Sequence-Function Relationships in *E. coli* Transcriptional Regulation

Thesis by

Daniel Jones

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

California Institute of Technology
Pasadena, California

2014
(Defended May 30, 2014)

Acknowledgements

All of the work described in this thesis involved colloborations with others, without whom much of this work would not have been possible. Rob Brewster has been an invaluable experimental partner throughout my time at Caltech, and one could not hope for a better companion on the journey from complete bench neophyte to occasional attainment of at least a minimal level of competence. Justin Kinney provided essential guidance on the sort-seq work which is discussed directly in chapter 3 but which also serves as backdrop throughout all the projects described here. Heun Jin Lee's assistance with microscopy was essential to the work done involving mRNA FISH, and his knowledge of the physiology and regulation of the MscL protein has played a key role in efforts to unravel MscL's transcriptional regulation. Of course, I want to thank my advisor, Rob Phillips, for asking skeptical questions about experimental measurements, for willingness to consider new ideas and approaches, and above all for creating a supportive and congenial lab environment. It has been an immense pleasure working with Rob and all the members of the Phillips lab.

On the non-scientific front, I want to extend a special thanks to Delores Bing, director of the student chamber music program at Caltech, as well as to all my chamber music partners from over the years. This program is one of the best things about Caltech and I'm grateful to have had the opportunity to work with some outstanding musicians.

Abstract

Understanding how transcriptional regulatory sequence maps to regulatory function remains a difficult problem in regulatory biology. Given a particular DNA sequence for a bacterial promoter region, we would like to be able to say which transcription factors bind there, how strongly they bind, and whether they interact with each other and/or RNA polymerase, with the ultimate objective of integrating knowledge of these parameters into a prediction of gene expression levels. The theoretical framework of statistical thermodynamics provides a useful framework for doing so, enabling us to predict how gene expression levels depend on transcription factor binding energies and concentrations. We used thermodynamic models, coupled with models of the sequence-dependent binding energies of transcription factors and RNAP, to construct a genotype to phenotype map for the level of repression exhibited by the *lac* promoter, and tested it experimentally using a set of promoter variants from *E. coli* strains isolated from different natural environments. For this work, we sought to "reverse engineer" naturally occurring promoter sequences to understand how variations in promoter sequence affects gene expression. The natural inverse of this approach is to "forward engineer" promoter sequences to obtain targeted levels of gene expression. We used a high precision model of RNAP-DNA sequence dependent binding energy, coupled with a thermodynamic model relating binding energy to gene expression, to predictively design and verify a suite of synthetic *E. coli* promoters whose expression varied over nearly three orders of magnitude. However, although thermodynamic models enable predictions of mean levels of gene expression, it has become evident that cell-to-cell variability or "noise" in gene expression can also play a biologically important role. In order to address this aspect of gene regulation, we developed models based on the chemical master equation framework and used them to explore the noise properties of a number of common *E. coli* regulatory motifs; these properties included the dependence of the noise on parameters such as transcription factor binding strength and copy number. We then performed experiments in which these parameters were systematically varied and measured the level of variability using mRNA FISH. The results showed a clear dependence of the noise on these parameters, in accord with model predictions.

Finally, one shortcoming of the preceding modeling frameworks is that their applicability is largely limited to systems that are already well-characterized, such as the *lac* promoter. Motivated by this fact, we used a high throughput promoter mutagenesis assay called Sort-Seq to explore the completely uncharacterized transcriptional regulatory DNA of the *E. coli* mechanosensitive channel of large conductance (MscL). We identified several candidate transcription factor binding sites, and work is continuing to identify the associated proteins.

Contents

Chapter 1

Introduction

1.1 The central dogma of molecular biology

The elucidation of the genetic code is one of the signal accomplishments of the field of molecular biology. It was the culmination of a decades-long search to unravel the mechanism by which genetic information is passed down from generation to generation in living organisms. Key insights along the way include Oswald Avery's discovery that DNA (and not protein) is the molecule by which genetic information is propagated [1]; Chargaff's observation that the fraction of As equals the fraction of T's in a DNA molecule, and likewise for Cs and Gs [2, 3]; and Watson and Crick's discovery of the structure of DNA, which "immediately suggests a possible copying mechanism for the genetic material," as Crick put it at the time [4]. Still, the precise mechanism by which a particular DNA sequence mapped to a particular protein remained unknown. In 1961, Crick, Brenner, and coworkers arrived at the now familiar result that a protein coding sequence consists of a series of trinucleotide codons [5], by showing that insertions of three base pairs into the phage T4 *rIIB* gene yielded a functioning protein (whereas insertions of one, two, or four bp yielded a non-functional protein). Researchers in various laboratories subsequently determined the mapping between codon sequence and amino acid identity; the ultimate result of these efforts was the codon table as seen in Figure 1.1. The actual act of translating from DNA sequence to protein sequence occurs in the ribosomes, where transfer RNAs recognize each codon of the messenger RNA in turn and add the appropriate amino acid to the growing polypeptide chain. This entire process by which genetic information flows from DNA to messenger RNA to protein was termed the "central dogma" of molecular biology by Crick in 1958.

The net result of these efforts was mastery of the genetic code by which DNA sequence is mapped to the chain of polypeptides that constitute a protein. However, encoding protein coding sequences is far from the only function of DNA. DNA sequence also encodes information about how much and at what times genes are expressed. Yet the nature of the code by which regulatory DNA sequence maps to regulatory function remains largely unknown. Given the DNA sequence of any particular protein

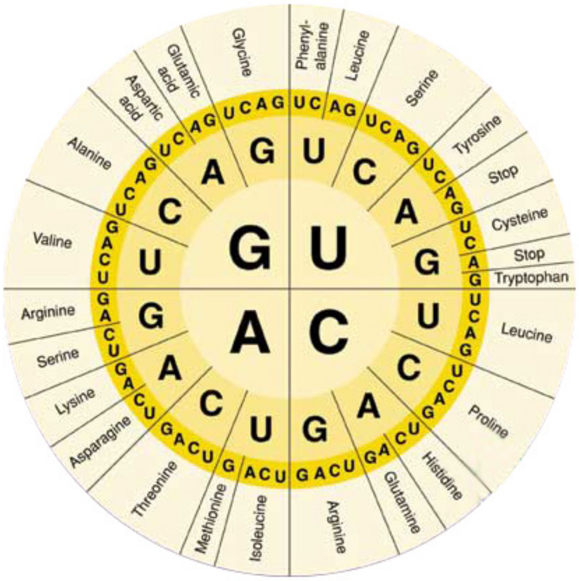


Figure 1.1: **Codon table.** The mapping between codon sequence and amino acid identity can be read off the table from the inner to outer rings.

coding sequence, it is trivial to predict the exact amino acid sequence of the resulting protein. But predicting the regulatory function of a given sequence of regulatory DNA is more or less completely infeasible except in a limited number of special cases. The principal thrust of this thesis, then, will be to arrive at a more detailed understanding of how DNA sequence maps to regulatory function. “Regulatory function” is an admittedly malleable term, and could be taken to mean the level of expression of a particular gene in terms of absolute numbers of proteins per cell; the degree to which expression of a gene is turned off under certain environmental conditions (“repression”), or the level of cell-to-cell variability or “noise” in expression. Each of these will be considered in detail in this work.

1.2 Gene regulation

Gene regulation is essential for the fitness of living organisms. While gene coding sequences encode the “raw materials” (proteins) out of which an organism is made, it is equally important that these proteins are expressed in the right amount at the right time. One aspect of expressing the right proteins at the right time is the fact that the cell simply needs more of some proteins than others. For instance, ribosomes, multi-protein complexes that translate messenger RNAs into protein, lie at the heart of all gene expression, to the extent that their production constitutes the rate-limiting step in cell division of exponentially growing bacteria. Consequently, genes coding for ribosomal RNA (in the *rrnA-E,G,H* operons) are some of the most highly transcribed in *E. coli*, to the extent that their transcription can account for the majority of RNAP activity [6, 7]. In contrast, a mere

10 copies per cell of the *lac* repressor are sufficient to repress expression of the *lac* operon by 1000 fold (see Chapter 2 of this work). To state the obvious, a cell with 10 ribosomes and 10,000 *lac* repressor molecules would be wholly dysfunctional. Ensuring that global stoichiometries of cellular components are appropriately balanced is thus an important aspect of gene regulation [6, 8].

Another important function of gene regulation is enabling cells to respond appropriately to environmental conditions. For instance, the canonical *lac* promoter system "turns on" production of the appropriate enzyme (LacZ) and membrane transporter (LacY) for metabolization of the dissacharide lactose when lactose is present in the environment and glucose is not. Production of these proteins is costly to the cell and thus cells that avoid expression when these conditions do not obtain have a fitness advantage over cells that do [9, 10]. In *B. subtilis*, an array of transcription factors and sigma factors is responsible for differentiation in times of nutritional stress into a sporulated state characterized by a tough external coating and virtually no energy consumption, allowing the cell to survive until environmental conditions are more favorable [11].

Gene regulation occurs at all the steps along the central dogma. In prokaryotes, transcription is regulated by DNA-binding proteins called transcription factors (TFs) that bind a gene's promoter region and activate or repress transcription of that gene. Unlike in the eukaryotic setting, where enhancers can be up to tens of kilobases away from the promoter, most prokaryotic transcription factors bind within approximately 100 bp of the gene they regulate (M. Rydenfelt, manuscript in preparation; data from [12]). In eukaryotes, it is well established that transcription is also regulated by the chromatin state, as DNA condensation mediated by nucleosomes can render regions of DNA inaccessible to transcription factors and RNA polymerase [13–16]. Although this mode of regulation is less thoroughly explored in prokaryotes, there is evidence that nucleoid associated proteins like H-NS, HU, Fis, and IHF, which are structurally similar to eukaryotic histone proteins, can also affect gene expression by structural modification of the chromosome [17–20]. Similarly, the supercoiling state of prokaryotic DNA can also play a role in determining gene expression [21, 22].

After transcription, translation is regulated by two principal mechanisms: small RNAs and the sequence of the ribosomal binding site. The ribosomal binding site is located in the 5' untranslated region (5' UTR) of an mRNA transcript, and its strength has important implications for how much a particular mRNA transcript is transcribed. Factors involved in ribosome binding strength include the interaction between the ribosome binding site and the 16s rRNA, the spacing between 16s rRNA binding site and the start codon, start codon sequence, and the free energy cost of unfolding any mRNA secondary structures in the RBS region [23, 24]. Small RNAs include both *cis* and *trans* encoded sRNAs. The former are transcribed from the antisense strand to a particular protein coding sequence and when paired to their complementary mRNA reduce gene expression by blocking translation. The latter have less extensive complementarity with their respective mRNA targets, and interactions between sRNA and mRNA are generally mediated by the RNA chaperone Hfq.

Such sRNAs can repress translation by blocking the ribosomal binding site or enhancing mRNA degradation; they can also increase translation by disrupting mRNA secondary structures that sequester the ribosomal binding site [25–27]. Finally, the activity of already-produced proteins is often modulated by post-translational modifications such as phosphorylation and dephosphorylation.

Given that gene regulation occurs at all steps along the central dogma, what determines the step at which a particular gene is regulated? And what are the relative prevalences of the different forms of gene regulation? Unfortunately, definitive answers are not available for either of these questions, though it is certainly possible to speculate. Possible considerations include speed of response and metabolic efficiency. Post-translational modifications require only a single phosphorylation reaction to take place and thus can operate on a relatively fast timescale. However, this type of regulation requires that the relevant proteins have already been produced. Transcriptional regulation occurs at the very beginning of the path from gene to protein, and thus is more efficient in the sense that cutting off gene expression at the source means that no resources need to be expended on unnecessary gene expression. However, the timescale at which transcriptional regulation can respond to changing environmental conditions is limited by the binding/unbinding kinetics of the relevant transcription factors (often on the order of minutes), the time needed for RNA polymerase to produce an mRNA transcript (tens of seconds), translation (tens of seconds), protein folding (variable; often < 1 second [28]), and, in some cases, protein maturation. Thus it seems reasonable to postulate a tradeoff between speed of response and efficiency between these modes of gene regulation. Another possible consideration is the nature of the response function to environmental stimuli: computational studies have found that small RNA regulation can produce qualitatively different response functions compared with TF-mediated transcriptional regulation [27, 29, 30]. As for the relative prevalence of different forms of regulation, despite the status of *E. coli* as a ubiquitous model organism for nearly a century, there is insufficient data to make a definitive pronouncement. For instance, of the roughly 4000 genes in *E. coli*, roughly half lack any transcriptional regulatory annotation whatsoever. It seems unlikely that all of these genes have no transcriptional regulation; far more likely is that this state of affairs reflects simple ignorance. In a subsequent chapter, we use a high throughput promoter mutagenesis experiment to explore the transcription of one such un-annotated gene, namely *mscL* (the mechanosensitive channel of large conductance). We identify three putative transcription factor binding sites. Of course, this $n = 1$ observation does not prove anything, but it lends support to the idea that vast swathes of *E. coli* gene regulation remain uncharacterized, and hence that statements about the relative importance of transcriptional vs. translational vs. post-translational regulation are necessarily speculative in the absence of much-needed data.

