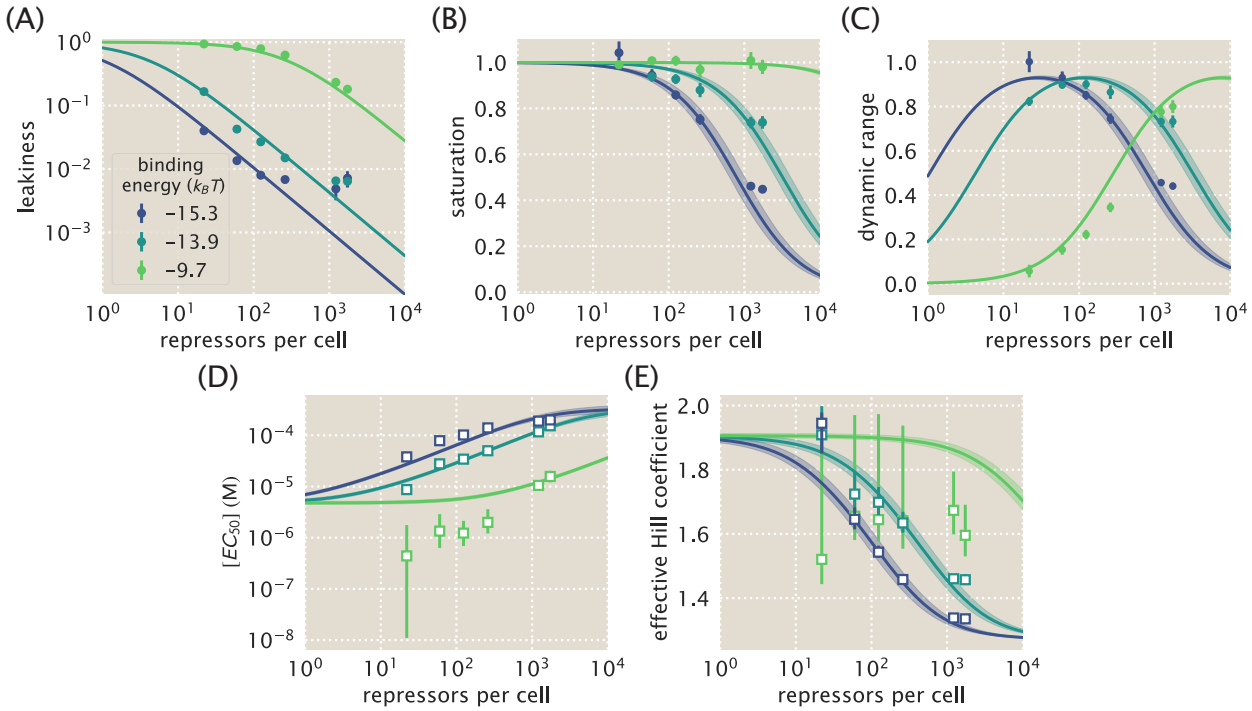

Figure 7. Predictions and experimental measurements of key properties of induction profiles. Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 6 in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-operator 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. 6(A)-(C) separately to Eq. (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. (9) and (10).

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. (5) as a Fermi function,

$$\text{fold-change} = \frac{1}{1 + e^{-F(c)}}, \tag{11}$$

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

$$F(c) = \frac{\Delta\varepsilon_{RA}}{k_BT} - \log \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n} - \log \frac{R}{N_{NS}}. \tag{12}$$

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. 8(A) demonstrates how the various induction curves from Fig. 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. 8(B) shows this data collapse for the 216 data points in Fig. 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. (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.

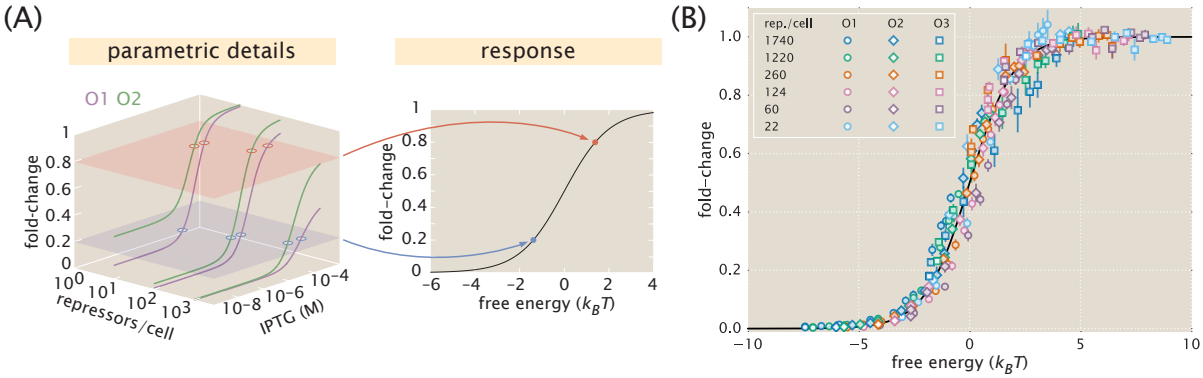

Figure 8. Fold-change data from a broad collection of different strains collapse onto a single master curve. (A) Any combination of parameters can be mapped to a single physiological response (i.e. fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from Fig. 6 collapse onto a single master curve as a function of the free energy Eq. (12). The free energy for each strain was calculated from Eq. (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.

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. 2005b; Garcia and Phillips 2011a), 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. 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. 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. 6(D) and Appendix I). 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. 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 G), 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. (2011b), 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. 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. (6) and (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 C and H), as well as quantitative Western blots (Garcia and Phillips 2011a) 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 2011a; Brewster et al. 2014) (see Appendix B) 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. (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. 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. (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. 2005b). We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli* (see Appendix K). 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.

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 (2011a) and Garcia et al. (2011b). 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. 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 $P_{LtetO\text{-}1}$ 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 (2011a). 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 Appendix L.

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 (2011a), 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 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 (2011a) 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 Appendix Table S1. 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.

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

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{13}$$

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

$$P(K_A, K_I \mid D) = \frac{P(D \mid K_A, K_I)P(K_A, K_I)}{P(D)}, \tag{14}$$

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, which we can ignore in our parameter estimation. Eq. (5) assumes a deterministic relationship between the parameters and the data, so in order to construct a probabilistic relationship as required by Eq. (14), we assume that the experimental fold-change for the i^{th} datum given the parameters is of the form

$$\text{fold-change}^{(i)}_{\text{exp}} = \left(1 + \frac{\left(1 + \frac{c^{(i)}}{K_A}\right)^2}{\left(1 + \frac{c^{(i)}}{K_A}\right)^2 + e^{-\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c^{(i)}}{K_I}\right)^2} \frac{R^{(i)}}{N_{NS}} e^{-\beta\Delta\varepsilon^{(i)}_{RA}}\right)^{-1} + \epsilon^{(i)}, \tag{15}$$

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

$$P(D|K_A, K_I, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\text{fold-change}^{(i)}_{\text{exp}} - \text{fold-change}(K_A, K_I, R^{(i)}, \Delta\varepsilon^{(i)}_{RA}, c^{(i)}))^2}{2\sigma^2}\right], \tag{16}$$

where $\text{fold-change}^{(i)}_{\text{exp}}$ is the experimental fold-change and $\text{fold-change}(\cdots)$ 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. (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 $10\,\mu$M to $1\,\mu$M leads to an equivalent increase in affinity as a change from $1\,\mu$M to $0.1\,\mu$M. 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

$$P(\tilde{k}_A, \tilde{k}_I, \sigma) \equiv \frac{1}{(\tilde{k}^{\text{max}}_A - \tilde{k}^{\text{min}}_A)} \frac{1}{(\tilde{k}^{\text{max}}_I - \tilde{k}^{\text{min}}_I)} \frac{1}{\sigma}. \tag{17}$$

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 as 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

This work 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. 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). Nathan Belliveau is a Howard Hughes Medical Institute International Student Research fellow.

Competing interests

The authors have declared that no competing interests exist.

Author contributions

MRM, SB, NB, GC, and TE contributed equally to this work. MRM, SB, NB, GC performed experiments. TE and MRM laid groundwork for the model. MRM, SB, NB, GC, and TE performed the data analysis. MRM, GC, NB, and SB wrote code used for all experimental analysis and parameter estimation. GC made the figures for the main text and GC, MRM, SB, and NB made figures for the supplemental information. MRM, SB, NB, GC, TE, and RP wrote the paper. ML and RP provided useful insight and advice in designing and executing the work.

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., Kuhlman, T., and Phillips, R. (2005a). Transcriptional regulation by the numbers: applications. *Current Opinion in Genetics & Development* 15.2, pp. 125–135.

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

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, e1002811.

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. *PNAS* 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.

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.

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.

Garcia, H. G. and Phillips, R. (2011a). 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. (2011b). Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. *Biophysical Journal* 101.3, pp. 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.

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.

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. 6043–8.

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.

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.
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., Cox, R. S., Garcia, H., and Phillips, R. (2014a). Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Physical Review E* 89 (1), p. 012702.
Rydenfelt, M., Garcia, H. G., Cox, R. S., and Phillips, R. (2014b). 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.
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.

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.

Supporting Information for Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction

Manuel Razo-Mejia[1,†], **Stephanie L. Barnes**[1,†], **Nathan M. Belliveau**[1,†], **Griffin Chure**[1,†], **Tal Einav**[2,†], **Mitchell Lewis**[3], **Rob Phillips**[1,4,*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125, USA

[2]Department of Physics, California Institute of Technology, Pasadena, CA, 91125, USA

[3]Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA

[4]Department of Applied Physics, California Institute of Technology, Pasadena, CA, 91125, USA

[†]Contributed equally

*Correspondence: phillips@pboc.caltech.edu

Contents

A Inferring Allosteric Parameters from Previous Data

The fold-change profile described by Eq. (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.

A.1 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. 5(C) to the fold-change Eq. (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 Appendix G.

In Fig. S1(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. (5) but with $\beta\Delta\varepsilon_{RA}$ modified to the form $\beta\Delta\tilde{\varepsilon}_{RA}$ in Eq. (S4) to account for the underlying assumptions used when fitting previous data (see Appendix A.2 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. S1(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. S1(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).

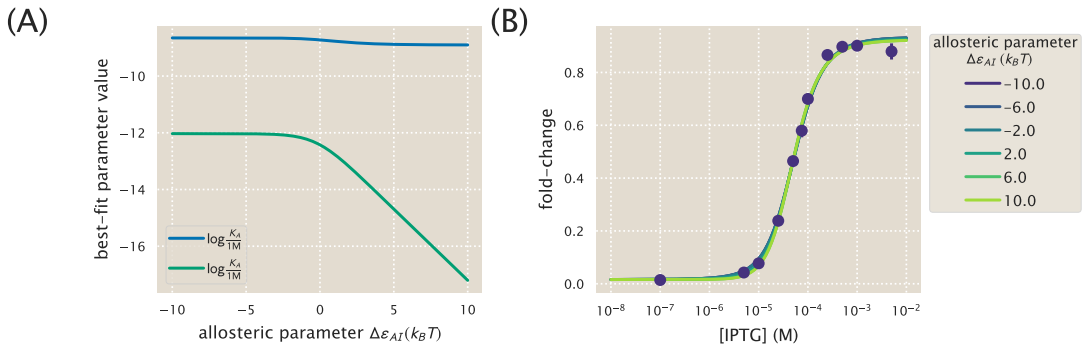

Figure S1. Multiple sets of parameters yield identical fold-change responses. (A) The data for the O2 strain ($\Delta\varepsilon_{RA} = -13.9\ k_BT$) with $R = 260$ in Fig. 5(C) was fit using Eq. (5) with $n = 2$. $\Delta\varepsilon_{AI}$ is forced to take on the value shown on the x-axis, while the 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.