In any case, it clear that transcriptional regulation is of profound importance in many biological contexts. Perhaps the most dramatic example of this lies in the developmental biology of eukaryotes, where highly conserved *hox* genes dictate body plan development across a vast array of species: while

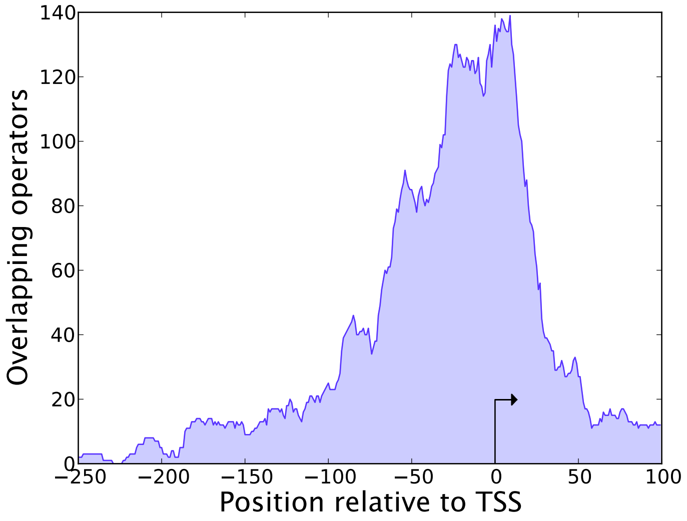


Figure 1.2: **Repressor binding site positions.** This histogram shows the number of repressor binding sites overlapping each nucleotide position, where nucleotide positions are reported with respect to the transcription start site. The plot was generated using data from the RegulonDB database [12]. The majority of repressor binding sites are within 50 bp of the transcription start site, although some are found as far as 200 bp upstream. Adapted from reference [34], courtesy of M. Rydenfelt.

the genes themselves remain largely the same, differences in their regulation encoded by non-protein-coding regulatory DNA yield the "endless forms most beautiful" remarked upon by Darwin. More closely related to the subject of this thesis, transcriptional regulation is integral to phenomena such as bacterial biofilm formation which depends on transcriptional activation of expression of polysaccharides to form a relatively impregnable extracellular matrix [31]. Various other behaviors related to quorum sensing and collective behavior are transcriptionally regulated as well [32, 33].

TF-mediated transcriptional regulation can be divided into two broad categories: repression and activation. Depending on the context, it is possible for the same transcription factor to act as both an activator and a repressor, sometimes even in regulation of the same gene. The principal mechanism of repression is simple steric hindrance. Typically this means that a transcription factor binds in the vicinity of the RNAP polymerase binding site and in doing so prevents the RNAP molecule from binding and/or successfully initiating transcription. The *lac* promoter O1 binding site, located immediately downstream of the transcription start site, is a good example of this type of regulation [35, 36]. A survey of binding site positions in the transcriptional regulatory database RegulonDB reveals that the majority of repressor binding sites are close to the RNAP binding site (see Figure 1.2), allowing the repressor to directly interact with the RNA polymerase [12, 34]. However, there are an appreciable number of repressor binding sites outside of the immediate vicinity of the transcription start site. An alternative mechanism of repression is what might be termed "second-order" repression, in which a repressor indirectly downregulates transcription by preventing

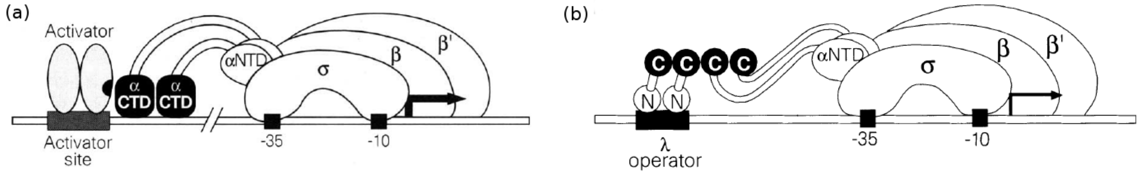


Figure 1.3: **(a)** Type I transcriptional activation. The activator protein (for instance, CRP) interacts with the α C-terminal domain of the RNA polymerase holoenzyme; this favorable energetic interaction increases the probability that RNAP binds the promoter. **(b)** Activation by arbitrary protein-protein contact. The wild-type α-CTD has been replaced by the CTD of the λ cI protein, which binds cooperatively to cI protein bound at an upstream binding site. The presence of cI at the upstream binding site increases transcription by approximately six fold. Adapted from [45].

the binding of an activator. Finally, it is worth noting that repression via steric exclusion can still be mediated by repressor binding sites away from the immediate vicinity of the transcription start site via the formation of DNA loops, as in the cases of the *lac* and *araC* promoters [37–39].

The mechanisms of transcriptional activation are somewhat less straightforward than repression. Activation can be brought about either by stabilizing the formation of the initial closed complex (and thus reducing the dissociation constant of the RNAP holoenzyme-promoter complex), by increasing the isomerization rate from the closed complex to the open complex, or by some combination of both [40–42]. Activation is often divided into two broad classes, namely, Class I and Class II [43, 44]. Class I transcriptional activation involves interactions between the RNAP alpha C terminal domain (α-CTD) and an activator TF bound upstream (10s of bp) of the RNAP binding site. Class II transcriptional activation involves activator TFs bound directly adjacent to or overlapping the promoter -35 region. In class II activation, TFs can (in addition to interactions with the α-CTD) interact with the RNAP at the alpha N terminal domain (α-NTD), or with region 4 of the sigma factor. It has been shown that in the context of Class II activation by CRP, a ubiquitous global TF, interactions between the α-CTD and CRP affect the equilibrium formation of the closed complex, while interactions between the α-NTD and CRP affect the isomerization rate to the open complex [41].

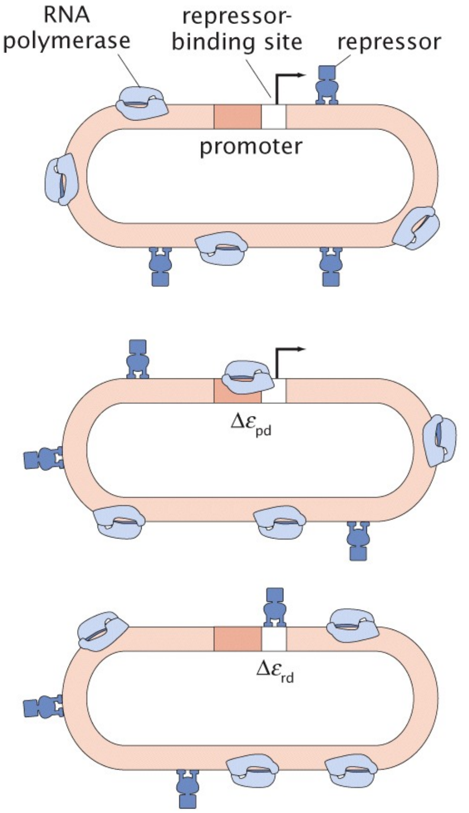
In addition to efforts to understand activation in the context of wild-type promoters, a number of elegant experiments involving synthetic transcriptional activators have been performed. A striking series of experiments from the lab of Ann Hochschild revealed that activation can be mediated by essentially arbitrary protein-protein contacts between RNAP and an activator TF [45, 46]. In Figure 1.3, one such experiment is shown, in which the wild-type RNAP α subunit was replaced with a chimeric protein containing a wild-type N-terminal domain and linker, but with a C-terminal domain replaced by the C-terminal domain of the lambda cI protein. The cI CTD mediates cI dimer-dimer interaction, and so the cI CTDs of the chimeric RNAP α subunit can reasonably be expected to interact with the CTDs of wild-type cI protein bound upstream of the promoter. Dove et al showed that this interaction did indeed occur and moreover was sufficient to cause transcriptional activation [45]. In addition to activation via such "molecular velcro" type interactions, activation can also be effected by "derepression," where a protein serves as an effective activator by blocking or inhibiting a repressor molecule. It has been shown that LacI itself, the canonical example of a repressor, can serve as an activator, in the context of the *E. coli bgl* promoter, by disruption of repression by H-NS [47].

1.3 Thermodynamic model of transcriptional regulation

The preceding discussion touched on many of the most important elements of transcriptional regulation. As physicists and quantitative scientists, we would like to go beyond descriptive cartoons and qualitative descriptions of the effects of various molecular players to construct a more quantitative picture. Adding urgency and relevance to this desire is the fact that many assays to measure gene expression yield quantitative results in terms of the number of mRNA and protein molecules produced. To fully engage with and and learn as much as possible from these measurements, we need to be able to make falsifiable quantitative predictions.

One class of models that has been successful in making quantitative predictions about gene regulation is that of thermodynamic or statistical mechanical models. This may seem surprising at first glance since biological systems are perhaps the single most salient example of systems that reside out of equilibrium; or, as the economist John Maynard Keynes put it, "In the long run we are all dead". Yet although in the long run biological systems are out of equilibrium, the separation of timescales involved in different cellular processes means that a quasi-equilibrium description of certain processes is not inappropriate. As we will see, many studies have used thermodynamic models successfully to quantitatively model gene expression.

Thermodynamic models of gene regulation take as their foundational assumption the idea that the level of gene expression is proportional to the equilibrium probability that RNA polymerase is bound to the promoter. Later, we will explore the conditions under which a quasiequilibrium view of transcription is accurate, using a simplified model of the kinetics of transcription initiation, but in brief, we expect this assumption to be valid in the limit that RNAP binding and unbinding is fast compared with the rate of transcription initiation. The procedure for constructing a thermodynamic model is straightforward in principle: one enumerates all the possible configurations of the system, and for each configuration computes the multiplicity (*i.e.,* the number of ways of realizing a particular configuration) and the energy of that configuration. The Boltzmann weight for a particular state is then simply the multiplicity times e raised to the energy of that configuration divided by k_BT, where k_B is Boltzmann's constant. According to the Boltzmann distribution, the probability of any particular configuration is then given by the Boltzmann weight of that configuration divided



State	Multiplicity	Energy
	$\frac{N_{NS}!}{P!R!(N_{NS}-P-R)!}$	$P\epsilon_{pd}^{NS}+R\epsilon_{rd}^{NS}$
$\Delta\varepsilon_{pd}$	$\frac{N_{NS}!}{(P-1)!R!(N_{NS}-P-R+1)!}$	$(P-1)\epsilon_{pd}^{NS}+\epsilon_{pd}^{S}+R\epsilon_{rd}^{NS}$
$\Delta\varepsilon_{rd}$	$\frac{N_{NS}!}{P!(R-1)!(N_{NS}-P-R+1)!}$	$P\epsilon_{pd}^{NS}+(R-1)\epsilon_{rd}^{NS}+\epsilon_{rd}^{S}$

Figure 1.4: **States, multiplicities, and energies for simple repression.** For the "simple repression" architecture, the promoter can take three possible states (shown top to bottom in the figure): neither repressor nor polymerase bound, polymerase bound, or repressor bound. By assumption, RNAP cannot bind the promoter when the repressor is bound. The multiplicities reflect the number of ways of distributing P polymerases and R repressors among N_{NS} non-specific binding sites, where N_{NS} is typically taken as the length of the genome. The nonspecific polymerase-DNA binding energy is given by ϵ_{pd}^{NS}, while the specific polymerase-DNA binding energy of the promoter of interest is given by ϵ_{pd}^{S}. Likewise, the nonspecific repressor-DNA binding energy is ϵ_{rd}^{NS}, and the specific repressor-DNA binding energy is ϵ_{d}^{S}. The Boltzmann weight of each state is given by $\Omega \times e^{-E/k_BT}$, where Ω is the multiplicity of the state, E is the energy of the state and k_B is Boltzmann's constant. Figure adapted from reference [48].

by the sum of the weights of all configurations (the partition function). The probabilities of all configurations in which RNAP is bound at the promoter are summed, and the overall level of gene expression is assumed to be proportional to this quantity. For instance, for the "simple repression" scenario depicted in Figure 1.4, the probability that RNAP is bound is given by [35]:

$$p_{\text{bound}} = \frac{\frac{P}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{pd}}{k_B T}\right)}{1 + \frac{P}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{pd}}{k_B T}\right) + \frac{R}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{rd}}{k_B T}\right)}, \tag{1.1}$$

where P is the number of polymerases, R is the number of repressors, N_{NS} is the number of nonspecific protein binding sites (usually taken to the be the length of the genome), $\Delta\epsilon_{pd} = \epsilon_{pd}^{S} - \epsilon_{pd}^{NS}$ is the difference between the specific and nonspecific binding energies for RNAP, and $\Delta\epsilon_{pd} = \epsilon_{rd}^{S} - \epsilon_{rd}^{NS}$ is the difference between specific and nonspecific binding energies for the repressor. More negative values of $\Delta\epsilon_{pd}$ and $\Delta\epsilon_{rd}$ indicate stronger binding.

The constant of proportionality between gene expression and p_{bound} depends on details such as the rate of transcription initiation while RNAP is in the bound state, the mRNA degradation rate, the translation rate, the protein degradation rate, and the rate of cell division. A convenient way to sidestep the need to know all these parameters is to simply consider the ratios of gene expression levels under different intracellular conditions. For instance, one can define the *Repression* as the ratio of gene expression in the presence of a repressor TF to gene expression in the absence of the repressor TF. Then (as long as the parameters mentioned above don't depend on TF concentration), the repression depends only on the ratio of p_{bound} in the presence of repressor to p_{bound} in the absence of repressor, eliminating the possibly unknown constant of proportionality. For the simple repression example in Figure 1.4, the repression is given by

$$\text{Repression} = \frac{1 + \frac{P}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{pd}}{k_B T}\right)}{1 + \frac{P}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{pd}}{k_B T}\right) + \frac{R}{N_{\text{NS}}} \exp\left(-\frac{\Delta\epsilon_{rd}}{k_B T}\right)}. \tag{1.2}$$

Thus, even if the exact constant of proportionality is unknown, direct comparisons between theory and experiment can be made by taking the ratio of two gene expression measurements and comparing with Equation 1.2 [35].