A.2 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. (5) for the induction of the Lac repressor. Specifically, we consider two previous sets of work from: (1) Garcia and Phillips (2011a) 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,

$$\text{fold-change}(c = 0) = \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\tilde{\varepsilon}_{RA}}\right)^{-1}. \tag{S1}$$

Note that this functional form does not exactly match our fold-change Eq. (5) in the limit $c = 0$,

$$\text{fold-change}(c = 0) = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{S2}$$

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,

$$\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \tag{S3}$$

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

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log\left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right). \tag{S4}$$

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. (S2) 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 (2011a) 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

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{S5}$$

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

$$R_{\text{tot}} = R_S + R_{NS}, \tag{S6}$$

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

$$R_S = N \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{S7}$$

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

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{S8}$$

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. (S7) and (S8) into the modified Eq. (S6) yields the form

$$p_A(0) R_{\text{tot}} = \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \left(N \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \tag{S9}$$

where we recall from Eq. (S4) 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. (S5) 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. S2(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. (S9) occurs near $p_A(0)R_{\text{tot}} = N$ (as shown by the triangles in Fig. S2(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 $\Delta\varepsilon_{AI} \gtrsim 5\ k_BT$, 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. S2(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. (S5), we extracted the parameter value $\Delta\varepsilon_{AI} = 4.5\ k_BT$. Substituting this value into Eq. (S3) 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 (2011a) and Brewster et al. (2014) were accurate.

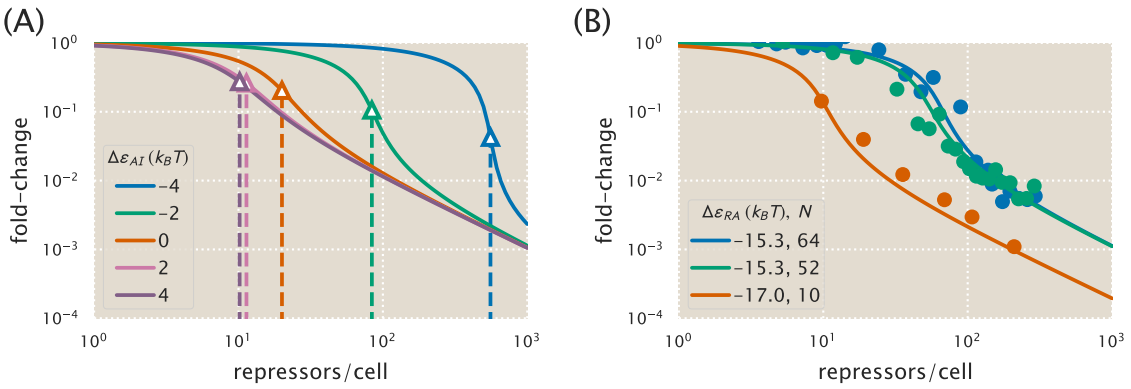

Figure S2. Fold-change of multiple identical genes. (A) In the presence of $N = 10$ identical promoters, the fold-change Eq. (S5) depends strongly on the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. The vertical dotted lines represent the 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.

B 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.

B.1 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 Appendix A, we can write the fold-change Eq. (2) in the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{S10}$$

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

$$R_{\text{tot}} = R_S + R_{NS} + R_C, \tag{S11}$$

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

$$R_S = N_S \frac{\lambda_r e^{-\beta \Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_{RA}}}, \tag{S12}$$

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

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{S13}$$

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

$$R_C = N_C \frac{\lambda_r e^{-\beta \Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta\varepsilon_C}}, \tag{S14}$$

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. (S11) 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. (4)),

$$p_A(c) = \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}. \tag{S15}$$

Substituting in Eqs. (S12)-(S14) into the modified Eq. (S11) yields the form

$$p_A(c)R_{\text{tot}} = N_S\frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1+\lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS}\frac{\lambda_r}{1+\lambda_r} + N_C\frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1+\lambda_r e^{-\beta\Delta\varepsilon_C}}. \tag{S16}$$

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. (S10) 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. (S16) 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.

B.2 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. 6). Here we predict the induction profiles for similar strains in which R is varied, but $N_S > 1$, as shown in Fig. S3. 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.

B.3 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. S4). 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.

B.4 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 regulated by a transcription factor that has multiple targets. This is highly relevant, as the majority of transcription factors in *E. coli* have

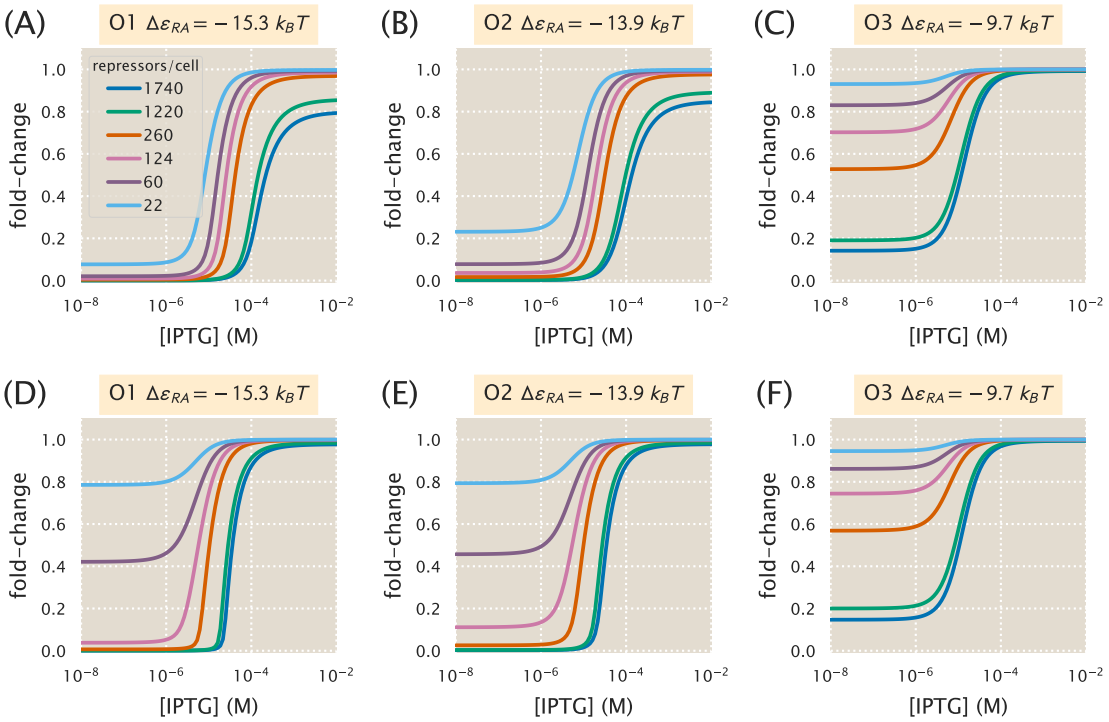

Figure S3. Induction with variable *R* and multiple specific binding sites. Induction profiles are shown for strains with variable R and $\Delta\varepsilon_{RA} = -15.3$, -13.9, or $-9.7\ k_BT$. (A-C) The number of specific sites, N_S, is held constant at 10 as R and $\Delta\varepsilon_{RA}$ are varied. (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.

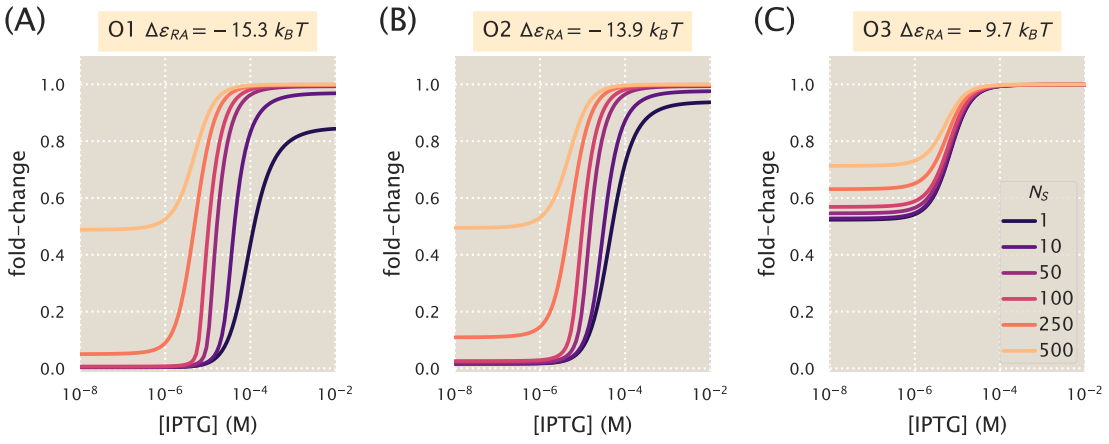

Figure S4. Induction with variable specific sites and fixed *R*. Induction profiles are shown for strains with $R = 260$ and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$. The number of specific sites N_S is varied from 1 to 500.

at least two known binding sites, with approximately 50 transcription factors having more than ten known binding sites (Rydenfelt et al. 2014b; 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.

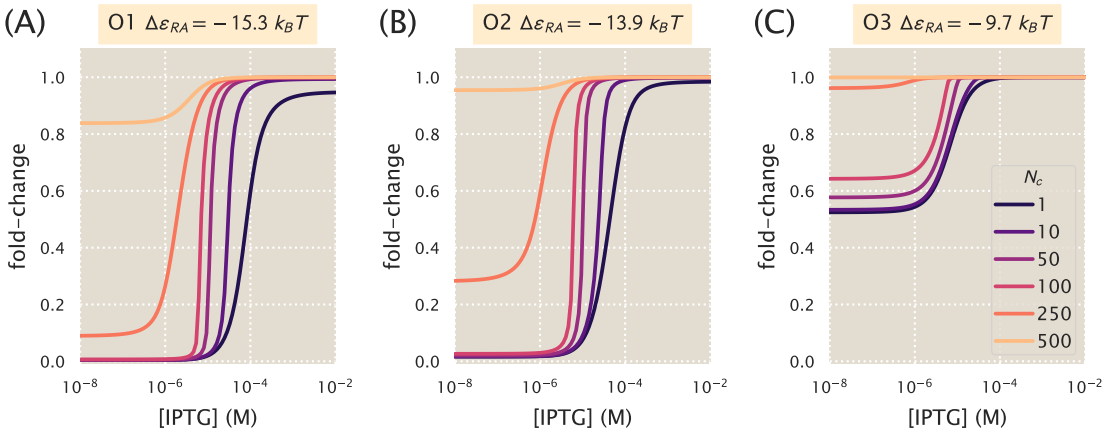

Figure S5. Induction with variable competitor sites, a single specific site, and fixed R. Induction profiles are shown for strains with $R = 260$, $N_s = 1$, and (A) $\Delta\varepsilon_{RA} = -15.3\ k_BT$ for the O1 operator, (B) $\Delta\varepsilon_{RA} = -13.9\ k_BT$ for the O2 operator, or (C) $\Delta\varepsilon_{RA} = -9.7\ k_BT$ 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.

B.5 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. 1) by numerically solving Eq. (S16) in the absence of inducer, $c = 0$, and in the presence of saturating inducer $c \to \infty$. Using Eq. (S15), the former case is given by

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}, \tag{S17}$$

whereupon substituting in the value of λ_r into Eq. (S10) will yield the leakiness. Similarly, the limit of saturating inducer is found by determining λ_r from the form

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^2} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}. \tag{S18}$$

In Fig. S6 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.

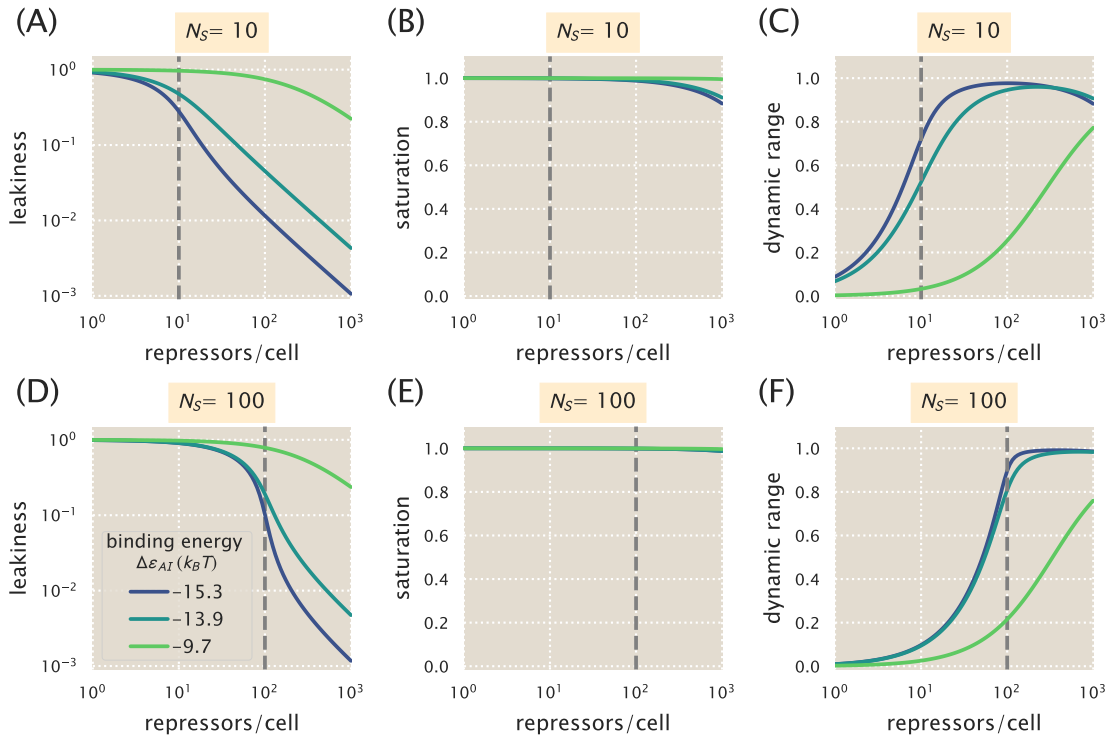

Figure S6. Phenotypic properties of induction with multiple specific binding sites. The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with number of specific binding sites $N_S = 10$ (A-C) or $N_S = 100$ (D-F). The dashed vertical line indicates the point at which $N_S = R$.

In Fig. S7 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. S6, with sharper transitions at $R = N_C$ due to the greater binding strength 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.

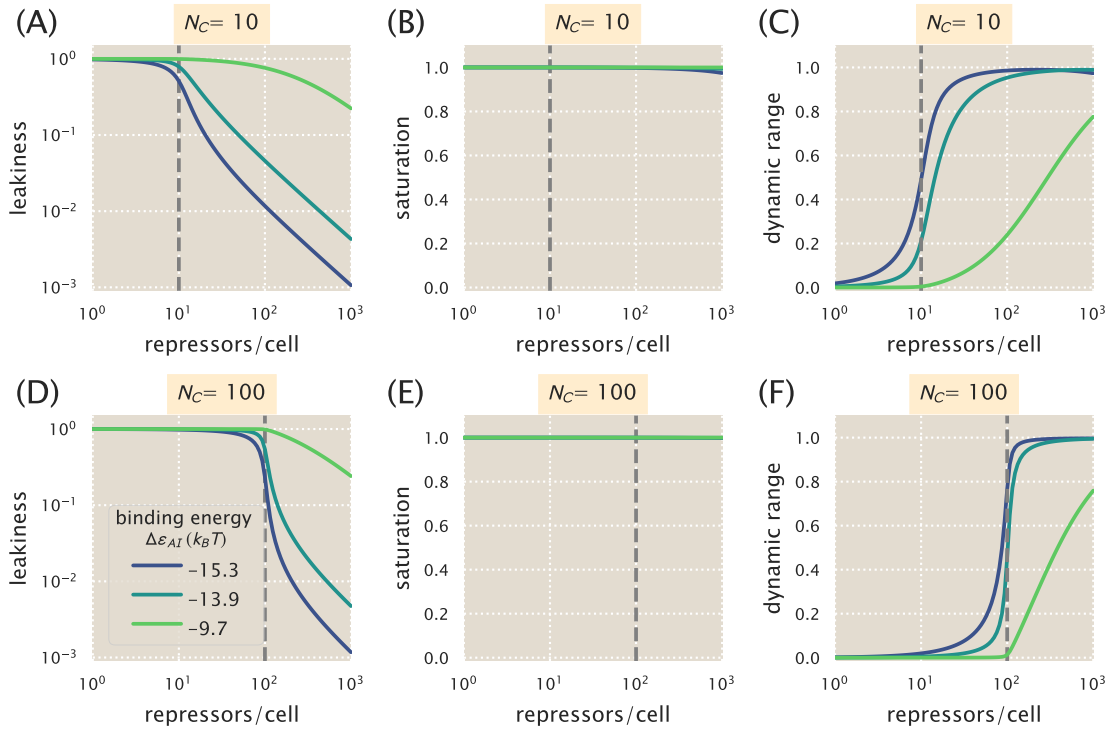

Figure S7. Phenotypic properties of induction with a single specific site and multiple competitor sites. The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with a single specific binding site $N_S = 1$ and a number of competitor sites $N_C = 10$ (A-C) or $N_C = 100$ (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 $N_C = R$.

C 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.

C.1 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 S1.

Table S1. Instrument settings for data collection using the Miltenyi Biotec MACSQuant flow cytometer. All experimental measurements were collected using these values.

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

C.2 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. S8(A) and Fig. S8(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.

C.3 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.

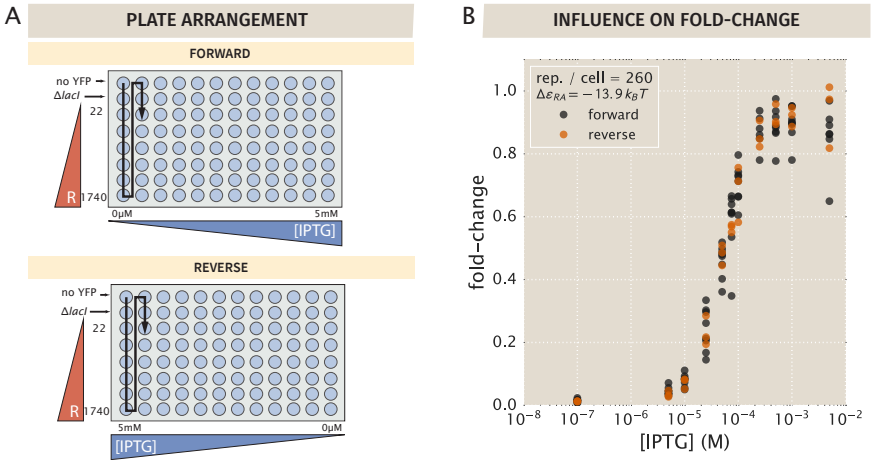

Figure S8. Plate arrangements for flow cytometry. (A) Samples were measured primarily in the forward arrangement with a subset of samples measured in reverse. The black arrow indicates the order in which samples were processed by the flow cytometer. (B) The experimentally measured fold-change values 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.

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

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_\alpha(p), \tag{S19}$$

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. S9 shows an example of different gating contours that would arise from different values of α in Eq. (S19). 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 Appendix D 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.

C.4 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. 2011b; 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 2011a). 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. S10 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

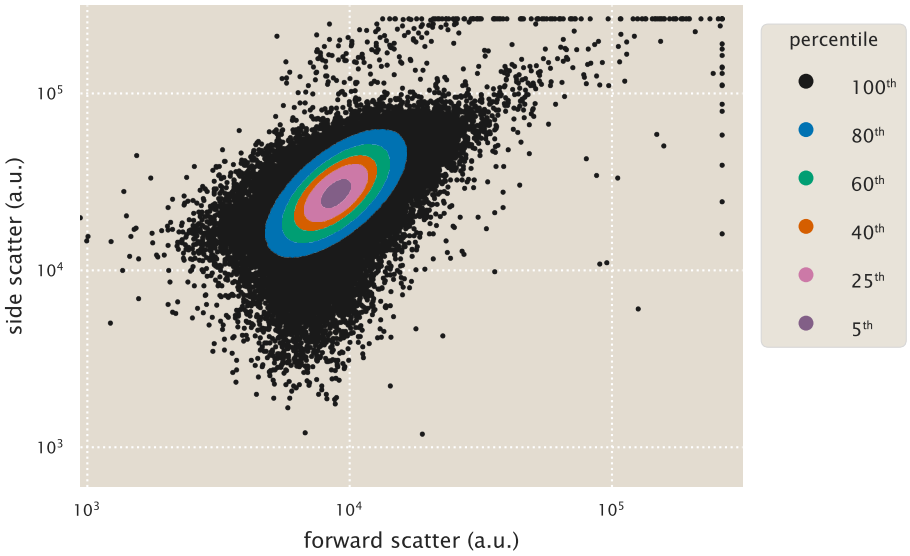

Figure S9. Representative unsupervised gating contours. Points indicate individual flow cytometry measurements of forward scatter and side scatter. Colored points indicate arbitrary gating contours ranging from 100% ($\alpha = 1.0$) to 5% ($\alpha = 0.05$). All measurements for this work were made computing the mean fluorescence from the 40^{th} percentile ($\alpha = 0.4$), shown as orange points.

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.

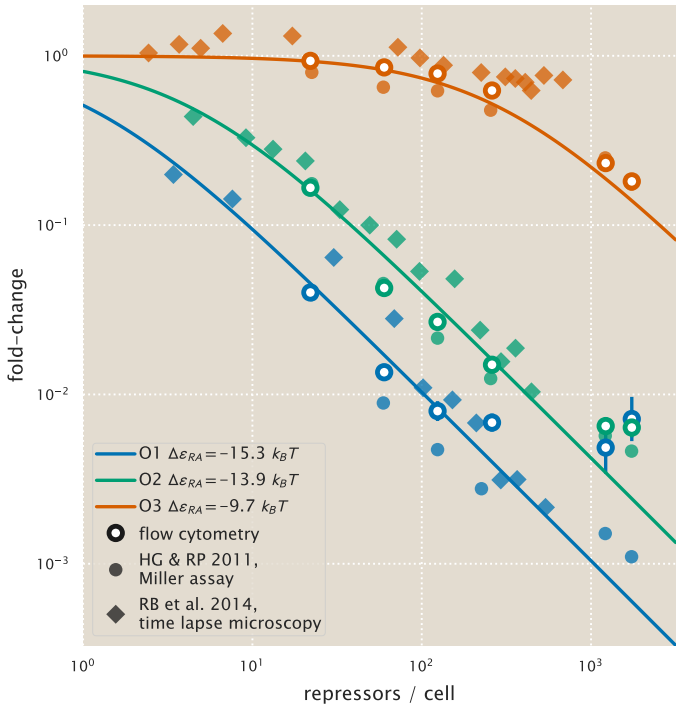

Figure S10. Comparison of experimental methods to determine the fold-change. The fold-change in gene expression for equivalent simple-repression constructs has been determined using three independent methods: flow cytometry (this work), colorimetric Miller assays (Garcia and Phillips 2011a), and video 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 (2011a).

D 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 2011a) 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. S11.

D.1 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 $100\,\mu\text{g}\cdot\text{mL}^{-1}$ spectinomycin in a deep-well 96 well plate at 37°C. These cultures were then diluted 1000-fold into $500\,\mu$L 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.

D.2 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 $500\,\mu$L aliquot was sandwiched between two glass coverslips and was allowed to solidify.

Once solid, the agarose substrates were cut into approximately $10\,\text{mm} \times 10\,\text{mm}$ 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 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.