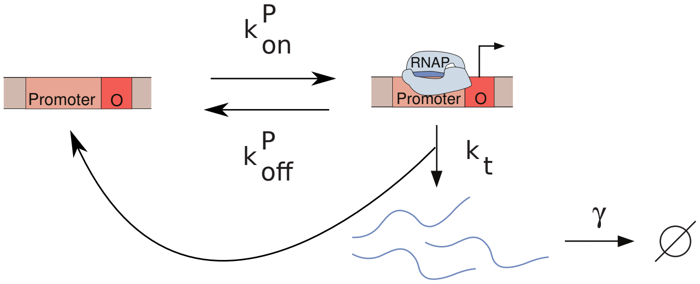
This framework has been successfully applied to gene regulation in a wide variety of contexts in both prokaryotes and eukaryotes. The earliest example is a pair of classic papers by Ackers, Johnson, and Shea in which transcriptional regulation of the lambda phage P_{R} and P_{RM} promoters was modeled using the approach outlined above [49, 50]. The same type of analysis has been applied to aspects of the *lac* promoter in *E. coli* including the energetics of DNA looping [51]. Thermodynamic analysis of the *lac* promoter culminated in the work of Kuhlman *et al* who presented a complete model of regulation at the *lac* promoter incorporating all known molecular interactions, and rigorously tested this model in a series of carefully designed experiments [52]. See also [53] and Chapter 2 of this work for an extension of these results. Bintu *et al* have systematized this overall approach and presented a thermodynamic analysis of 10 common regulatory motifs in prokaryotes [54, 55].

Thermodynamic models have also been applied successfully in the eukaryotic context including yeast promoters [56] and the expression of genes responsible for segmentation in *Drosophila* development [57]. In eukaryotes, an additional complication arises from the presence of nucleosomes, which form the basic building block of chromatin and consist of 147 bp of DNA wrapped around a histone protein octamer. DNA that is part of a nucleosome is inaccessible to binding by TFs or RNAP, and hence the presence or absence of nucleosomes in the promoter region has a significant effect on transcription. Like TFs, nucleosomes have distinct sequence preferences and thus nucleosome positioning can be encoded by the DNA sequence of the genome [58, 59]. Nucleosome occupancy is thus an additional molecular species to be accounted for when enumerating the possible states in a thermodynamic model [16]. Proteins similar to eukaryotic histones appear in prokaryotes, including H-NS, HU, and StpA. Although it appears that they too can play a significant role in transcriptional regulation, this aspect of prokaryotic regulation has garnered somewhat less attention than in eukaryotes. Later on in this thesis, we will explore the regulation of an *E. coli* promoter in which H-NS appears to play a role, the *mscL* promoter.

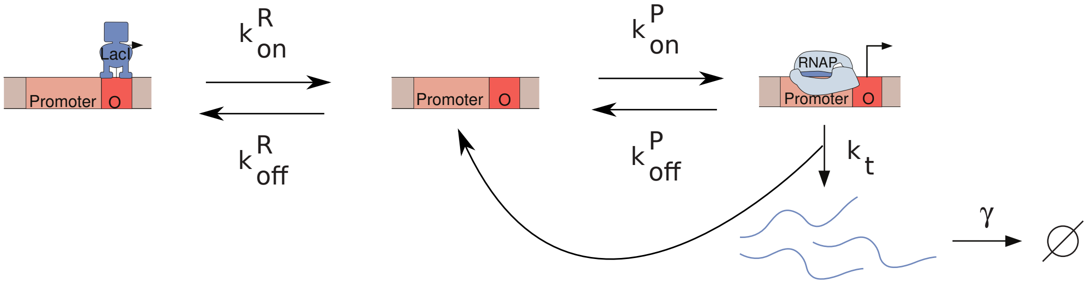
A commonly used three-step model of transcription initiation comprises the following steps: (1) (reversible) closed complex formation; (2) (irreversible) open complex formation; and (3) (irreversible) promoter clearance (which itself is comprised of rounds of abortive initiation followed by clearance and RNA chain elongation) [60–62]. We will consider a slightly simplified two-step version of the model in which closed complex formation occurs reversibly with association rate k_{on}^P and dissociation rate k_{off}^P, and transcription is initiated from the closed complex at a rate k_t. This simplification is appropriate in the limit that the the promoter escape rate is much larger than the open complex formation rate (*i.e.*, in the limit that transcription initiation has open complex formation as a single rate-limiting step, which appears empirically to be the case [40, 61]). Figure 1.5a shows a schematic of this two-step model of transcription initiation. This schematic can be applied directly to the case of constitutive expression, where transcription occurs independently of any transcription factors.

The standard assumption in equilibrium statistical mechanical models of gene expression is that gene expression is proportional to the probability that RNA polymerase is bound at the promoter, given equilibrium between the unbound and bound (*i.e.*, closed complex) states. In the language of rate constants in Figure 1.5a, this probability is given by

$$p_{\text{bound}} = \frac{k_{on}^P}{k_{on}^P + k_{off}^P}. \tag{1.3}$$



(a) Schematic of the kinetics of transcription for constitutive expression. The polymerase binds the promoter at rate k_{on}^P, forming a closed complex, and unbinds at rate k_{off}^P. Transcripts are produced from the bound state at rate k_t, and the system reverts to the unbound state upon production of an mRNA transcript. All intermediate steps between formation of the closed complex and production of a transcript (*e.g.*, open complex formation, abortive initiation) are subsumed into the single effective rate k_t, which is appropriate as long as these intermediates have a single rate-limiting step (generally taken to be open complex formation).



(b) Schematic of the kinetics of transcription for simple repression. The two states on the right of the schematic (promoter unbound and promoter bound by RNAP) are the same as in part (a); additionally, the promoter can be bound by a repressor such as LacI (leftmost state in schematic). Transcription cannot occur when the repressor is bound. The repressor binds at rate k_{on}^R and unbinds at rate k_{off}^R.

Figure 1.5: Schematics of the kinetics of transcription for constitutive expression and simple repression.

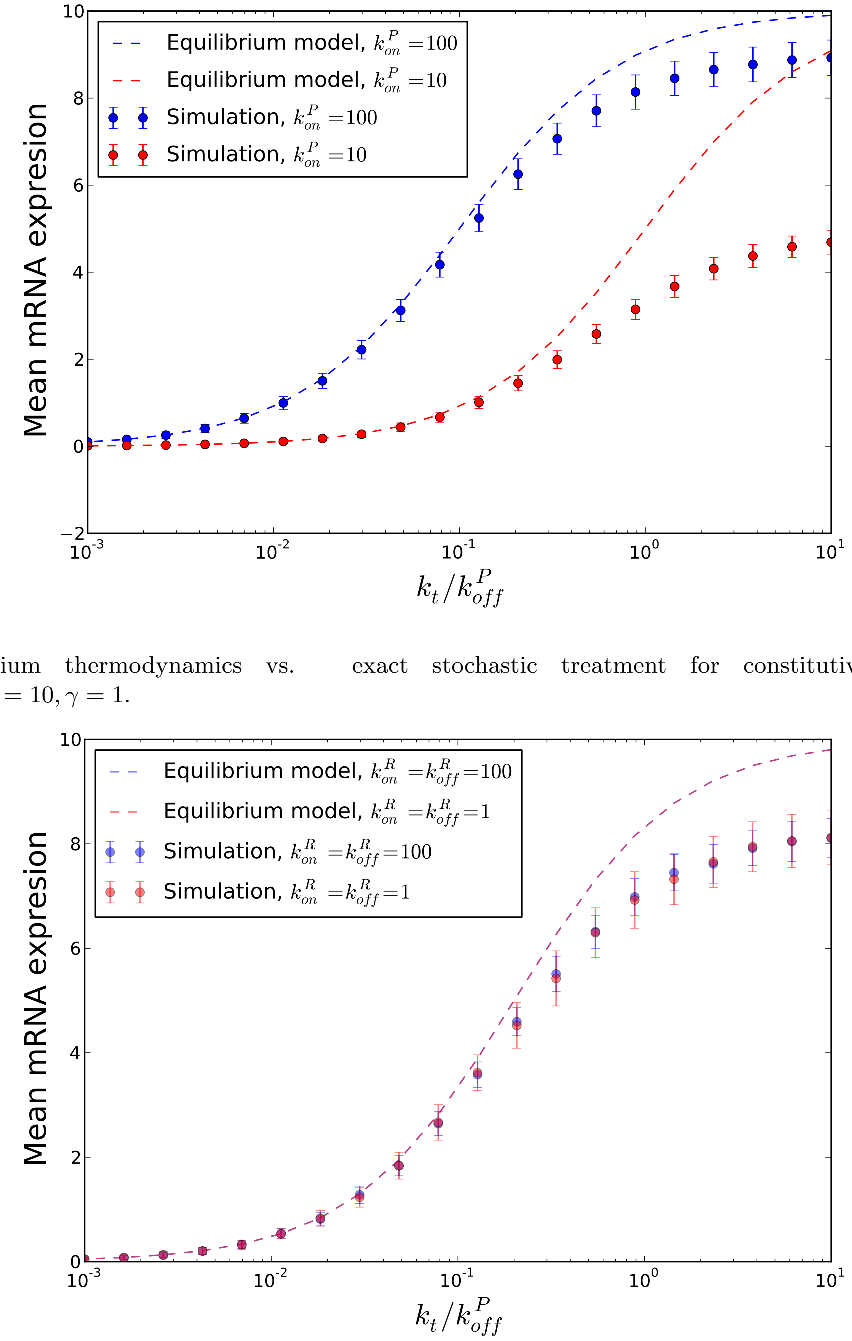
The constant of proportionality is given by the transcription rate k_t divided by the degradation rate γ, so that the overall predicted level of mean gene expression is given by

$$\langle \text{mRNA} \rangle = \frac{k_t}{\gamma} \frac{k_{on}^P}{k_{on}^P + k_{off}^P}. \tag{1.4}$$

With respect to our model of constitutive transcription, under which conditions do we expect this assumption to hold? A moment's reflection reveals that as long as the initiation rate from the closed complex k_t is much less than the dissociation rate of the closed complex k_{off}^P (*i.e.*, $k_t << k_{off}^P$) there will be many association and dissociation events for each transcription event, and the quasiequilibrium description will hold. If k_t is comparable to or larger than k_{off}^P, the system will be continually driven out of equilibrium by irreversible transcription events, and an equilibrium description is not appropriate. In Figure 1.6a, we compare the predictions of the equilibrium model (given in Equation 1.4) with the gene expression levels obtained by performing Gillespie simulations of the scenario shown in Figure 1.5a, for a range of values of k_t/k_{off}^P. The Gillespie algorithm is a well-known algorithm for performing exact stochastic simulations of chemical reaction networks [63]. Briefly, the algorithm entails enumerating all possible reactions the system can undergo (*e.g.* production of an mRNA transcript, degradation of a transcript, association and dissociation of RNAP). At each time step of the simulation, the total rate k_{tot} for all possible reactions is calculated as the sum of the rates of the possible reactions, and the length of time until the next reaction occurs is drawn from an exponential distribution with mean $1/k_{tot}$. The particular reaction that occurs is chosen randomly weighted by the rates of the possible reactions. See Chapter 5.2.3 for additional information.

As expected, the simulations show that the equilibrium model accurately predicts gene expression when $k_t << k_{off}^P$, but diverges when $k_t \sim k_{off}^P$. There is good (although indirect) experimental evidence that this condition does hold for *E. coli* promoters. For instance, Hawley and McClure performed abortive initiation assays in an *in vitro* transcription reaction, and found that the delay time between addition of RNAP to the reaction and open complex formation was consistent with rapid equilibrium of the closed complex and open complex formation as a single rate-limiting step [61]. In a related assay, the same authors added a repressor TF (λ cI) to a similar *in vitro* reaction and found that the formation of open complexes was immediately (within measurement precision) halted, consistent with a picture in which rapid dissociation of RNAP immediately allows the repressor to bind, while existing open complexes are unaffected [40]. To date, the rates of closed complex formation and dissociation have not been directly measured *in vivo.*

What happens when transcription factors are involved? We next examine a slightly more complicated scenario where transcription can be turned off by the binding of a repressor TF. A schematic of the possible transitions for this system is shown in Figure 1.5b. The predicted equilibrium occupancy



(a) Equilibrium thermodynamics vs. exact stochastic treatment for constitutive expression. $k_{on}^P = 100, k_t = 10, \gamma = 1$.

(b) Equilibrium thermodynamics vs. exact stochastic treatment for simple repression. $k_{on}^P = 100, k_t = 10, \gamma = 1$, k_{on}^R and k_{off}^R as indicated in legend.