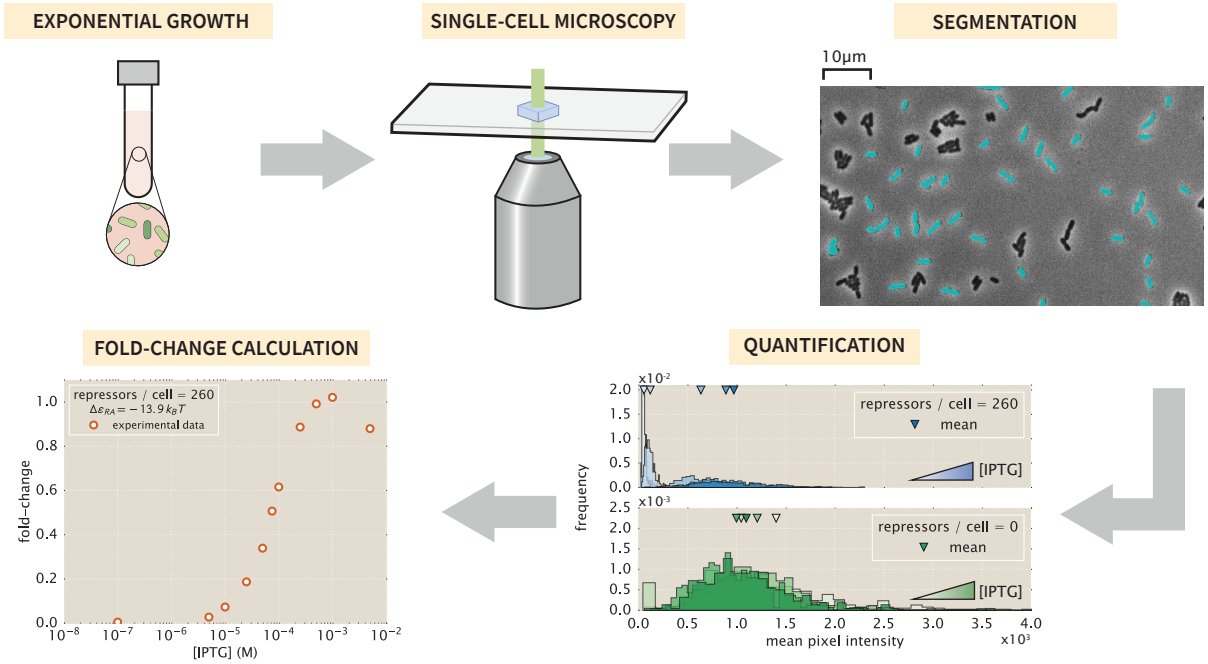

Figure S11. Experimental workflow for single-cell microscopy. For comparison with the flow cytometry results, the cells were grown in an identical manner to those described in the main text. Once cells had reached mid to late exponential growth, the cultures were diluted and placed on agarose substrates 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.

D.3 Image Processing

D.3.1 Correcting Uneven Illumination

The excitation laser has a two-dimensional gaussian profile. To minimize non-uniform 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

$$I_{\mathrm{flat}} = \frac{I - I_{\mathrm{dark}}}{I_{\mathrm{YFP}} - I_{\mathrm{dark}}} \langle I_{\mathrm{YFP}} - I_{\mathrm{dark}} \rangle, \tag{S20}$$

where I_{flat} is the renormalized image and I is the original fluorescence image. An example of this correction can be seen in Fig. S12.

D.3.2 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

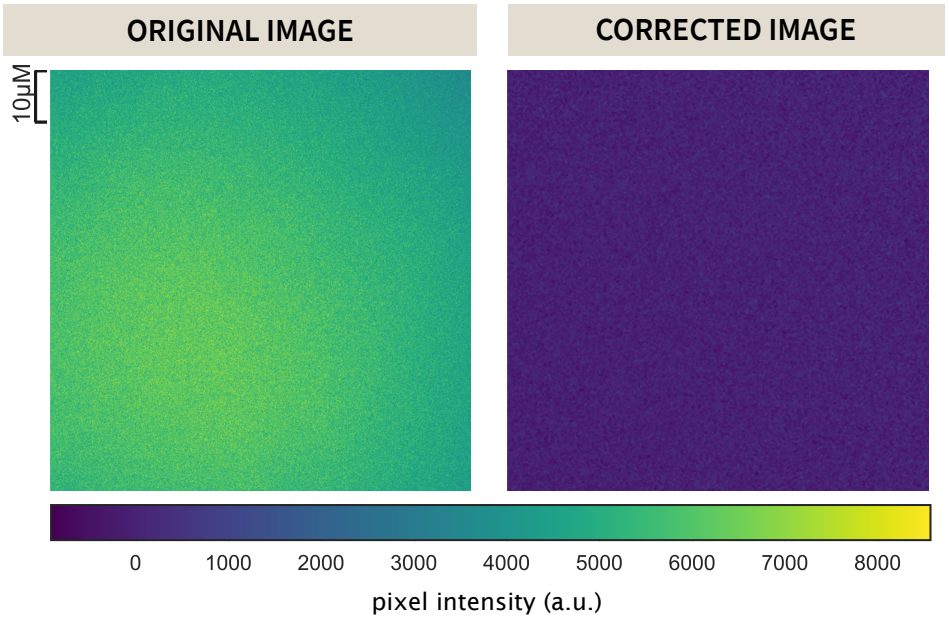

Figure S12. Correction for uneven illumination. A representative image of the illumination profile of the 512 nm excitation beam on a homogeneously fluorescent slide is shown in the left panel. This is corrected for using equation Eq. (S20) and is shown in the right panel.

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

$$\text{eccentricity} = \sqrt{1 - \left(\frac{b}{a}\right)^2}, \tag{S21}$$

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. S13(A)) and eccentricity filter (Fig. S13(B)), the remaining objects were considered cells of interest (Fig. S13(C)) and the mean fluorescence intensity of each cell was extracted.

D.3.3 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

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{S22}$$

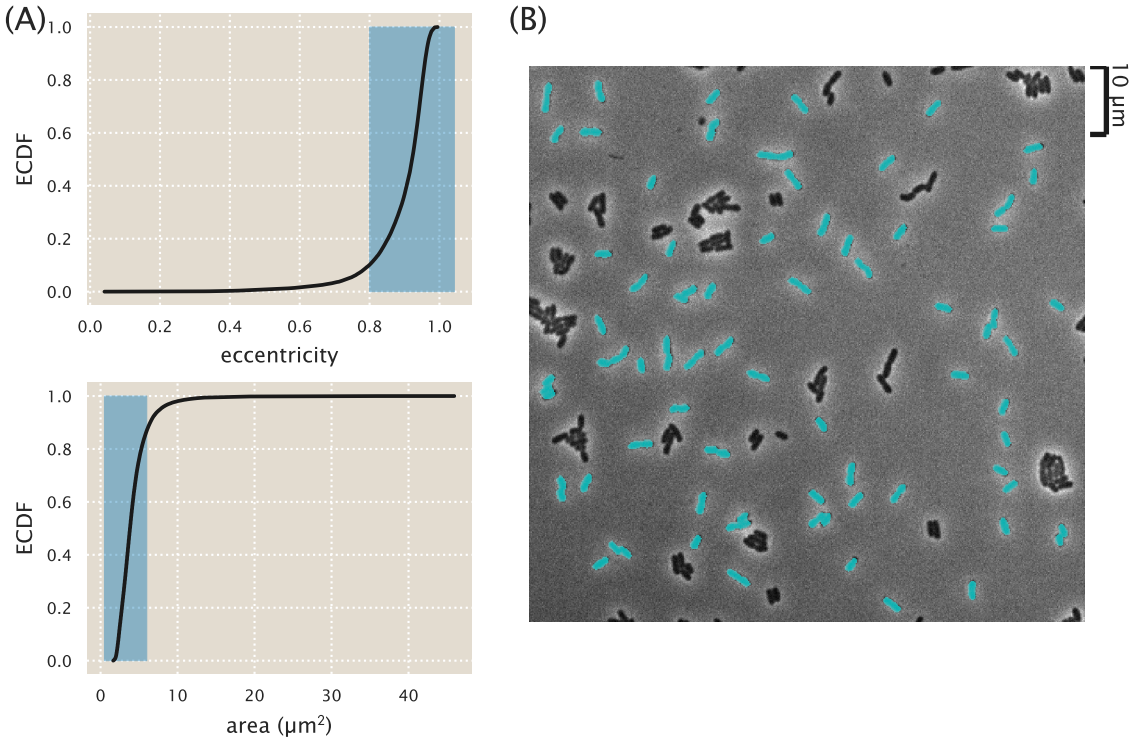

Figure S13. Segmentation of single bacterial cells. (A) Objects were selected if they had an eccentricity greater than 0.8 and an area between $0.5\,\mu\text{m}^2$ and $6\,\mu\text{m}^2$. Highlighted in blue are the regions considered to be representative of single cells. The black lines correspond to 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.

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, 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.

D.4 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. S14(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 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. S14(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. S14(B) are much broader than those from flow cytometry due to the fewer number of replicates performed.

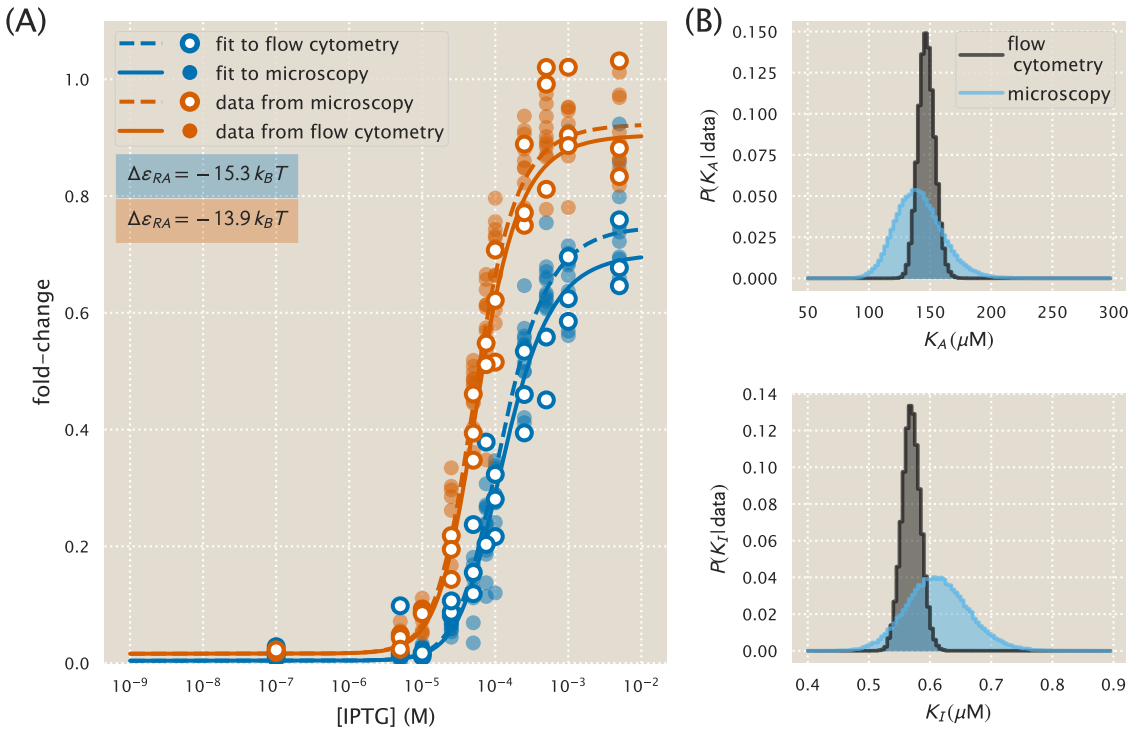

Figure S14. Comparison of measured fold-change between flow cytometry and single-cell microscopy. (A) Experimentally measured fold-change values obtained through single-cell microscopy and flow cytometry are shown as white filled and solid colored circles, respectively. Solid and dashed lines indicate 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 2011a). (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).

E Fold-Change Sensitivity Analysis

In Fig. 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. (7). While it is straightforward to consider the width of the credible regions at any other inducer concentration, Fig. 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

$$\Delta\,\text{fold-change}_{K_A} \equiv \text{fold-change}(K_A, K_I^{\text{fit}}) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}), \tag{S23}$$

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

$$\Delta\,\text{fold-change}_{K_I} = \text{fold-change}(K_A^{\text{fit}}, K_I) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}) \tag{S24}$$

to measure the sensitivity of the fold-change to K_I at a fixed K_A^{fit}. Fig. S15 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

$$\frac{\partial\,\text{fold-change}}{\partial K_A} = \frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n-1}}{\left(1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \tag{S25}$$

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

$$\frac{\partial\,\text{fold-change}}{\partial K_I} = -\frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n}}{\left(1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \tag{S26}$$

From Eqs. (S25) and (S26), 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. S15.