Figure 1.6: **Gillespie simulations of scenarios depicted in Figure 1.5.** For both constitutive expression and simple repression, the thermodynamic model accurately predicts expression when $k_t << k_{on}^P$ but diverges when $k_t \sim k_{on}^P$. The fact that the thermodynamic model overestimates expression is due to the fact that transcription drives the system into the RNAP unbound state, and thus the thermodynamic model overestimates the time spent in the RNAP bound state. In the case of simple repression, the rates of repressor binding and unbinding do not affect the accuracy of the thermodynamic model. To compute each data point, 100 Gillespie simulations are initiated at each value of k_t/k^{off} and run until reaching equilibrium, at which the mean mRNA levels are computed.

of the promoter is slightly more complex in this case and given by (compare with Equation 1.1)

$$p_{\text{bound}} = \frac{\frac{k_{on}^P}{k_{off}^P}}{1 + \frac{k_{on}^P}{k_{off}^P} + \frac{k_{on}^R}{k_{off}^R}}, \tag{1.5}$$

yielding a predicted gene expression level of

$$\langle \text{mRNA} \rangle = \frac{k_t}{\gamma} \frac{\frac{k_{on}^P}{k_{off}^P}}{1 + \frac{k_{on}^P}{k_{off}^P} + \frac{k_{on}^R}{k_{off}^R}}. \tag{1.6}$$

As in the previous example, we can perform Gillespie simulations to exactly simulate the stochastic reactions schematized in Figure 1.5b, and compare the results of these exact simulations with the equilibrium binding prediction of Equation 1.6. We do so for two sets of values of the repressor association and dissociation rates, one set exhibiting kinetics on the same timescale as RNAP binding/unbinding, and one set exhibiting slower kinetics than RNAP binding/unbinding. In Figure 1.6b, we see that again, it is the ratio of the RNAP dissociation rate k_{off}^P to the transcription initiation rate that determines the accuracy of the equilibrium description. Interestingly, the relative magnitude of the transcription factor kinetics does not play a role - the equilibrium thermodynamics description accurately predicts mean gene expression even for TF kinetics substantially slower than RNAP kinetics. Thus, it is not necessary to postulate that all relevant timescales in the system be faster than transcription initiation kinetics - only that the RNAP unbinding rate be faster than k_t.

Of course, the conclusions drawn from these models are only valid insofar as the schematics in Figure 1.5 are an accurate depiction of reality. Nonetheless, it seems reasonable to conclude that equilibrium statistical mechanics descriptions of gene regulation are likely to be applicable in a broad range of situations. Of particular noteworthiness is the fact that the applicability of equilibrium assumptions is determined only by the relative rates of RNAP dissociation (k_{off}^P) and transcription initiation from the closed complex (k_t), and not by the rates of transcription factor kinetics. This is a good thing since the timescales of transcription factor association and dissociation are frequently on the order of minutes, whereas transcription events can occur on the timescale of seconds. Thus it would be unpromising indeed for the general applicability of thermodynamic models if TF kinetics were required to occur on faster timescales than transcription.

1.4 Stochastic chemical kinetics model of transcriptional regulation

In a preceding paragraph, I wrote that “the equilibrium thermodynamics description accurately predicts mean gene expression even for TF kinetics substantially slower than RNAP kinetics.” This statement is correct, but somewhat incomplete. The reason why it is incomplete is immediately evident upon inspection of the full probability distribution functions for mRNA expression for each of the two cases considered (slow and fast TF kinetics). As seen in Figure 1.7, while it is true that both distributions have the same mean, the distribution corresponding to slow TF kinetics is much broader than the distribution corresponding to fast TF kinetics. In fact, in the limit that TF kinetics are much slower than mRNA degradation, the distribution will be bimodal. For fast TF kinetics, the distribution is reasonably well-characterized by the mean value: the distribution is unimodal and centered on the mean value. For slow TF kinetics, the mean value does a poor job of characterizing the distribution: the distribution is not centered on the mean, and is very broad. Although the thermodynamic model correctly predicts the mean expression for slow TF kinetics, this prediction is arguably not a particularly useful or relevant characterization of gene expression from the promoter. We are thus motivated to consider models of gene expression that allow us to make predictions about higher moments of the probability distribution function for gene expression.

The fundamental theoretical tool for characterizing stochastic gene expression is the *master equation*. The master equation is essentially a way of keeping track of the transitions between states for a Markov process, and can be applied to any Markov process. A Markov process is a stochastic process with no memory, for which the state of the system at time t_3 depends only on the state at time t_2, and not on any previous history of the system. In its most general form, the master equation is written

$$\frac{dP_n}{dt} = \sum_{n'} \left(W_{n'n} p_{n'} - W_{nn'} p_n \right), \tag{1.7}$$

where $W_{n'n}$ is the rate of transitions from state n' to state n, and $W_{nn'}$ is the rate of transitions from state n to state n'. The interpretation of this equation is quite straightforward: to determine the rate at which the probability of being in state n changes, simply add up the transition rates from all other states n into state n', weighted by the current probability of being in state n, and subtract the sum of the transition rates from state n' to all other states n, again weighted by the current probability of being in state n. In this most general form, the master equation can be applied to any Markov process. It might be reasonable to ask whether modeling gene expression as a Markov process is appropriate. After all, it is often the case that delay times introduced by processes such as mRNA processing, protein folding, and protein maturation are an important part of the dynamics of gene expression. The resolution to this apparent dilemma is that although on a macroscopic scale non-memoryless phenomena like delay times are manifested, on a microscopic scale each individual protein molecule is still undergoing memoryless transitions from one state to another. For instance, in the context of eukaryotic transcription a number of processes have to take place before a messenger RNA can be translated, such as splicing, addition of a polyadenylated tail, and export from the nucleus to the cytoplasm; these steps cumulatively create a delay between transcription and transla-

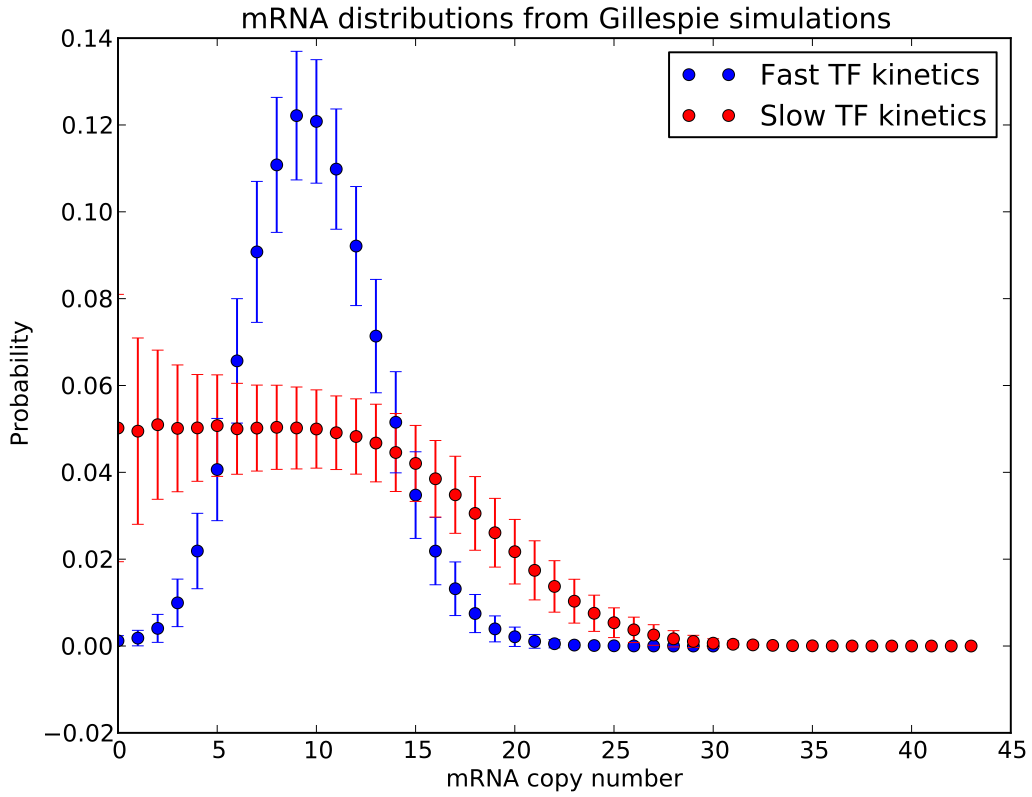


Figure 1.7: **mRNA copy number distributions for fast and slow repressor kinetics.** Gillespie simulations were performed of the scenario depicted in Figure 1.5b, and the resulting steady-state mRNA distributions were computed from the results of the simulations. For slow TF kinetics, $k_{on}^R = k_{off}^R = 1$; for fast kinetics, $k_{on}^R = k_{off}^R = 100$; $\gamma = 1$. Both distributions have the same mean value (10 mRNA), but the distribution corresponding to fast TF kinetics is much more peaked around the mean value than the broad, long-tailed distribution for slow TF kinetics.

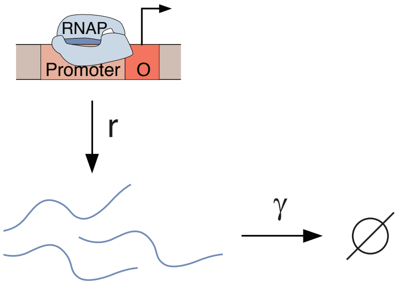


Figure 1.8: **Constitutive expression schematic.** This schematic is a simplified version of Figure 1.5 in which the kinetics of RNAP binding, unbinding, and transcript initiation are subsumed into a single effect transcription rate r as described in Equation 1.8. mRNA transcripts are degraded with probability γ per unit time per transcript.

tion. But at a molecular level, each individual mRNA is simply diffusing around until it encounters the appropriate enzyme or transporter to effect each of these reactions. From the perspective of an individual mRNA, the time that has passed since transcription is irrelevant to the probability of encountering the spliceosome in the next instant dt.

Equation 1.7 gives the most general form of the master equation. While dealing with gene expression, we will generally be working with a particular class of Markov processes called "birth and death processes" [64]. For this class of processes, the state of the system can only increase in increments of one: *i.e.*, $W_{ij} = 0$ if $|i - j| > 1$. For instance, if n refers to the number of mRNA transcripts present in the cell, n, can increase by one via a transcription event, or can decrease by one via a degradation event. In Figure 1.8, we show a schematic of a simple birth and death model of constitutive (unregulated) transcription. mRNA transcripts are produced with constant probability per unit time at rate r, and are degraded with constant probability per mRNA per unit time at rate γ. Note that in Figure 1.8 we are no longer explicitly considering the transitions between the states in which RNAP is bound and unbound. As long as these transition rates (k^P_{off} and k^P_{on}) are fast compared to the initiation rate from the closed complex k_t, these pre-initiation dynamics can be modeled without loss of accuracy using a single effective transcription rate r, which is related to k^P_{off}, k^P_{on} and k_t by the following expression:

$$r = \frac{k^P_{on}}{k^P_{on} + k^P_{off}} k_t. \tag{1.8}$$

This expression for r can be interpreted as the fraction of time for which the promoter is bound by RNAP times the transcription rate k_t from the closed complex.

The master equation corresponding to Figure 1.8 can be written:

$$\frac{dp(m,t)}{dt} = rp(m-1,t) + \gamma(m+1)p(m+1,t) - rp(m,t) - \gamma mp(m,t). \tag{1.9}$$

The first two terms of the master equation have positive signs and are concerned with transitions to the state of having m mRNA: one could start with $m-1$ mRNA, and produce one (first term), or one could start with $m+1$ mRNA, and degrade one (second term). The third and fourth terms deal with transitions away from m mRNA and hence have negative signs: one could start with m mRNA and produce one (third term), or start with m mRNA and degrade one (fourth term). In this work we will principally be concerned with the steady state probability distribution. It can be shown (by setting the left hand side of Equation 1.9 to zero and directly substituting the following expression) that the steady-state solution to Equation 1.9 is a Poisson distribution with mean r/γ:

$$p(m) = \frac{(r/\gamma)^m}{m!} e^{-r/\gamma}. \tag{1.10}$$

For more complicated scenarios involving regulation by transcription factors, a closed form solution to the master equation will not in general be available. However, we can still make progress by calculating the various moments of the distribution analytically. For instance, to compute the steady-state mean of Equation 1.9, we set the left hand side to zero, multiply by m, and sum from $m = 0$ to infinity:

$$\begin{aligned}
0 &= r\sum_{m=0}^{\infty} mp(m-1) + \gamma\sum_{m=0}^{\infty} m(m+1)p(m+1) - r\sum_{m=0}^{\infty} mp(m) - \gamma\sum_{m=0}^{\infty} m^2 p(m), && (1.11)\\
0 &= r\sum_{m=0}^{\infty} (m+1)p(m) + \gamma\sum_{m=0}^{\infty} m(m-1)p(m) - -r\sum_{m=0}^{\infty} mp(m) - \gamma\sum_{m=0}^{\infty} m^2 p(m), && (1.12)\\
0 &= r\langle m\rangle + r + \gamma\langle m^2\rangle - \gamma\langle m\rangle - \gamma\langle m^2\rangle, && (1.13)\\
\gamma\langle m\rangle &= r, && (1.14)\\
\langle m\rangle &= \frac{r}{\gamma}, && (1.15)
\end{aligned}$$

where we have invoked the normalization condition that $\sum_{m=0}^{\infty} p(m) = 1$. A similar procedure can be carried out for all moments of the distribution. This means that, although we may not be able to obtain analytic solutions for the full probability distribution function for more complicated regulatory scenarios, we can still analytically compute useful properties of the distribution such as the noise strength (standard deviation divided by mean) and Fano factor (variance divided by mean).

The existence of noise in gene expression has been noted as early as 1976 [65, 66], and was vividly placed in a quantitative framework in a 2002 paper by Elowitz et al [67]. In this work, accompanied by theoretical work providing the mathematical justification [68], Elowitz and coworkers showed how to experimentally decompose variability in gene expression into so-called “intrinsic” and “extrinsic” components. “Intrinsic” variability refers to the variability resulting from the inherent stochasticity of molecular reactions such as transcription factor binding and unbinding and transcription initiation.