Lastly, we note that Eqs. (S25) and (S26) 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.

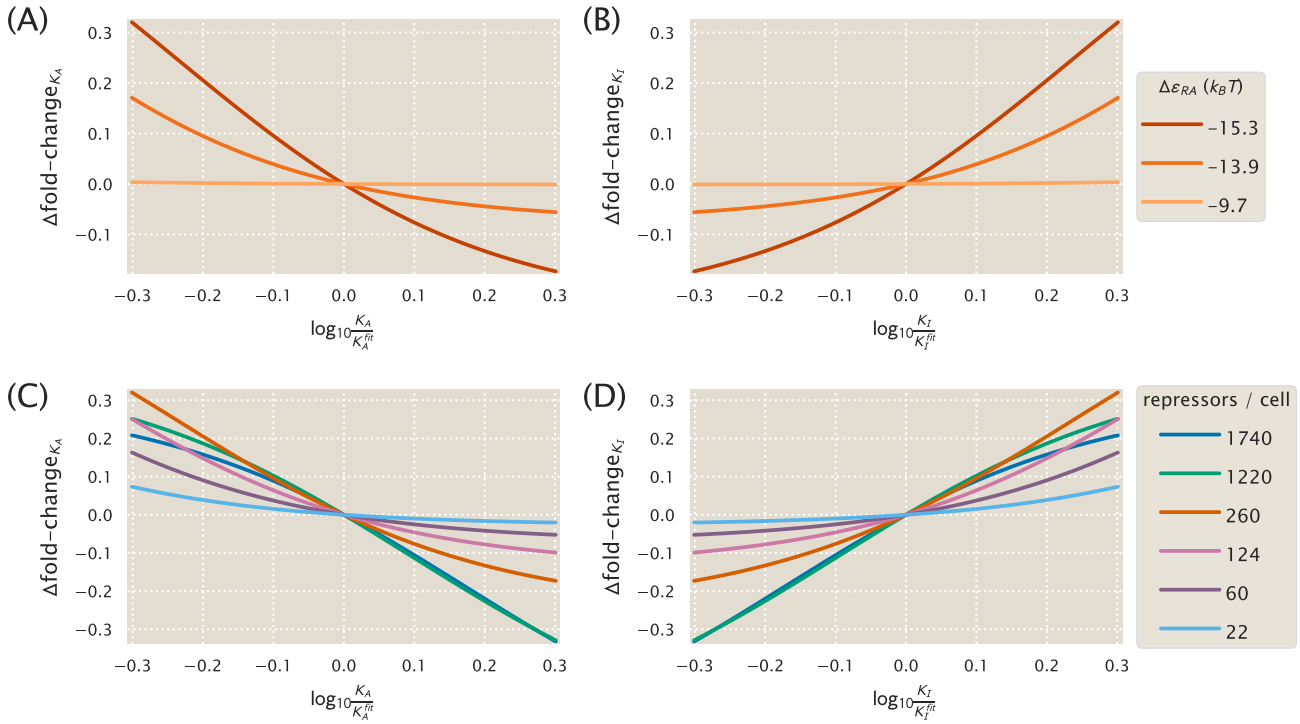

Figure S15. Determining how sensitive the fold-change values are to the fit values of the dissociation constants.(A) The difference Δ fold-change$_{K_A}$ in fold change when the dissociation constant K_A is slightly offset from its best-fit value $K_A = 139^{+29}_{-22} \times 10^{-6}$ M, as given by Eq. (S23). Fold-change is computed in the limit of saturating inducer concentration ($c \to \infty$, see Eq. (7)) where the credible regions in Fig. 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. Hence, 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. 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.

F 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.

F.1 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

$$\text{fold-change} = (\text{leakiness}) + (\text{dynamic range})\frac{\left(\frac{c}{K}\right)^n}{1+\left(\frac{c}{K}\right)^n}, \tag{S27}$$

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. S16 shows how the individual induction profiles can be fit (using the same Bayesian methods as described in Appendix G and the Methods) to this Hill response, yielding a similar response to that shown in Fig. 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. S17. 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. (S27) 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.

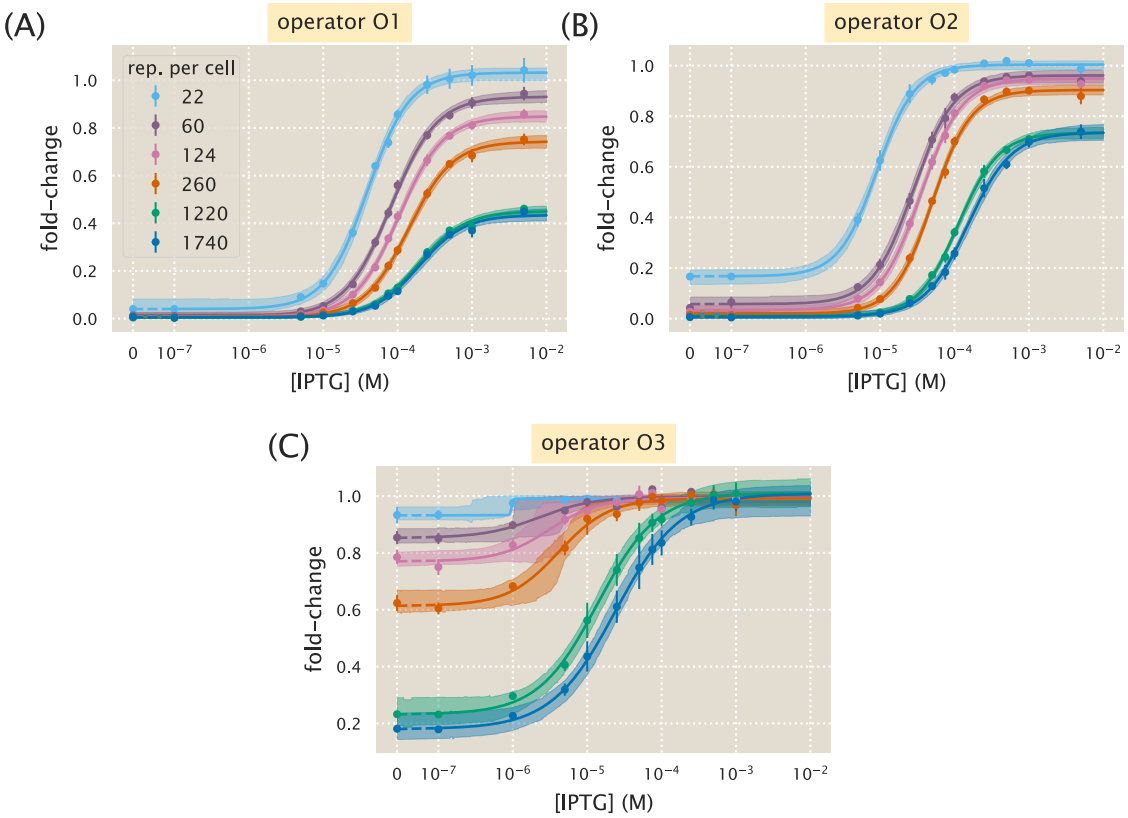

Figure S16. Hill function and MWC analysis of each induction profile. Data for each individual strain was fit to the general Hill function in Eq. (S27). (A) strains with O1 binding site, (B) strains with O2 binding site, and (C) strains with O3 binding site. Shaded regions indicate the bounds of the 95% credible region.

F.2 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(A) but substitute the statistical model in Fig. 2(B) with the phenomenological Hill function Eq. (S27).

Following Eqs. (1)-(3), fold-change is given by

$$\text{fold-change} = \left(1 + p_A(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} \tag{S28}$$

where the Hill function

$$p_A(c) = p_A^{\max} - p_A^{\text{range}}\frac{\left(\frac{c}{K_D}\right)^n}{1+\left(\frac{c}{K_D}\right)^n} \tag{S29}$$

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 $p_A^{\max} - p_A^{\text{range}}$ 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.

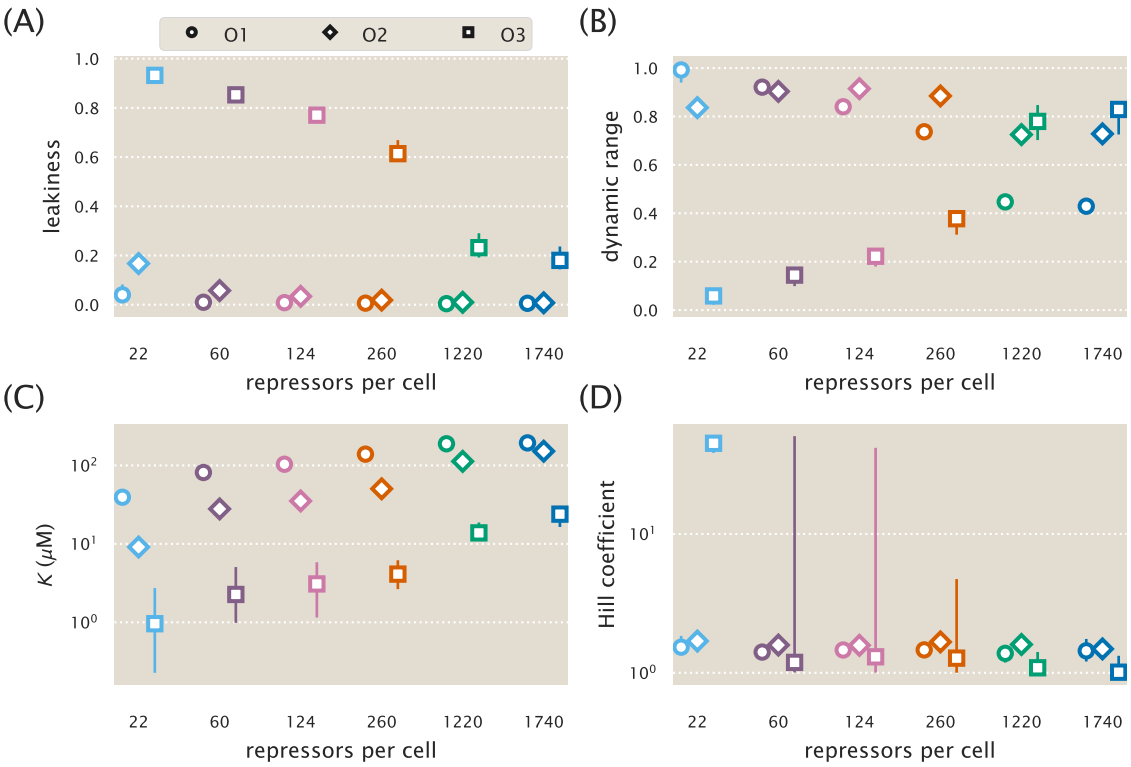

Figure S17. Parameter values for the Hill equation fit to each individual titration. The resulting fit parameters from the Hill function fits of Fig. S16 are summarized. The large parameter intervals for many of the O3 strains are due to the flatter induction profile (as seen by its smaller dynamic range), and the ability for a large range of K and n values to describe the data.

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. (S28) to describe the response of each of the eighteen strains. Fig. S18 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. S18 are nearly identical to those in Fig. 5 (which were made using the MWC model Eq. (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 $K_D = 4^{+2}_{-1} \times 10^{-6}\,\text{M}$ 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. (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 difference $\Delta\varepsilon_{AI}$ between the repressor's active and inactive conformations. For our purposes, the Hill function Eq. (S29) 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.