"Extrinsic" variability refers to the variability resulting from the fact that each of the molecular rates depicted in Figure 1.5b is itself subject to variation due to *e.g.* fluctuations in repressor or RNA polymerase copy numbers. Notably, the mathematical breakdown of intrinsic vs extrinsic noise given in reference [68] relies on the assumption that extrinsic fluctuations are slower than intrinsic fluctuations which is probably true at the level of mRNA expression, but not necessarily at the protein expression level [69]. Other noteworthy experimental investigations of variability in gene expression include a 2005 experiment by Ido Golding and coworkers, in which the MS2 mRNA tagging system was used to monitor the production of mRNA molecules in essentially "real time," allowing the authors to observe the distribution of waiting times between mRNA production events [70]. One of the more surprising results to emerge from these experiments was the observation that even in the fully induced state, the promoter still exhibited pronounced periods of inactivity, even though in principle the repressor TF should have been inactivated by the presence of the inducer molecule IPTG.

More recently, a series of publications have advanced the hypothesis that noise is "universal" in prokaryotes in the sense that the level of variability is dictated solely by the mean level of gene expression and not by the specific molecular details of promoter architecture such as transcription factor binding site locations and strengths. This hypothesis was advanced most explicitly in a 2011 paper by So *et al*, in which the authors measured the level of variability in mRNA copy number for a variety of *E. coli* promoters under a variety of induction conditions [71]. Similar results were obtained in a study by Taniguchi *et al* of transcription from a library of some 2000 genes in *E. coli* [72]. This observation of universality was extended to other microorganisms in a work by Salman *et al* [73]. These experimental observations, combined with the observation from reference [70] of "burst-like" mRNA production even in the fully induced case, have led to speculation that some as-yet uncharacterized mechanism universally causes *E. coli* promoters to exhibit periods of activity and inactivity, regardless of transcription factor binding. One possible mechanism is derived from the fact that transcriptional silencing by nucleosomes is a well-known phenomenon in eukaryotes, and thus it seems plausible that nucleoid-associated prokaryotic proteins homologous to eukaryotic histones could be playing a similar role in prokaryotes. Later in this thesis, this question of universality will be addressed directly; I will briefly note here that our experimental results argue rather strongly against universality.

In any case we have seen that noise in gene expression is a fact of life, in both prokaryotes and eukaryotes. Given that this is so, the question naturally arises whether noise is simply an uncomfortably reality that life simply has to deal with, or whether it can play an adaptive or functional physiological role in living organisms [74]. One intriguing hypothesis is that phenotypic diversity resulting from gene expression noise could confer fitness on a genetically identical population by serving as a "bet-hedging" strategy against fluctuating environmental conditions [75–77]. Examples of environmental fluctuations include shifts from well-mixed to stagnant liquid growth conditions, or changes in the availability of sugars to metabolize [78, 79]. The idea is that if a small subpopulation of an overall population adopts a phenotype suitable for a different environmental condition, if the environment changes rapidly the subpopulation will be poised to rapidly succeed in the new environment. A key element of the bet-hedging concept is incomplete information about the environment. With perfect knowledge of environmental conditions, bet-hedging would be suboptimal, as the best strategy would be to deterministically adopt the the phenotype most suited to environmental conditions. However, if conditions change unpredictably, or if acquiring information about the environment is too costly, then bet-hedging can make sense.

As discussed above, theoretical efforts to characterize stochasticity in gene expression have centered around solving the chemical master equation. The majority of this work has been done starting around the year 2000, reflecting the fact that it is only relatively recently that single-cell techniques have allowed noise in gene expression to be characterized at the experimental level. In the early 2000s, theorists calculated the gene expression probability distributions resulting from "bursty" gene expression [80, 81]. This corresponds to the scenario illustrated in Figure 1.5b in the limit that the periods of active gene expression are short compared to the lifetime of an mRNA. In that case, it is appropriate to model the transcripts as being produced essentially all at once. The resulting distribution is a gamma distribution (or equivalently its discrete counterpart, the negative binomial distribution), and is characterized by having a longer tail in the positive direction than a Poisson distribution. Later, Raj and coworkers found an analytic expression for the resulting probability distribution for the general case (*i.e.*, the length of active period is not necessarily short compared with the mRNA lifetime) [82, 83]. In work published in 2008, and also in Chapter 5 of this thesis (published in 2010 in PLoS Computational Biology), Sanchez *et al* systematized and extended these efforts to compute the variability for a variety of promoter architectures, much as Bintu *et al* had previously done for mean gene expression using thermodynamic models [36, 54]. Notably, these theoretical efforts yield predictions for relationships between noise and mean expression that distinctly depend on the details of promoter architecture, in contrast with some experimental results described above.

While the work described in the preceding paragraph has been largely focused on computing how gene expression variability depends on the details of regulation of a particular isolated gene, other noteworthy efforts have examined how noise propagates through networks of genes [84], and have derived fundamental limits on the ability of networks of interacting genes to suppress fluctuations in gene expression [85, 86]. At the same time, Cox and Munsky have showed how fluctuations in gene expression can be used to infer properties of gene expression networks [87, 88]. Another related line of inquiry has been pursued by William Bialek and coworkers including Gaspar Tkacik, Thierry Mora, and Aleksandra Walczak. These researchers have looked in-depth at the implications of noise in gene expression for information processing and signal transduction. The basic idea here is that the concentration(s) of various transcription factor(s) encode some information about the environment. In the process of converting from transcription factor concentrations (the input) to gene expression level (the output), some information is inevitably lost due to noise in gene expression. These efforts have sought to characterize how information is integrated and transduced by regulatory DNA, and to examine the flow of information through genetic regulatory networks [89–92]. Finally, the effect of the partitioning of proteins between daughter cells at cell division on cell-to-cell variability in protein copy numbers has been examined. The authors concluded that in many cases, the variability due to cell partitioning can be as important or more than the variability due to transcription [93], highlighting the need to exercise caution in interpreting the results of gene expression variability measurements.

The remainder of this thesis can be briefly summarized as follows. In **Chapter 1**, we employ the thermodynamic modeling framework in the regulatory context of the wild-type *lac* operon. We use models of the sequence-dependent binding energies of the relevant proteins (CRP, LacI, and RNAP) to directly construct a genotype to phenotype map for the level of repression exhibited by the *lac* promoter. In **Chapter 2**, we examine how a model of the sequence-dependent binding energy of RNAP can be used in conjunction with a thermodynamic model of gene expression to design promoter to yield targeted levels of gene expression, in the regulatory contexts of constitutive expression and simple repression.In **Chapter 3**, we introduce a high-throughput promoter mutagenesis assay called Sort-Seq and use it to explore the transcriptional regulation of the mechanosensitive channel of large conductance, whose transcriptional regulation had previously been almost completely uncharacterized. In **Chapter 4**, we introduce a theoretical modeling framework for predicting the level of cell-to-cell variability in gene expression as a function of the promoter architecture, and use this framework to explore the noise properties of a number of common *E. coli* regulatory motifs. Finally, in **Chapter 5**, we use mRNA FISH to test these theoretical predictions, and find that the noise does distinctly depend on the promoter architectures in play.

Bibliography

[1] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of Experimental Medicine*, 79(2):137–158, 1944.

[2] Erwin Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209, 1950.

[3] D. Elson and E. Chargaff. On the desoxyribonucleic acid content of sea urchin gametes. *Experientia*, 8(4):143–145, 1952.

[4] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

[5] F. H. C. Crick, Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, Dec 1961.

[6] H. Bremer and P. P. Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In Frederick C. Neidhardt et al., editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1553–1569. ASM Press, Washington DC, 1996.

[7] Brian J. Paul, Wilma Ross, Tamas Gaal, and Richard L. Gourse. rRNA transcription in *Escherichia coli*. *Annual Review of Genetics*, 38(1):749–770, 2004. PMID: 15568992.

[8] Matthew Scott, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(November), 2010.

[9] L Perfeito, S Ghozzi, Johannes Berg, Karin Schnetz, and M Lässig. Nonlinear fitness landscape of a molecular pathway. *PLoS genetics*, 7(7):1–10, 2011.

[10] E. Dekel and U. Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–92, 2005.

[11] Jeffery Errington. *Bacillus subtilis* sporulation: Regulation of gene expression and control of morphogenesis. *Microbiological Reviews*, 57(1), March 1993.

[12] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muiz-Rascado, Jair S. Garca-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martnez-Flores, Alejandra Medina-Rivera, Gerardo Salgado-Osorio, Shirley Alquicira-Hernndez, Kevin Alquicira-Hernndez, Alejandra Lpez-Fuentes, Liliana Porrn-Sotelo, Araceli M. Huerta, Csar Bonavides-Martnez, Yalbi I. Balderas-Martnez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Vernica Jimnez-Jacinto, Leticia Vega-Alvarado, Victor del Moral-Chvez, Alfredo Hernndez-Alvarez, Enrique Morett, and Julio Collado-Vides. Regulondb v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[13] Geeta J. Narlikar, Hua-Ying Fan, and Robert E. Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, Feb 2002.

[14] Guo-Cheng Yuan, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309(5734):626–630, 2005.

[15] Felix H. Lam, David J. Steger, and Erin K. O/'Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–250, May 2008.

[16] E. Segal and J. Widom. From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat Rev Genet*, 10(7):443–56, 2009.

[17] M. S. Luijsterburg, M. C. Noom, G. J. Wuite, and R. T. Dame. The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: A molecular perspective. *J Struct Biol*, 2006.

[18] Wenqin Wang, Gene-Wei Li, Chongyi Chen, X. Sunney Xie, and Xiaowei Zhuang. Chromosome organization by a nucleoid-associated protein in live bacteria. *Science*, 333(6048):1445–1449, 2011.

[19] Charles J Dorman. H-NS: A universal regulator for a dynamic genome. *Nature Reviews Microbiology*, 2(5):391–400, 2004.

[20] Ferric C Fang and Sylvie Rimsky. New insights into transcriptional regulation by h-ns. *Current Opinion in Microbiology*, 11(2):113 – 120, 2008. Cell Regulation.

[21] H. M. Lim, D. E. Lewis, H. J. Lee, M. Liu, and S. Adhya. Effect of varying the supercoiling of DNA on transcription and its regulation. *Biochemistry*, 42(36):10718–25, 2003.

[22] A. Travers and G. Muskhelishvili. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol*, 3(2):157–69, 2005.

[23] H. M. Salis, E. A. Mirsky, and C. A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*, 27(10):946–50, 2009.

[24] I. Chen and D. Dubnau. DNA uptake during bacterial transformation. *Nat Rev Microbiol*, 2(3):241–9, 2004.

[25] Susan Gottesman and Gisela Storz. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology*, 3(12):a003798, 2011.

[26] Susan Gottesman. The small rna regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, 58:303–328, 2004.

[27] Chase L Beisel and Gisela Storz. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiology Reviews*, 34(5):866–82, September 2010.

[28] Karen L. Maxwell, David Wildes, Arash Zarrine-Afsar, Miguel A. De Los Rios, Andrew G. Brown, Claire T. Friel, Linda Hedberg, Jia-Cherng Horng, Diane Bona, Erik J. Miller, Alexis Valle-Blisle, Ewan R.G. Main, Francesco Bemporad, Linlin Qiu, Kaare Teilum, Ngoc-Diep Vu, Aled M. Edwards, Ingo Ruczinski, Flemming M. Poulsen, Birthe B. Kragelund, Stephen W. Michnick, Fabrizio Chiti, Yawen Bai, Stephen J. Hagen, Luis Serrano, Mikael Oliveberg, Daniel P. Raleigh, Pernilla Wittung-Stafshede, Sheena E. Radford, Sophie E. Jackson, Tobin R. Sosnick, Susan Marqusee, Alan R. Davidson, and Kevin W. Plaxco. Protein folding: Defining a standard set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Science*, 14(3):602–616, 2005.

[29] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa. Quantitative characteristics of gene regulation by small RNA. *PLoS Biol*, 5(9):e229, 2007.

[30] Yishai Shimoni, Gilgi Friedlander, Guy Hetzroni, Gali Niv, Shoshy Altuvia, Ofer Biham, and Hanah Margalit. Regulation of gene expression by small non-coding RNAs: A quantitative view. *Molecular systems biology*, 3(138):138, January 2007.

[31] M R Parsek and E P Greenberg. Acyl-homoserine lactone quorum sensing in gram-negative bacteria: A signaling mechanism involved in associations with higher organisms. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):8789–93, August 2000.

[32] Michael E Hibbing, Clay Fuqua, Matthew R Parsek, and S Brook Peterson. Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*, 8(1):15–25, January 2010.

[33] Marcus Miethke and Mohamed a Marahiel. Siderophore-based iron acquisition and pathogen control. *Microbiology and Molecular Biology Reviews*, 71(3):413–51, September 2007.

[34] Mattias Rydenfelt. *The Combinatorics of Transcriptional Regulation.* PhD thesis, California Institute of Technology, 2014.

[35] H. G. Garcia, A. Sanchez, T. Kuhlman, J. Kondev, and R. Phillips. Transcription by the numbers redux: Experiments and calculations that surprise. *Trends Cell Biol*, 2010.

[36] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Computational Biology*, 7(3):e1001100, 2011.

[37] James Q. Boedicker, Hernan G. Garcia, and Rob Phillips. Theoretical and experimental dissection of dna loop-mediated repression. *Physical Review Letters*, 110(1):018101, January 2013.

[38] Axel Cournac and Jacqueline Plumbridge. DNA looping in prokaryotes: Experimental and theoretical approaches. *Journal of Bacteriology*, 195(6):1109–1119, 2013.

[39] RB Lobell and RF Schleif. DNA looping and unlooping by AraC protein. *Science*, 250(4980):528–532, 1990.

[40] D. K. Hawley and W. R. McClure. Mechanism of activation of transcription initiation from the lambda PRM promoter. *J Mol Biol*, 157(3):493–525, 1982.