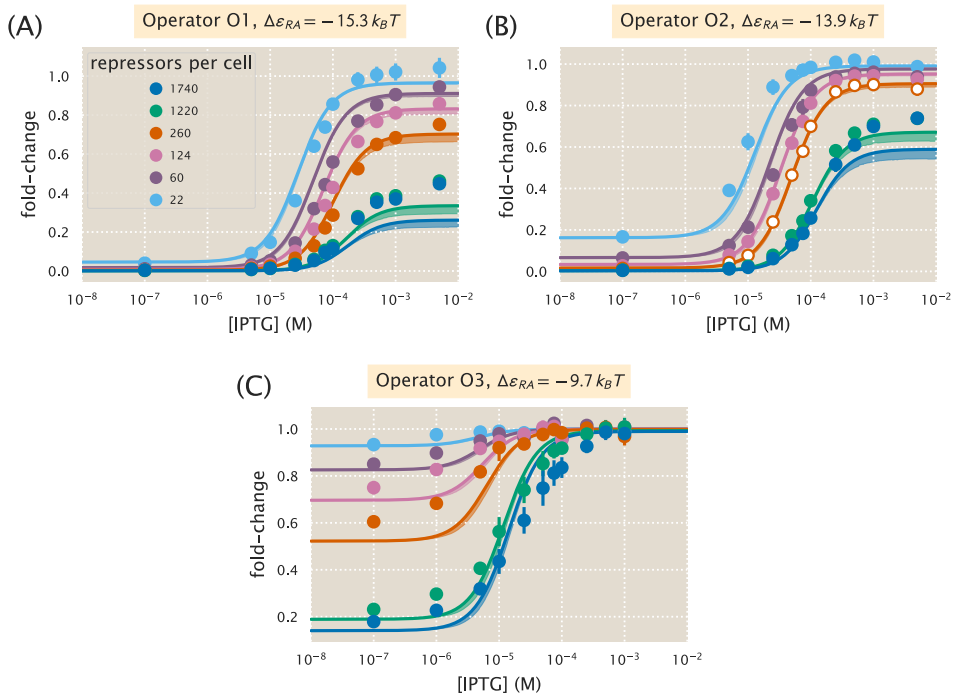

Figure S18. A thermodynamic model coupled with a Hill analysis can characterize induction. Combining a thermodynamic model of simple repression with the Hill function to characterize the repressor-inducer binding successfully characterizes the induction profiles of all eighteen strains. As in the main text, data was only fit for the O2 $R = 260$ strain using Eqs. (S28) and (S29) and the parameters $p_A^{\max} = 0.90^{+0.03}_{-0.01}$, $p_A^{\text{range}} = -0.90^{+0.02}_{-0.03}$, $n = 1.6^{+0.2}_{-0.1}$, and $K_D = 4^{+2}_{-1} \times 10^{-6}$ M. Shaded regions indicate bounds of the 95% credible region.

G 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 (2011a). 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. 6 along with microscopy data for the synthetic operator Oid (see Appendix H). 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 S2.

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 $\tilde{k}_A = -\log \frac{K_A}{1\,\mathrm{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\mathrm{M}}$ of the dissociation constants which grants better numerical stability.

As in Eqs. (15) and (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,

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma \mid D) = \frac{P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma)}{P(D)}, \tag{S30}$$

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. (S30) can be written as

$$P(D \mid \tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\mathrm{fc}_{\mathrm{exp}}^{(i)} - \mathrm{fc}(\tilde{k}_A, \tilde{k}_I, R^{(i)}, \Delta\varepsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2}\right], \tag{S31}$$

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. (S30) represents the prior information of the parameter values. We assume that all parameters are independent of each other, so that

$$P(\tilde{k}_A, \tilde{k}_I, \boldsymbol{R}, \boldsymbol{\Delta\varepsilon_{RA}}, \sigma) = P(\tilde{k}_A) \cdot P(\tilde{k}_I) \cdot \prod_i P(R^{(i)}) \cdot \prod_j P(\Delta\varepsilon_{RA}^{(j)}) \cdot P(\sigma), \tag{S32}$$

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 (2011a). This prior knowledge is included

within our model using an informative prior for these two parameters, which we assume to be Gaussian. Hence each of the $R^{(i)}$ repressor copy numbers to be fit satisfies

$$P(R^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_{R_i}^2}} \exp\left(-\frac{(R^{(i)} - \bar{R}^{(i)})^2}{2\sigma_{R_i}^2}\right), \tag{S33}$$

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 (2011a). 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

$$P(\Delta\varepsilon_{RA}^{(j)}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_j}^2}} \exp\left(-\frac{(\Delta\varepsilon_{RA}^{(j)} - \Delta\bar{\varepsilon}_{RA}^{(j)})^2}{2\sigma_{\varepsilon_j}^2}\right), \tag{S34}$$

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 (2011a) .

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. (S33) 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. S19 shows the result of this global fit. When compared with Fig. 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 S3 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 (2011a). 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 Appendix H.

Fig. S20 shows the same key properties as in Fig. 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.

Table S2. Key model parameters for induction of an allosteric repressor.

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 S3. Global fit of all parameter values using the entire data set in Fig. 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 $\pm$ 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 $\tilde{k}_A = -\log \frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\text{M}}$ of the dissociation constants were fit for numerical stability.

	Reported Values (Garcia and Phillips 2011a)	Global Fit
$\tilde{k}_A$	–	$-5.33^{+0.06}_{-0.05}$
$\tilde{k}_I$	–	$0.31^{+0.05}_{-0.06}$
K_A	–	$205^{+11}_{-12}\,\mu$M
K_I	–	$0.73^{+0.04}_{-0.04}\,\mu$M
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 $\Delta\varepsilon_{RA}$	$-17.0 \pm 0.2\ k_BT$	$-17.7^{+0.2}_{-0.1}\ k_BT$

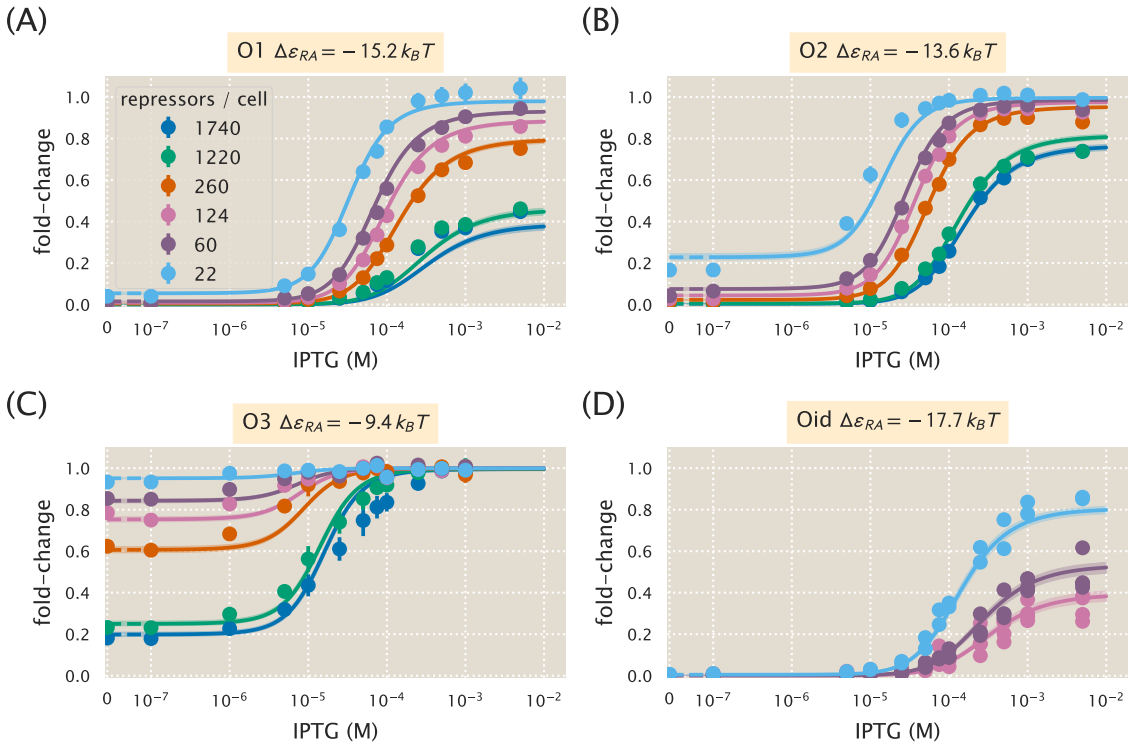

Figure S19. Global fit of dissociation constants, repressor copy numbers and binding energies. Theoretical predictions resulting from simultaneously fitting the dissociation constants K_A and K_I, the six repressor copy numbers R, and the four repressor-DNA binding energies $\Delta\varepsilon_{RA}$ using the entire data set from Fig. 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. 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.

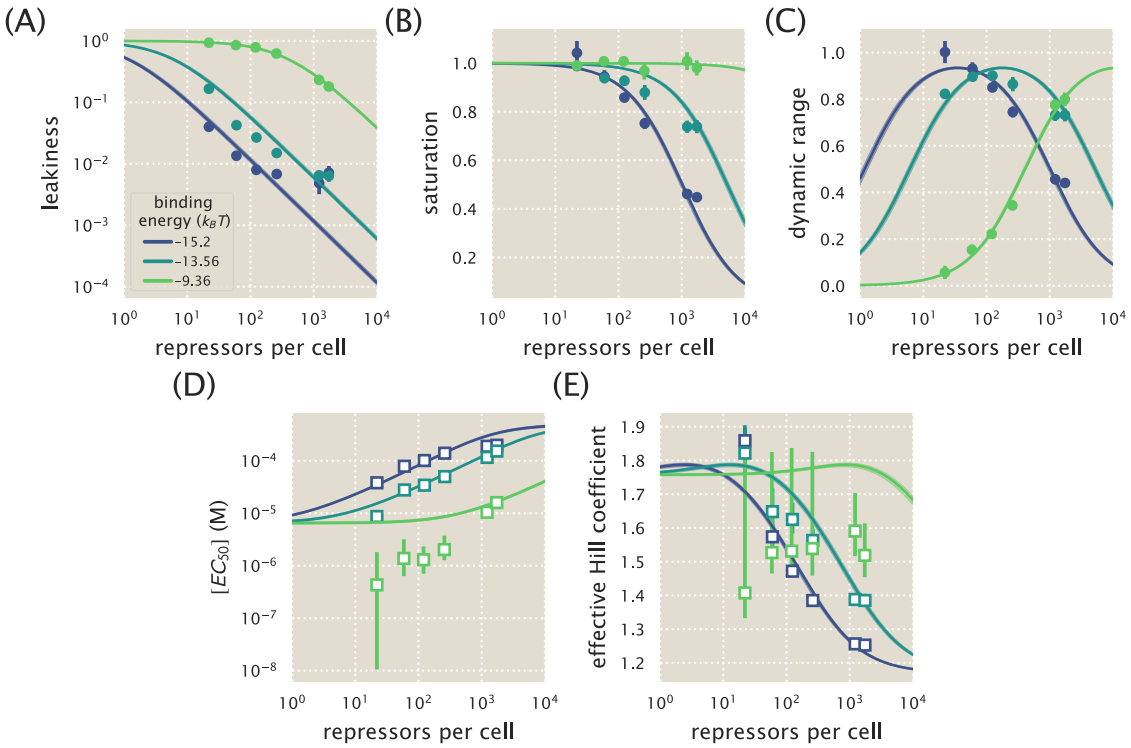

Figure S20. Key properties of induction profiles as predicted with a global fit using all available data. Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 6 in the absence and presence of IPTG. All prediction curves were generated using the parameters listed in S3. 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. 6(A)-(C) separately to Eq. (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.

H 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. 2011b). Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry. Appendix D 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. S21(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 Appendix G) to see whether we might find better agreement with the observed data. Interestingly, we findthat 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 (2011a). In Fig. S21(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.