[41] Virgil A Rhodius, David M West, Christine L Webster, Stephen J W Busby, and Nigel J Savery. Transcription activation at Class II CRP-dependent promoters : The role of different activating regions. *Nucleic Acids Research*, 25(2):326–332, 1997.

[42] W Niu, Y Kim, G Tau, T Heyduk, and R H Ebright. Transcription activation at class II CAP-dependent promoters: Two interactions between CAP and RNA polymerase. *Cell*, 87(6):1123–34, December 1996.

[43] D. F. Browning and S. J. Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, 2004.

[44] T. H. Lee and N. Maheshri. A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol. Syst. Biol.*, 8:576, 2012.

[45] S. L. Dove and A. Hochschild. Use of artificial activators to define a role for protein-protein and protein-DNA contacts in transcriptional activation. *Cold Spring Harb Symp Quant Biol*, 63:173–80, 1998.

[46] SL Dove and Ann Hochschild. Conversion of the ω subunit of Escherichia coli RNA polymerase into a transcriptional activator or an activation target. *Genes & development*, pages 745–754, 1998.

[47] Angela Caramel and Karin Schnetz. Lac and λ repressors relieve silencing of the *Escherichia coli bgl* promoter. activation by alteration of a repressing nucleoprotein complex. *Journal of molecular biology*, 284(4):875–883, 1998.

[48] Rob Phillips, Jane Kondev, Julie Theriot, and Hernan Garcia. *Physical Biology of the Cell*. Garland Science, New York, 2 edition, 2012. Illustrated by Nigel Orme; with problems, solutions, and editorial assistance of Hernan G. Garcia.

[49] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982.

[50] M. A. Shea and G. K. Ackers. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol*, 181(2):211–30, 1985.

[51] G. R. Bellomy, M. C. Mossing, and M. T. Record, Jr. Physical properties of DNA in vivo as probed by the length dependence of the lac operator looping process. *Biochemistry*, 27(11):3900–6, 1988.

[52] T. Kuhlman, Z. Zhang, Jr. Saier, M. H., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6043–8, 2007.

[53] M Razo-Mejia, J Q Boedicker, D Jones, A DeLuna, J B Kinney, and R Phillips. Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. *Physical Biology*, 11(2):026005, 2014.

[54] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[55] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Current Opinion in Genetics and Development*, 15(2):125–35, 2005.

[56] J. Gertz, E. D. Siggia, and B. A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–8, 2009.

[57] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–40, 2008.

[58] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thastrom, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, Aug 2006.

[59] Ilya P. Ioshikhes, Istvan Albert, Sara J. Zanton, and B. Franklin Pugh. Nucleosome positions predicted through comparative genomics. *Nat Genet*, 38(10):1210–1215, Oct 2006.

[60] M. Thomas Record Jr., William S. Reznikoff, Maria L. Craig, Kristi L. McQuade, and Paula J. Schlax. *Escherichia coli* RNA polymerase (Eσ^{70}), promoters, and the kinetics of the steps of transcription initiation. In Frederick C. Neidhardt et al., editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1553–1569. ASM Press, Washington DC, 1996.

[61] W R McClure. Rate-limiting steps in RNA chain initiation. *Proceedings of the National Academy of Sciences of the United States of America*, 77(10):5634–8, October 1980.

[62] WR McClure. Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry*, pages 171–204, 1985.

[63] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

[64] Nicolaas Godfried Van Kampen. *Stochastic Processes in Physics and Chemistry*, volume 1. Elsevier, 1992.

[65] J. L. Spudich and Jr. Koshland, D. E. Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467–71, 1976.

[66] Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[67] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[68] P.S. Swain, M.B. Elowitz, and E.D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99:12795., 2002.

[69] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–72, July 2011.

[70] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[71] L. H. So, A. Ghosh, C. Zong, L. A. Sepulveda, R. Segev, and I. Golding. General properties of transcriptional time series in *Escherichia coli*. *Nature Genetics*, 43(6):554–60, 2011.

[72] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[73] Hanna Salman, Naama Brenner, Chih-kuan Tung, Noa Elyahu, Elad Stolovicki, Lindsay Moore, Albert Libchaber, and Erez Braun. Universal protein fluctuations in populations of microorganisms. *Physical Review Letters*, 108:238105, Jun 2012.

[74] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73, 2010.

[75] D Cohen. Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12(1):119–&, 1966.

[76] Montgomery Slatkin. Hedging one's evolutionary bets. *Nature*, 250:704–705, 1974.

[77] BN Danforth. Emergence dynamics and bet hedging in a desert bee, Perdita portalis. *Proceedings of the Royal Society B: Biological Sciences*, 266(1432):1985–1994, OCT 7 1999.

[78] Hubertus J. E. Beaumont, Jenna Gallie, Christian Kost, Gayle C. Ferguson, and Paul B. Rainey. Experimental evolution of bet hedging. *Nature*, 462(7269):90–93, Nov 2009.

[79] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–6, 2008.

[80] J Paulsson and M Ehrenberg. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review Letters*, 84(23):5447–50, June 2000.

[81] N. Friedman, L. Cai, and X. S. Xie. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Physical Review Letters*, 97(16):–, 2006.

[82] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

[83] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci U S A*, 105(45):17256–61, 2008.

[84] J. M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–9, 2005.

[85] I. Lestas, G. Vinnicombe, and J. Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174–8, 2010.

[86] A Grönlund, P Lötstedt, and Johan Elf. Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nature Communications*, (May):1–5, 2013.

[87] R. S. Cox III, M. G. Surette, and M. B. Elowitz. Programming gene expression with combinatorial promoters. *Mol Syst Biol*, 3:145, 2007.

[88] B. Munsky, B. Trinh, and M. Khammash. Listening to the noise: Random fluctuations reveal gene network parameters. *Mol Syst Biol*, 5:318, 2009.

[89] Samuel F Taylor, Naftali Tishby, and William Bialek. Information and Fitness. 2007.

[90] Gašper Tkačik, Curtis Callan, and William Bialek. Information capacity of genetic regulatory elements. *Physical Review E*, 78(1):1–17, July 2008.

[91] Gašper Tkačik, Aleksandra Walczak, and William Bialek. Optimizing information flow in small genetic networks. *Physical Review E*, 80(3):1–18, September 2009.

[92] Aleksandra M. Walczak, Gašper Tkačik, and William Bialek. Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Physical Review E*, 81(4):1–16, April 2010.

[93] Dann Huh and Johan Paulsson. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100, February 2011.

Chapter 2

Comparison of the theoretical and real-world evolutionary potential of a genetic circuit.

A version of this chapter originally appeared as: M. Razo-Mejia, J.Q. Boedicker, D. Jones, A. DeLuna, J.B. Kinney, and R. Phillips (2014) Comparison of the theoretical and real-world evolutionary potential of a genetic circuit . Physical Biology 11: 026005. doi:10.1088/1478-3975/11/2/026005

Author contribution note: for this chapter, I (DLJ) performed Sort-Seq experiments, wrote data analysis code, and analyzed data to obtain the sequence-dependent binding energy models for CRP, RNAP, and LacI.

2.1 Introduction.

Despite efforts to understand genotypic variability within natural populations [1] and recent interest in fine-tuning genetic circuits for synthetic biology [2], it still remains unclear how, with base pair resolution, the sequence of a gene regulatory region can be translated into output levels of gene expression [3]. Generally, classical population genetics has treated regulatory architectures as changeless parameters, rather than potential evolutionary variables, focusing on changes in protein structure rather than gene regulation. However, genetic regulatory architecture can also determine the variation of traits, and thus the evolutionary potential of these genes [4]. After all, the structure of bacterial promoters dictates interactions among the transcriptional apparatus, and through the modification of this structure, regulatory circuits can be modified to potentially allow cells to occupy different niches [5, 6].

Thermodynamic models of gene regulation have been widely used as a theoretical framework to dissect and understand genetic architectures [7–11]. Such dissections have led to a quantitative understanding of how parameters such as binding energies, transcription factor copy numbers, and the mechanical properties of the DNA dictate expression levels. Recently the development of experimental techniques combining these types of models with cell sorting and high-throughput sequencing have made it possible to understand gene regulation at single-base pair resolution [12–14], as well as to deliberately design promoter architectures with desired input-output functions [15]. These models connect the sequence of a promoter to the output phenotype, making it possible to predict variability and evolutionary potential of gene regulatory circuits.

The *lac* operon has served as a paradigm of a genetic regulatory system for more than 60 years [16, 17]. This operon contains the molecular machinery that some bacterial species, including the model organism *E. coli*, use to import and consume lactose. Extensive quantitative characterization of the regulation of this genetic circuit [18, 19], as well as of the link between fitness and expression of the operon [20–24] make it an ideal system for exploring the evolutionary potential of a regulatory circuit. With previous exhaustive description and quantification of the parameters controlling the expression level of this genetic circuit [19, 25–27] we now have what we think is a nearly complete picture of the regulatory *knobs* that can modify the expression level, shown schematically in Figure 2.1(a). In this article we build upon this understanding by directly linking the sequence of the promoter region with these control parameters, thereby creating a map from genotype to transcriptional output.

Within a collection of *E. coli* isolated from different host organisms we observe significant variability for the regulation of the *lac* operon, as shown in Figure 2.1(b). By characterizing the variability of the regulatory control parameters shown in Figure 2.1(a) within these strains, we identified evolutionary trends in which certain parameters or subsets of parameters are seen to vary more often than others within this collection of natural isolates. Using the map of promoter sequence to transcriptional output, we demonstrated that the regulatory input-output function for the *lac* promoter could account for most of the natural variability in regulation we observed. We then implement the map to explore the theoretical potential for this regulatory region to evolve. This level of analysis gives us clues as to how selection could fine tune gene expression levels according to the environmental conditions to which cells are exposed.

2.2 Results.

2.2.1 Quantitative model of the natural parameters that regulate gene expression

Thermodynamic models of gene regulation have become a widely used theoretical tool to understand and dissect different regulatory architectures [3, 12, 19, 26, 27, 31]. The *lac* promoter is one such regulatory architecture that has been studied in detail [32]. Models have been constructed and experimentally validated for both the wild-type *lac* promoter and synthetic promoter regions built up from the *lac* operon's regulatory components [12, 15, 19, 26, 27, 32–37]

In a simple dynamical model of transcription the number of messenger RNA (mRNA) is propor-

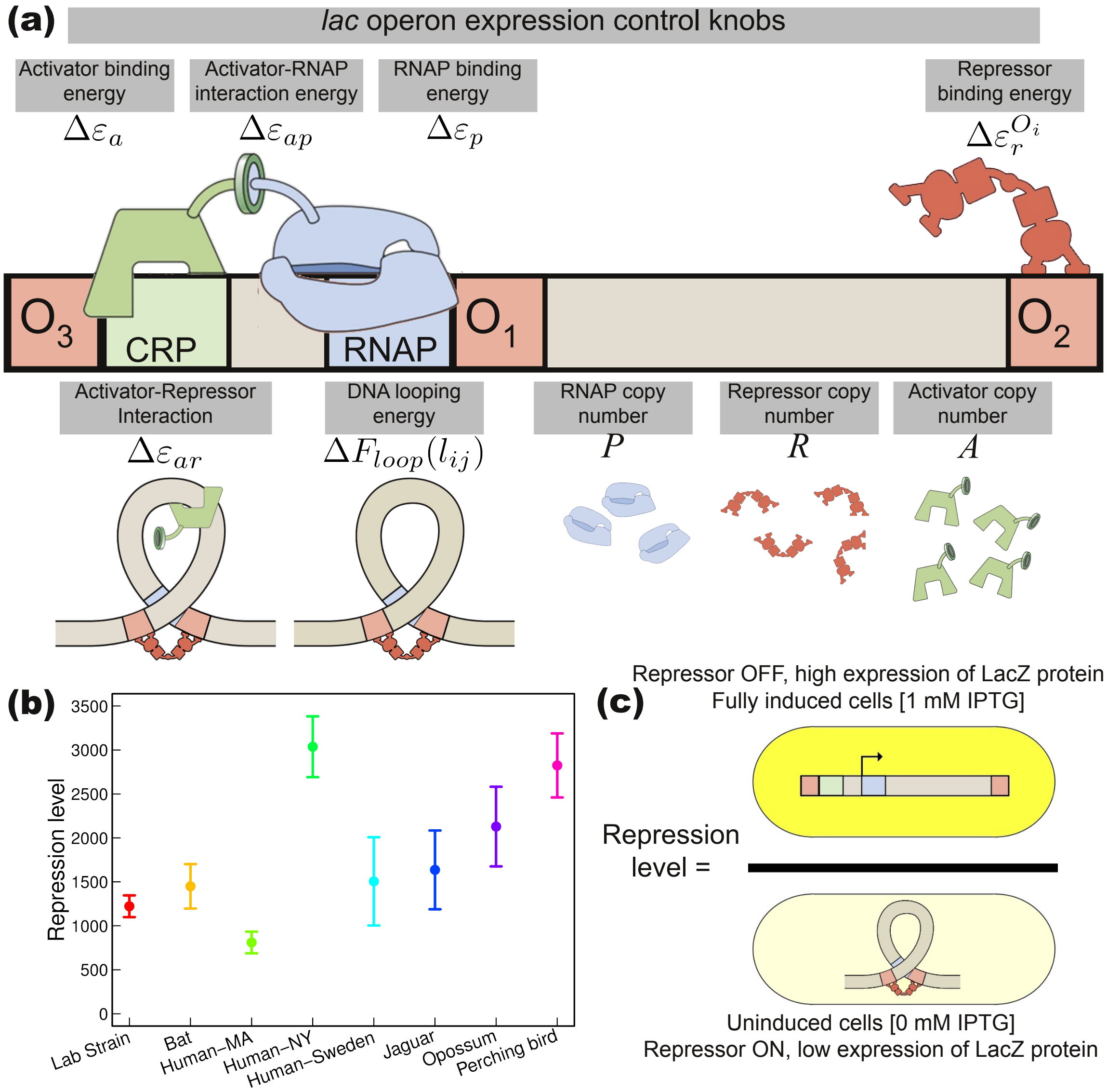


Figure 2.1: (a) Regulatory knobs that control the expression of the *lac* operon and the symbols used to characterize these knobs in the thermodynamic model. The activator CRP increases expression, the Lac repressor binds to the three operators to decreases expression, and looping can lock the repressor onto O_1 leading to increased repression. The interaction energy between RNAP and CRP reflects the stabilization of the open complex formation due to the presence of the activator [28], and the interaction between the Lac repressor and CRP stabilizes the formation of the upstream loop [29]. (b) Variability in the repression level of *E. coli* natural isolates and the lab control strain MG1655. Strains are named after the host organism from which they were originally isolated [30]. Error bars represent the standard deviation from at least three independent measurements. (c) Schematic representation of the repression level, in which the role of the repressor in gene regulation is experimentally measured by comparing the ratio of LacZ proteins in cells grown in the presence of 1 mM IPTG to cells grown in the absence of IPTG. LacZ protein concentrations were measured using a colorimetric assay.

tional to the transcription rate and the degradation rate of the mRNA,

$$\frac{dm}{dt} = -\gamma \cdot m + \sum_i r_i \cdot p_i, \tag{2.1}$$

where γ is the mRNA degradation rate and m is the number of transcripts of the gene per cell;

r_i and p_i are the transcription rate and the probability of state i respectively. We can think of p_i as a measure of the time spent in the different transcriptionally active states. Thermodynamic models assume that the gene expression level is dictated by the probability of finding the RNA *polymerase* (RNAP) bound to the promoter region of interest [7–9]. With a further quasi-equilibrium assumption for the relevant processes leading to transcription initiation, we derive a statistical mechanics description of how parameters such as transcription factor copy number and their relevant binding energies, encoded in the DNA binding site sequence, affect this probability [10]. Quantitative experimental tests of predictions derived from equilibrium models have suggested the reasonableness of the assumption [15, 19, 26, 27], although caution should be used as the equilibrium assumption is not necessarily valid in all cases. The validity of this equilibrium assumption relies on the different time-scales of the processes involved in the transcription of a gene. Specifically the rate of binding and unbinding of the transcription factors and the RNAP from the promoter region should be faster than the open complex formation rate; if so, the probability of finding the RNAP bound to the promoter is given by its equilibrium value [9, 38]. For the case of the Lac repressor, the rate of unbinding from the operator is 0.022 1/s [39], and the binding of an unoccupied operator with 10 repressors per cell occurs at a similar rate [40]. Open complex formation, a rate limiting step in promoter escape, has been measured at a rate of 2×10^{-3} 1/s [41]. Promoter escape is about an order of magnitude slower than the binding and unbinding of the Lac repressor, and this separation of time scales supports the equilibrium assumption for this particular case. We enumerate the possible states of the system and assign statistical weights according to the Boltzmann distribution as shown in Figure 2.2.

From these states and weights we derive an equation describing the probability of finding the system in a transcriptionally active state, and therefore the production term from Equation 2.1,

$$\sum_i r_i p_i = \sum_i r_i \frac{W_i}{Z_{tot}}, \tag{2.2}$$

where W_i is the statistical weight of states in which the polymerase is bound, which are assumed to lead to the transcription of the operon (shaded blue in Figure 2.2), and $Z_{tot} = \sum_{\text{All states}} W_{state}$ is the partition function, or the sum of the statistical weights of all states. We connect this model to experimental measurements of repression, that is the ratio of gene expression in the absence of the active repressor to gene expression in the presence of active repressor, using:

$$\text{repression} = \frac{\text{gene expression}\,(R = 0)}{\text{gene expression}\,(R \neq 0)}, \tag{2.3}$$

where R is the number of repressor molecules per cell. The experimental equivalent of repression is depicted in Figure 2.1(c). In experiments, isopropyl β-D-1-thiogalactopyranoside (IPTG) is used to inactivate the Lac repressor, preventing it from binding to the genome with high affinity [19].

Repression, as defined in Equation 2.3, has been a standard metric for the role of transcription factors, including the Lac repressor, on gene expression [7, 42]. By measuring the ratio of steady-state levels of a gene reporter protein, here LacZ, we are able to isolate the role of the repressor in gene regulation, as described further in Section 2.5.11.

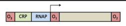
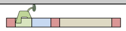
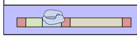
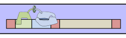
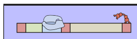
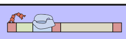
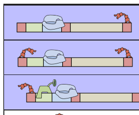
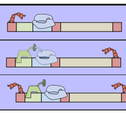
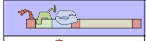
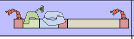
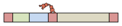
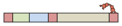
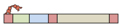
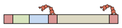
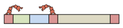
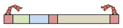
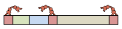
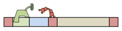
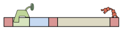
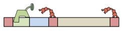
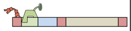
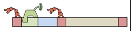
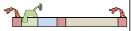
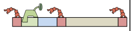
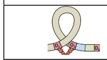
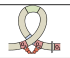
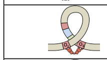
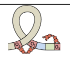
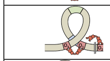
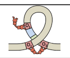
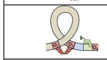
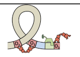
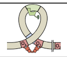
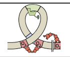
State	Weight	State	Weight
	1		$\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_a}$
	$\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_p}$		$\frac{(A)(P)}{N_{NS}^2}e^{-\beta(\Delta\varepsilon_a+\Delta\varepsilon_p+\Delta\varepsilon_{ap})}$
	$\frac{2R(P)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O2}+\Delta\varepsilon_p\right)}$		$\frac{2R(P)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O3}+\Delta\varepsilon_p\right)}$
	$\frac{4R(R-1)(P)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta\varepsilon_p\right)}$		$\frac{2R(A)(P)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_p+\Delta\varepsilon_{ap}\right)}$
	$\frac{2R(A)(P)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_p+\Delta\varepsilon_r^{O3}\right)}$		$\frac{4R(R-1)(A)(P)}{N_{NS}^4}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_p+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}\right)}$
	$\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_r^{O1}}$		$\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_r^{O2}}$
	$\frac{2R}{N_{NS}}e^{-\beta\Delta\varepsilon_r^{O3}}$		$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}\right)}$
	$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O3}\right)}$		$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}\right)}$
	$\frac{8R(R-1)(R-2)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}\right)}$		$\frac{2R(A)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O1}\right)}$
	$\frac{2R(A)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O2}\right)}$		$\frac{4R(R-1)(A)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}\right)}$
	$\frac{2R(A)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O3}\right)}$		$\frac{4R(R-1)(A)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O3}\right)}$
	$\frac{4R(R-1)(A)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}\right)}$		$\frac{8R(R-1)(R-2)(A)}{N_{NS}^4}e^{-\beta\left(\Delta\varepsilon_a+\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}\right)}$
	$\frac{2R}{N_{NS}}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta F_{loop}(l_{12})\right)}$		$\frac{2R}{N_{NS}}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O3}+\Delta F_{loop}(l_{13})\right)}$
	$\frac{2R}{N_{NS}}e^{-\beta\left(\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta F_{loop}(l_{23})\right)}$		$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta F_{loop}(l_{12})\right)}$
	$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta F_{loop}(l_{13})\right)}$		$\frac{4R(R-1)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta F_{loop}(l_{23})\right)}$
	$\frac{2R(A)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_a+\Delta F_{loop}(l_{12})\right)}$		$\frac{4R(R-1)(A)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta\varepsilon_a+\Delta F_{loop}(l_{12})\right)}$
	$\frac{2R(A)}{N_{NS}^2}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O3}+\Delta\varepsilon_a+\Delta\varepsilon_{ar}+\Delta F_{loop}(l_{13})\right)}$		$\frac{4R(R-1)(A)}{N_{NS}^3}e^{-\beta\left(\Delta\varepsilon_r^{O1}+\Delta\varepsilon_r^{O2}+\Delta\varepsilon_r^{O3}+\Delta\varepsilon_a+\Delta\varepsilon_{ar}+\Delta F_{loop}(l_{13})\right)}$

Figure 2.2: Thermodynamic model of gene regulation. The table shows all states permitted within the model and their respective statistical weights as obtained using statistical mechanics. In these weights P = number of RNAP per cell, R = number of repressor molecules per cell, A = number of activator molecules per cell, $\Delta\varepsilon_r^{O_i}$ = binding energy of Lac repressor to the i^{th} operator, $\Delta\varepsilon_p$ = binding energy of RNA polymerase to the promoter, $\Delta\varepsilon_a$ = activator binding energy, $\Delta F_{loop}(l_{ij})$ = looping free energy between operator O_i and O_j, N_{NS} = number of nonspecific binding sites on the genome, $\Delta\varepsilon_{ap}$ = interaction energy between the activator and the RNAP, $\Delta\varepsilon_{ar}$ = interaction energy between the activator and the repressor, and β = inverse of the Boltzmann constant times the temperature (see Supplementary Information for further discussion). States with blue background are assumed to lead to transcription of the operon.

Various models of the wild-type *lac* promoter have been reported in the past using this simple structure. Our work builds upon the work by Kinney *et al.* [12]. Kinney and collaborators combined a thermodynamic model of regulation with high-throughput sequencing to predict gene expression from statistical sequence information of the cAMP-receptor protein (CRP) and the RNAP binding sites. To predict how the sequence of the entire regulatory region influences expression, we adapted this model to account for how the binding site sequence and copy number of the Lac repressor modulate gene expression. Our model also takes into account growth rate effects, captured in the RNAP copy number [43, 44].

Based on previous work done on the *lac* operon [12, 19], we assumed that the presence of the activator does not affect the rate of transcription (r_i from Equation 2.1), but instead influences the probability of recruiting the polymerase to the promoter (p_i from Equation 2.1). Previous experimental characterization of the repressor binding energy to the different operators [26], the looping free energy for the upstream loop between $O_1 - O_3$ [27], activator concentration and its interaction energy with RNAP [19], RNAP binding energy [15] and RNAP copy number as a function of the growth rate [44], left us only with three unknown parameters for the model. One of these missing parameters, a decrease in the looping free energy when CRP and Lac repressor are bound at the same time, is a consequence of the experimental observation that the presence of CRP stabilizes the formation of the loop between $O_1 - O_3$ [29, 45]. The remaining two parameters, the looping energies for the $O_1 - O_2$ and $O_3 - O_2$ loops, are not well characterized. These looping energies may differ from upstream loops due to the absence of the RNAP binding site which modifies the mechanical properties of the loop [46]. We fit these parameters for our model using Oehler *et al.* repression measurements on *lac* operon constructs with partially mutagenized or swapped binding sites [42, 47] (see section S5 of the Supplementary Information for further details). Using these parameters the model is consistent with previous measurements (Figure 2.12). We emphasize that having the 14 parameters of the model characterized (see Table 2.2) provides testable predictions without free parameters that we compare with our experimental results.

2.2.2 Sensitivity of expression to model parameters

As an exploratory tool, the model can predict the change in regulation due to modifications in the promoter architecture. Figure 2.3 shows the fold-change in the repression level as a function of each of the parameters, using the lab strain MG1655 as a reference state (see Supplementary Information for further detail on these reference parameters). We have reported parameters using strain MG1655 as a reference strain because this strain served as the basis for which most parameter values were determined and the gene expression model was derived.

From this figure we see that within the confines of this model, modifications in the O_1 binding energy have the most drastic effect on the repression of the operon. For the case of O_2 we see that increasing its affinity for the repressor does not translate into an increased ability to turn off the operon; but by decreasing this operator affinity the model predicts a reduction in the repression with respect to the reference strain.

Surprisingly the repression level is predicted to be insensitive to activator copy number. The same cannot be said about the affinity of the activator, since decreasing the activator binding energy greatly influences the repression level.