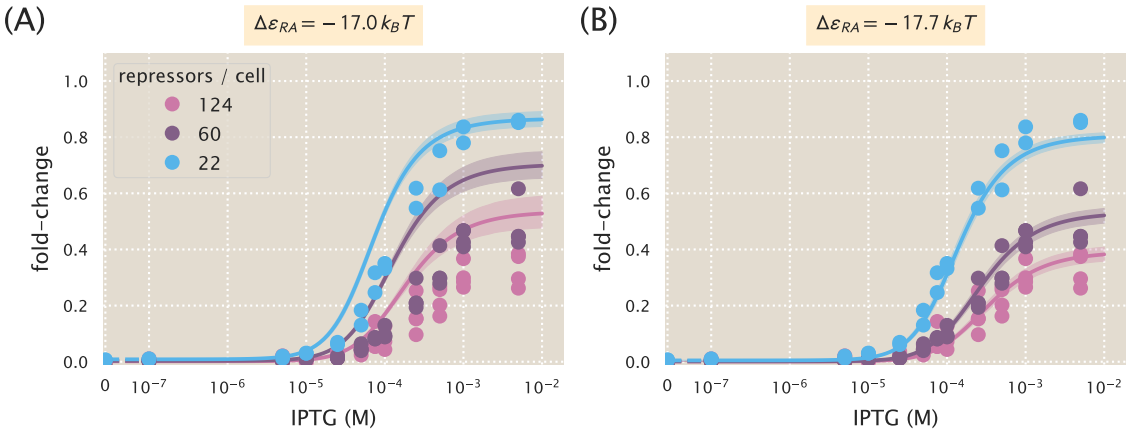

Figure S21. Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers. (A) Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with $R = 260$. Here we use the previously measured binding energy $\Delta\varepsilon_{RA} = -17.0\ k_BT$ (Garcia and Phillips 2011a). (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 Appendix G). 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.

Fig. S22 shows the cumulative data from Garcia and Phillips (2011a) 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.

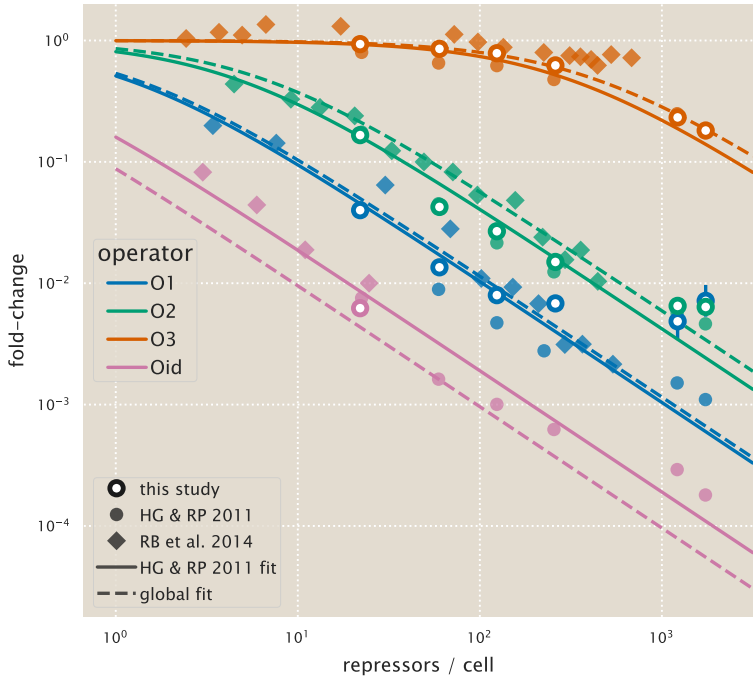

Figure S22. Comparison of fold-change predictions based on binding energies from Garcia and Phillips and those inferred from this work. Fold-change curves for the different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ are plotted as a function of repressor copy number when IPTG concentration $c = 0$. Solid curves use the binding energies determined from Garcia and Phillips (2011a), 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 (2011a) (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. S21 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.

I 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. 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. (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. S23 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. S24 and Fig. S25, 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. 6(D).

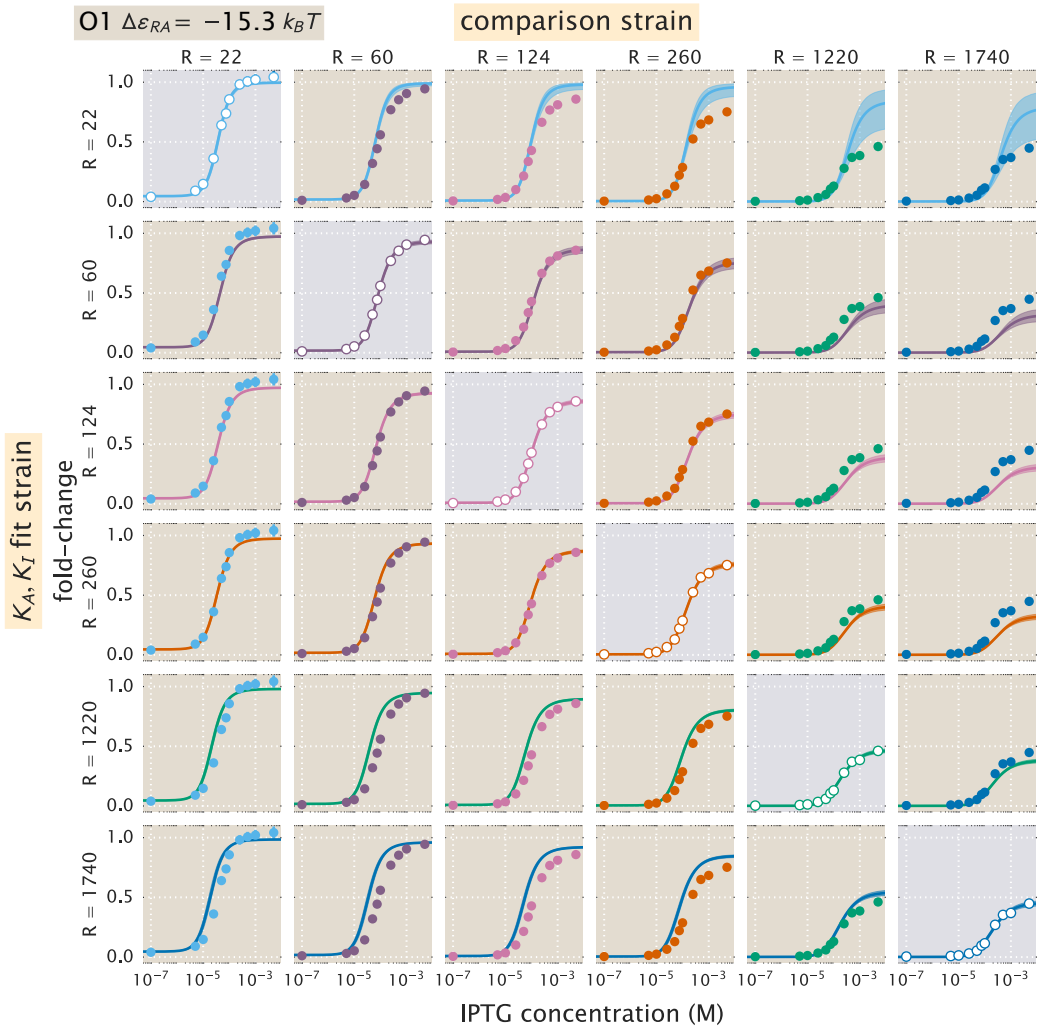

Figure S23. O1 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I. Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O1 operator. The solid points correspond to the mean experimental value. The solid lines correspond to Eq. (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 (2011a). The shaded region on the curve represents the uncertainty from our parameter estimates and reflect the 95% highest probability density region of the parameter predictions.

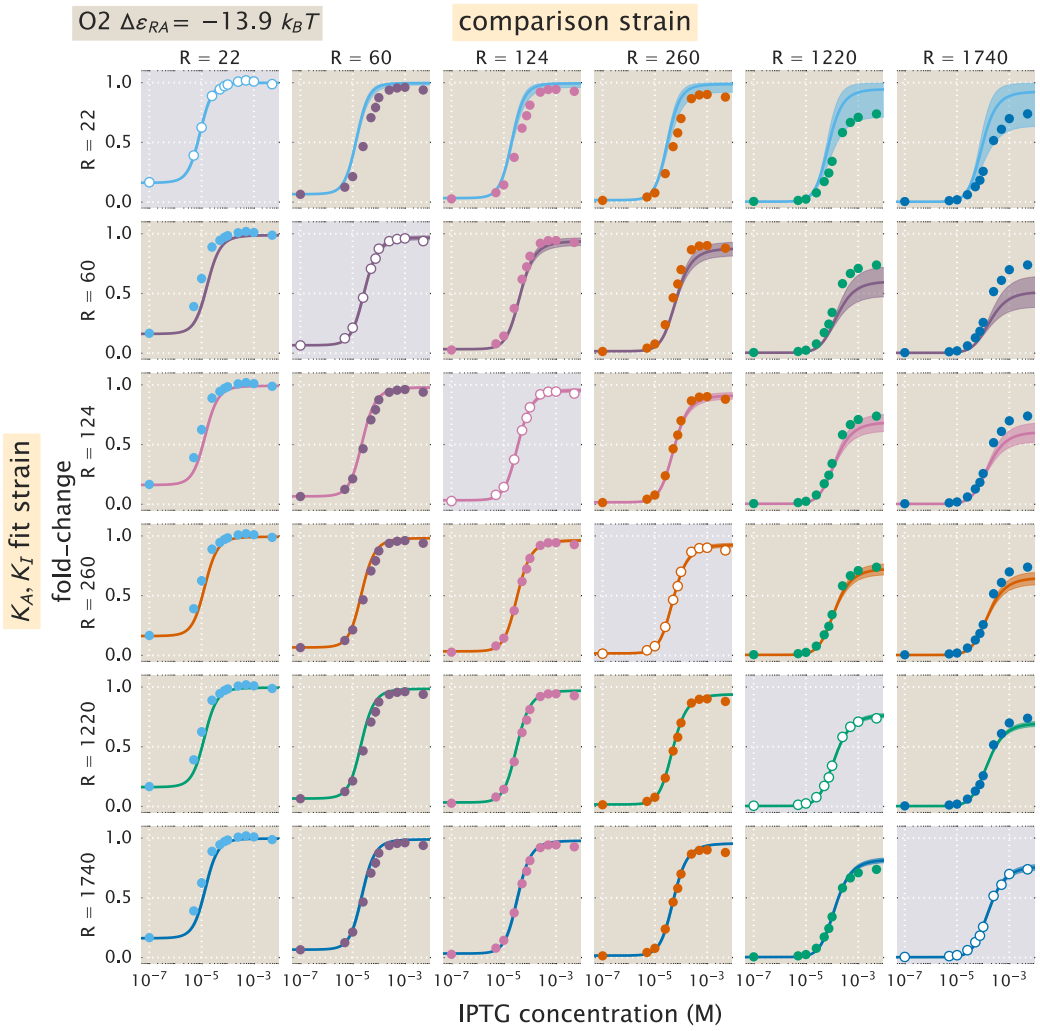

Figure S24. O2 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I. Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O2 operator. The plots and data shown are analogous to Fig. S23, but for the O2 operator.

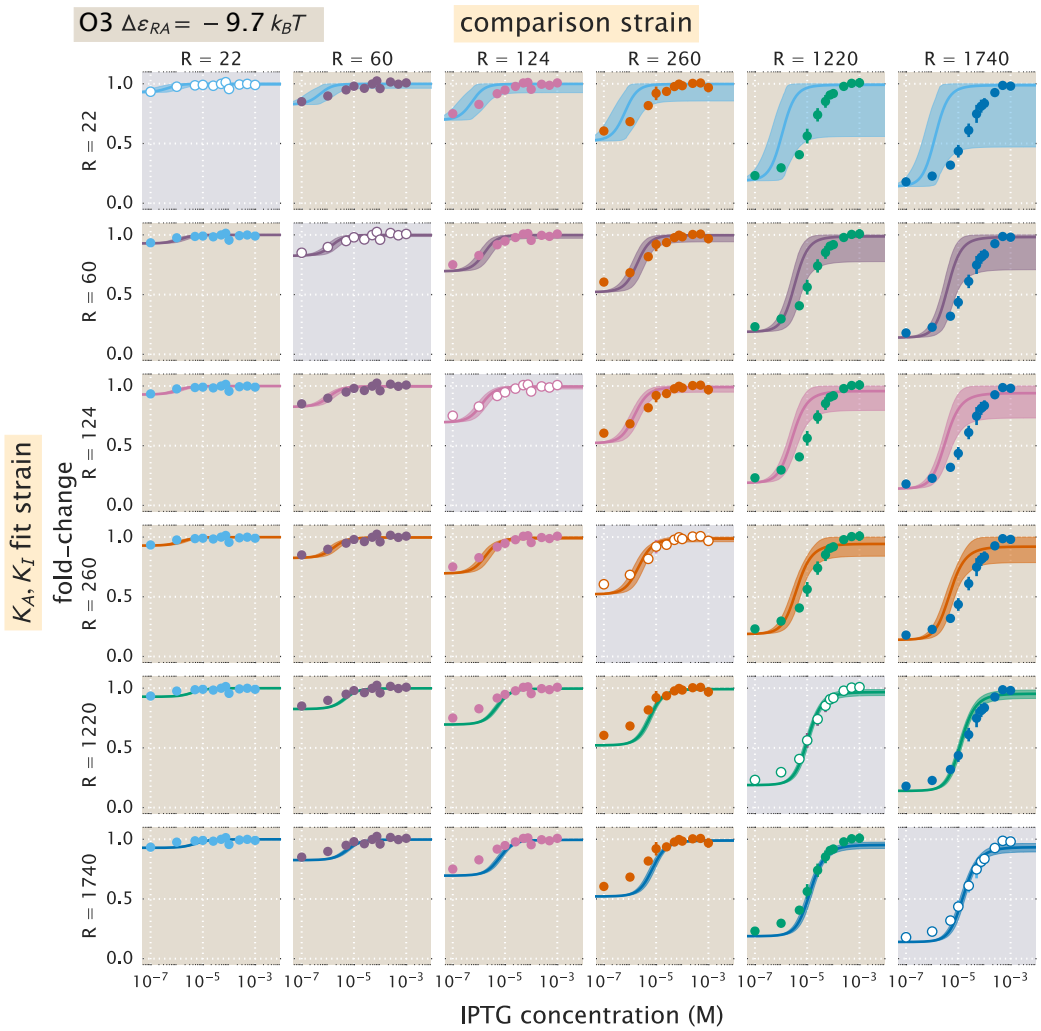

Figure S25. O3 strain fold-change predictions based on strain-specific parameter estimation of K_A and K_I. Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O3 operator. The plots and data shown are analogous to Fig. S23, but for the O3 operator. 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. 6(D)) leads to correspondingly wider credible regions.

J 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. 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. (6)-(8), the dynamic range is given by

$$\text{dynamic range} = \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} - \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}. \tag{S35}$$

The dynamic range, along with saturation and leakiness were plotted with our experimental data in Fig. 7(A)-(C) as a function of repressor copy number. Fig. S26 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. 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, $Re^{-\beta\Delta\varepsilon_{RA}}$. Hence, increasing R on a logarithmic scale (as in Fig. 7(A)-(C)) is equivalent to decreasing $\Delta\varepsilon_{RA}$ on a linear scale (as in Fig. S26).

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. (S35) and setting it equal to zero, we find that this peak occurs at

$$\frac{R^*}{N_{NS}} = e^{-\beta(\Delta\varepsilon_{AI}-\Delta\varepsilon_{RA})}\sqrt{e^{\Delta\varepsilon_{AI}}+1}\sqrt{e^{\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n}. \tag{S36}$$

The magnitude of the peak is given by

$$\text{max dynamic range} = \frac{\left(\sqrt{e^{\Delta\varepsilon_{AI}}+1}-\sqrt{e^{\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n}\right)^2}{\left(\frac{K_A}{K_I}\right)^n - 1}, \tag{S37}$$

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.

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. (9). For the simple repression system, the $[EC_{50}]$ is given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I}-1}{\frac{K_A}{K_I} - \left(\frac{\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+\left(\frac{K_A}{K_I}\right)^n\left(2e^{-\beta\Delta\varepsilon_{AI}}+\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)\right)}{2\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+e^{-\beta\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\varepsilon_{AI}}}\right)^{\frac{1}{n}}} - 1. \tag{S38}$$

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. (10)). In Fig. 7(D)-(E) we show how these two properties vary with repressor copy number, and in Fig. S27 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.

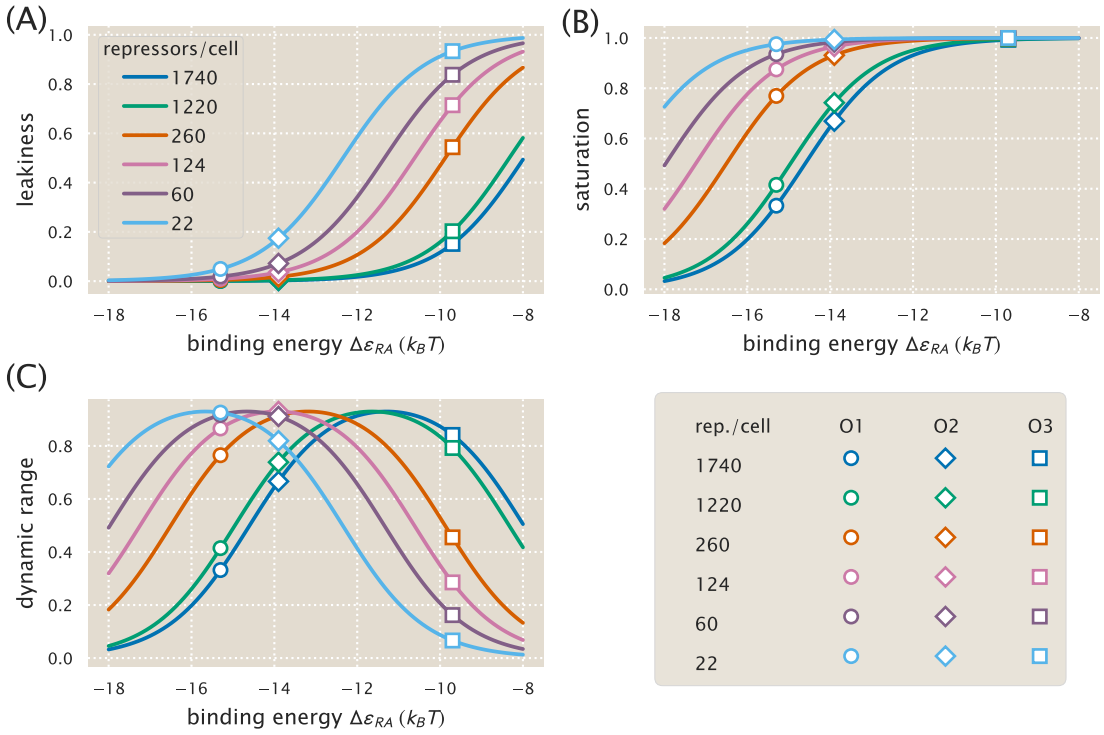

Figure S26. Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number. Increasing repressor copy number or decreasing the repressor-operator binding energy suppresses gene expression and decreases both the (A) leakiness and (B) saturation. (C) The 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. (S37)). 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.

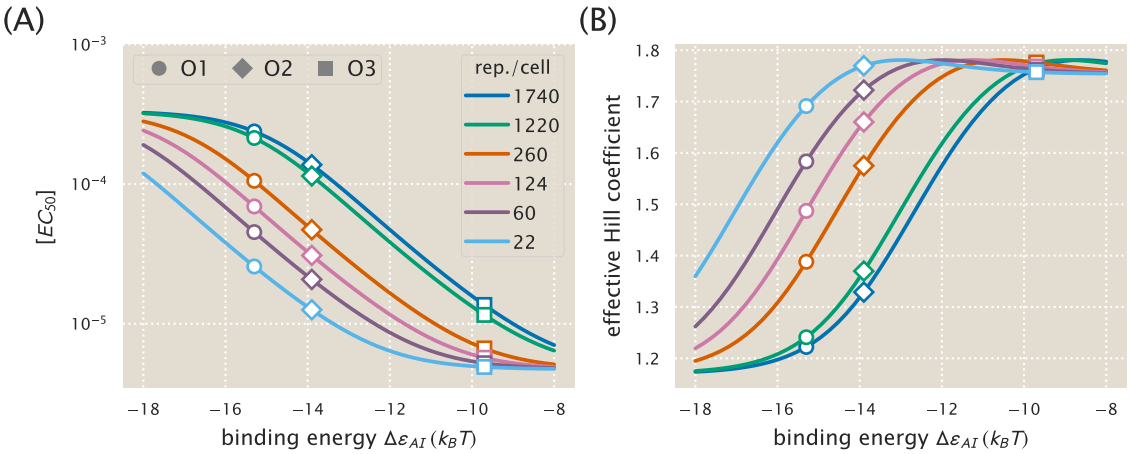

Figure S27. $[EC_{50}]$ and effective Hill coefficient depend strongly on repressor copy number and operator binding energy. (A) $[EC_{50}]$ values range from very small and tightly clustered at weak operator binding energies (e.g. O3) to relatively large and spread out for stronger operator binding 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.

K 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. (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. (2005b) and Marzen et al. (2013).

K.1 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 and yield the same fold-change equation as Eq. (5),

$$\text{fold-change} \approx \left(1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{S39}$$

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. 5). Much as was performed in the main text, we can make some predictions about the how the response of a corepressor. In Fig. S28(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.

K.2 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

$$\text{fold-change} = \frac{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_{AA}}e^{-\beta\varepsilon_{AP}}}{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_{AA}}}, \tag{S40}$$

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. S28(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. (S40) 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.

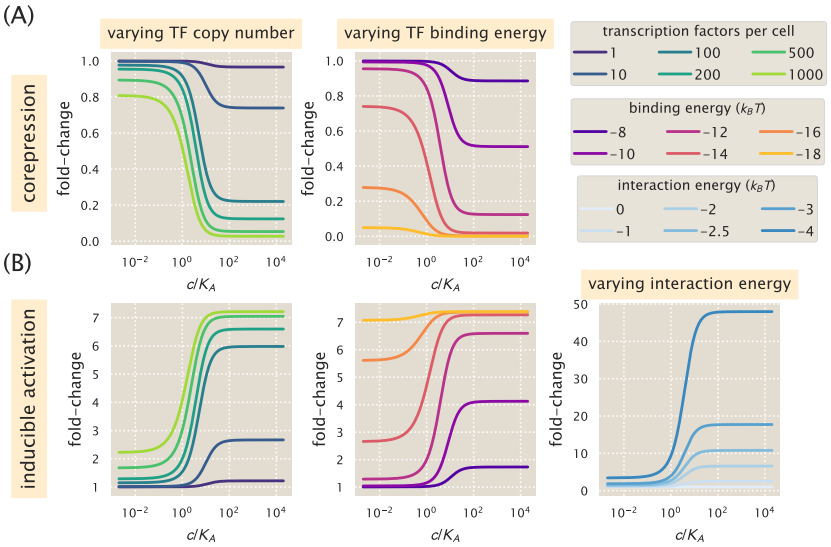

Figure S28. Representative fold-change predictions for allosteric corepression and activation. (A) Contrary to the case of induction described in the main text, addition of a corepressor decreases fold-change in gene expression. The left and right panels demonstrate how varying the values of the repressor 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.

L *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 (2011a) using the primers listed in Table S4. 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 (2011a). 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 S5. 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.

Table S4. 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.

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*)	gtttgcgcgcagtcagcgatatccattttcgcgaatccgg agtgtaagaaACTAGCAACACCAGAACAGCC	Reverse primer for 25x+11-yfp with homology to *galK* locus.
HG6.3 (*galK*)	ttcatattgttcagcgacagcttgctgtacggcaggcacc agctcttccgGGCTAATGCACCCAGTAAGG	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*)	acctctgcggaggggaagcgtgaacctctcacaagacggc atcaaattacACTAGCAACACCAGAACAGCC	Reverse primer for 3*1x-lacI with homology to *ybcN* locus.
HG11.3 (*ybcN*)	ctgtagatgtgtccgttcatgacacgaataagcggtgtag ccattacgccGGCTAATGCACCCAGTAAGG	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 S5. *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 (2011a).

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)

Supplemental References

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.

Chroma Technology Corporation (2016). Chroma Spectra Viewer.

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.

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

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.

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

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.

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

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