2.2.3 Mapping from sequence space to level of regulation

Recent developments of an experimental technique called Sort-Seq, involving cell sorting and high-throughput sequencing, have proved to be very successful in revealing how regulatory information is encoded in the genome with base pair resolution [12]. This technique generates energy matri-

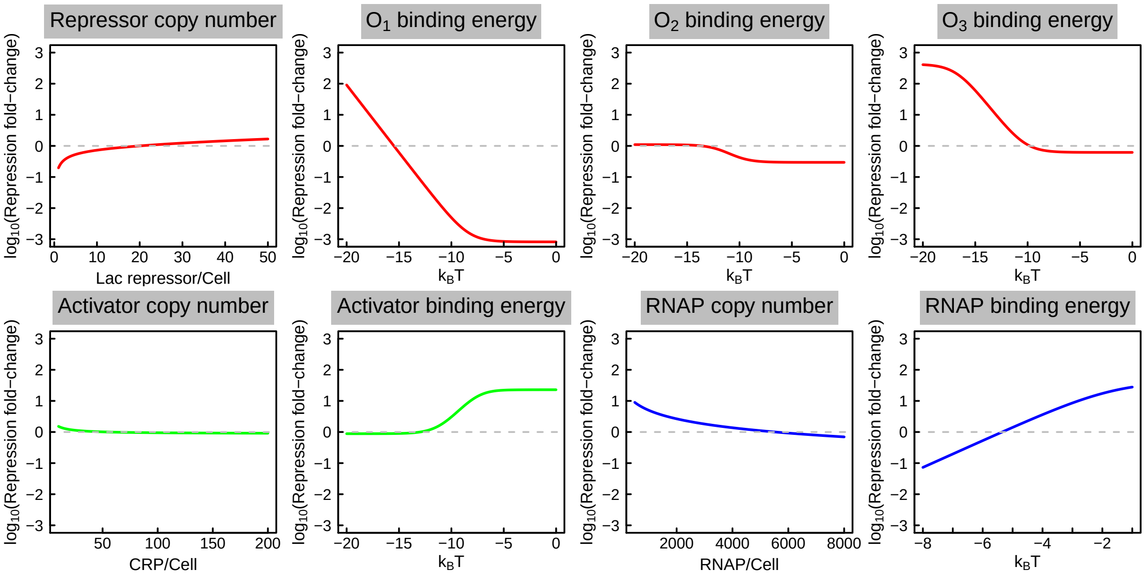


Figure 2.3: Sensitivity of phenotype to the parameters controlling the gene expression level. Each graph shows how a specific model parameter changes the level of gene expression. The $\log_{10}$ ratio of repression is calculated with respect to the predicted repression for the lab strain MG1655. The vertical axis spans between 1000 fold decrease to 1000 fold increase in repression with respect to this strain. The gray dotted line indicates the reference value for the lab strain MG1655. Values above this line indicate the operon is more tightly repressed and values below this line have a leakier expression profile (see Table 2.2 for further detail on the reference parameters).

ces that make it possible to map from a given binding site sequence to its corresponding binding energy for a collection of different proteins and binding sites. Combining these energy matrices with thermodynamic models enables us to convert promoter sequence to the output level of gene expression. Recently these energy matrices have been used to deliberately design promoters with a desired expression level, demonstrating the validity of these matrices as a design tool for synthetic constructs [15]. We use the matrices for CRP and RNAP published previously [12]. We experimentally determined the matrix for the LacI operator using previously published methods [12], as discussed in Materials and Methods. Figure 2.4(a) shows a schematic representation of the relevant protein binding sites involved in the regulation of the *lac* operon and their respective energy matrices. Implementing these matrices into the thermodynamic model gives us a map from genotype to phenotype. We use this map to calculate the fold-change in repression relative to MG1655 for all possible point mutations in this region. Figure 2.4(b) shows the fold-changes in repression levels for the two base pair substitutions at each position that result in the largest predicted increase or decrease in repression.

Again we see that mutations in the O_1 binding site have the largest effect on regulation since a single base pair change can lower the ability of the cell to repress the operon by a factor of ≈ 20. With only two relevant mutations that could significantly increase the repression level, this map reveals how this operator and its corresponding transcription factor diverged in a coordinated fashion; the wild-type sequence has nearly maximum affinity for the repressor [48]. It is known that

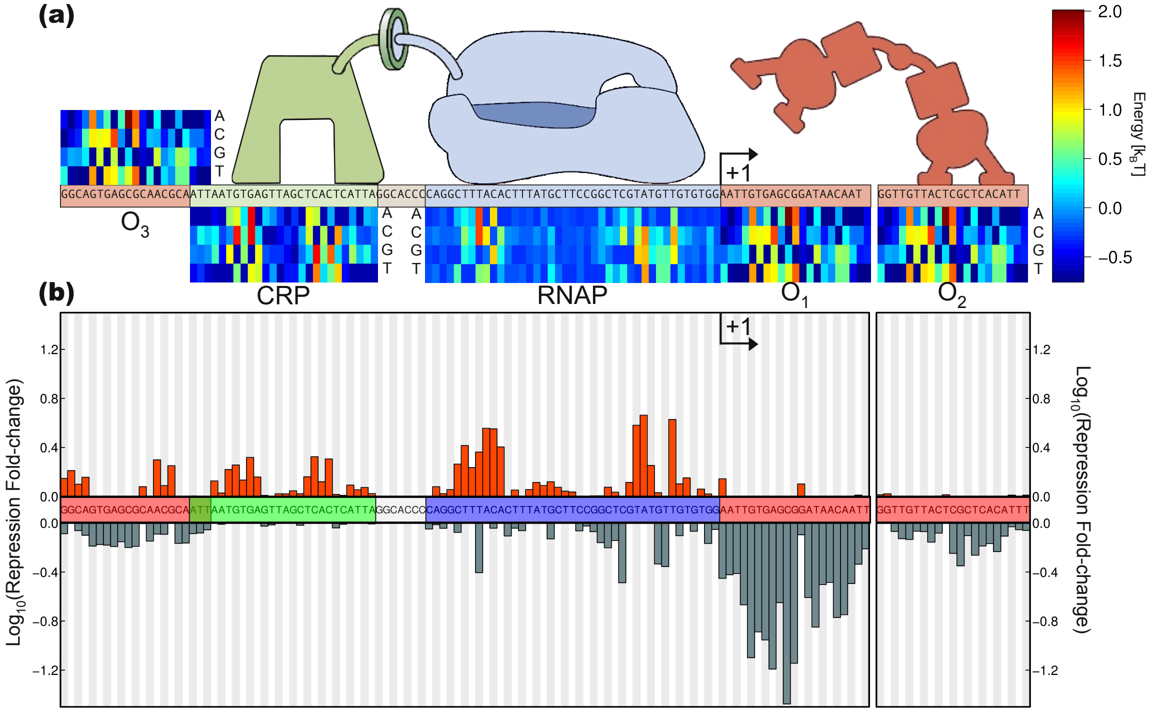


Figure 2.4: Mapping from promoter sequence to regulatory level. (a) Energy matrices for the relevant transcription factors (Blue - RNAP, green - CRP, red - Lac repressor). These matrices allow us to map from sequence space to the corresponding binding energy. The contribution of each base pair to the total binding energy is color coded. The total binding energy for a given sequence is obtained by adding together the contribution of each individual base pair. (b) Using the energy matrices from (a) and the model whose states are depicted in Figure 2.2, the $\log_{10}$ repression change was calculated for all possible single point mutations of the promoter region. The height of the bars represents the biggest possible changes in the repression level (gray bars for biggest predicted decrease in repression, orange bar for biggest predicted increase in repression) given that the corresponding base pair is mutated with respect to the reference sequence (*lac* promoter region of the lab strain MG1655). The black arrows indicate the transcription start site.

the non-natural operator O_{id} binds more strongly than O_1 [42]. O_{id} is one base pair shorter than O_1 and current maps made with Sort-Seq cannot predict changes in binding affinity for binding sites of differing length, although accounting for length differences in binding sites is not a fundamental limitation of this method.

For the auxiliary binding sites, the effect discussed in section 2.2.2 is reflected in this map: increasing the Lac repressor affinity for the O_2 binding site does not increase repression. Mutations in almost all positions can decrease repression, and no base pair substitutions significantly increase the repression level. Mutations in the O_3 binding site have the potential to either increase or decrease the repression level. With respect to the RNAP binding site, we can see that, as expected, the most influential base pairs surround the well characterized -35 and -10 boxes. The CRP binding site overlaps three base pairs with the upstream Lac repressor auxiliary operator. As the heat-map reveals, the binding energy is relatively insensitive to changes in those base pairs, so we assume independence when calculating the binding energy and capture the synergy between the Lac repressor bound to O_3 and CRP with an interaction energy term.

The construction of the sequence to phenotype map enables us to predict the evolvability of the *lac* promoter region. We calculated the effect that all possible double mutations would have in the regulation of the operon, again with respect to the predicted repression level of the reference strain MG1655. Figure 2.5 shows what we call the "phenotype change distribution" obtained by mutating one or two base pairs from the reference sequence, under the assumption of same growth rate and transcription factor copy numbers as the reference strain. The distribution peaks at zero for both cases, meaning that the majority of mutations are predicted not to change the repression level with respect to the reference strain, and would result in genetic drift. However it is interesting to note that the range of repression values predicted by the model with only one mutation varied between 30 times lower and 4.6 times higher than the reference value, and with two mutations the repression varied between 345 times lower and 15 times higher than the reference value. This suggests that regulation of this operon could rapidly adapt and fine tune regulation given appropriate selection.

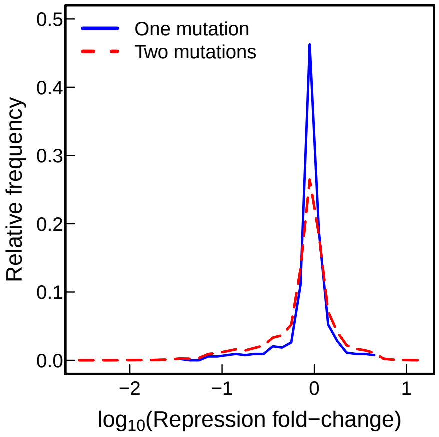


Figure 2.5: Phenotype change distribution. Relative frequency of the predicted changes in repression level by mutating one (solid blue line) or two (dashed red line) base pairs from the reference sequence (MG1655 promoter region).

2.2.4 Promoter sequence variability of natural isolates and available sequenced genomes

In order to explore the natural variability of this regulatory circuit, we analyzed the *lac* promoter region of 22 wild-type *E. coli* strains which were isolated from different organisms [30], along with 69 fully sequenced *E. coli* strains (including MG1655) available online (`http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html`). Figure 2.6 summarizes the sequencing results; for comparison, we plot the "genotype to phenotype map" from Figure 2.4(b) to gain insight into how the sequence variability influences regulation in these strains. Figure 2.6(b) shows the relative frequency of single nucleotide polymorphisms (SNPs) with respect to the consensus sequence. Qualitatively we can appreciate that the mutations found in these strains fell mostly within base pairs which, according to the model, weakly regulated expression. To quantify this observation we mapped the sequences to their corresponding binding energies. As shown in Figure 2.6(c) the distribution of parameters is such that the observed mutations result in relatively small changes to the binding energies, less than 1 k_BT relative to the reference sequence, except for the O_3 binding energy that is predicted to increase >1 k_BT in 16 strains.

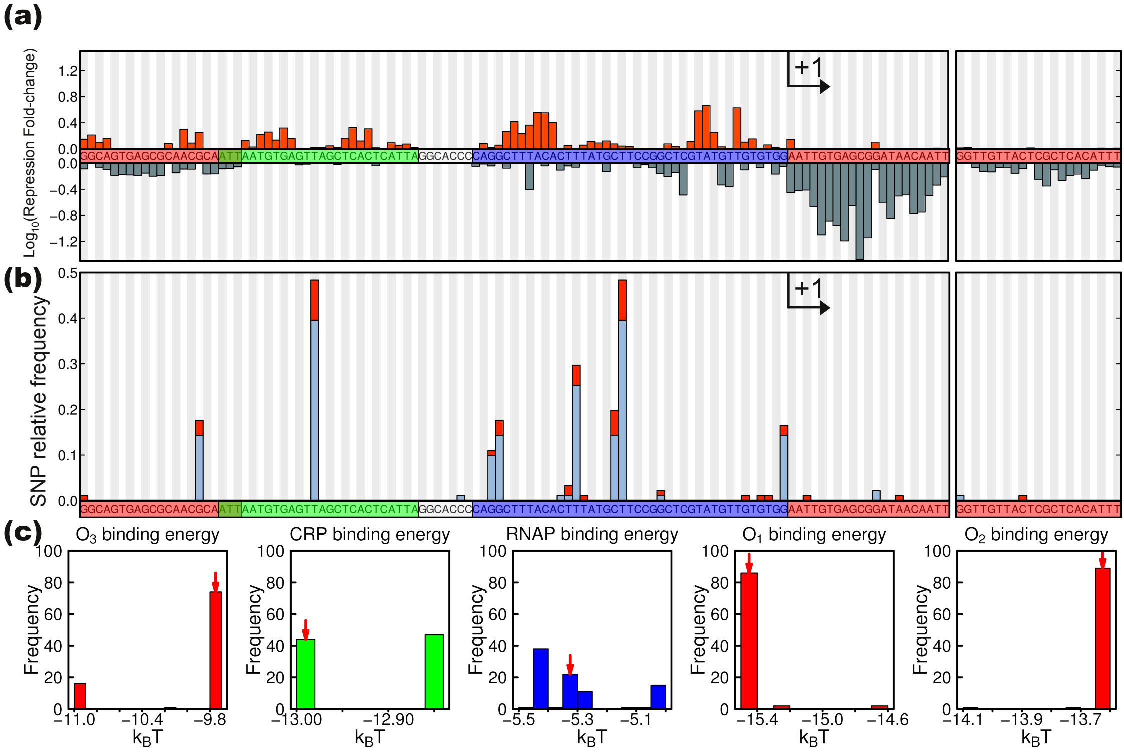


Figure 2.6: Mutational landscape of the regulatory region of the *lac* operon. (a) The genotype to phenotype map is reproduced from Figure 2.4(b) in order to show how each base pair in the region influences gene regulation. (b) Comparing the sequence of the *lac* promoter from 91 *E. coli* strains identifies which base pairs were mutated in this region. The heights of the bars represent the relative frequency of a mutation with respect to the consensus sequence. The red part of each bar represents the 22 natural isolates from different hosts [30] and the light blue part of these bars represents the 69 fully sequenced genomes (`http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html`). Color coding of the binding sites and the transcription start site is as in Figure 2.4. (c) Using the energy matrices of Figure 2.4(a), we calculate the variability of protein binding energies for all sequences. The red arrow indicates reference binding energies for control strain MG1655.

2.2.5 Does the model account for variability in the natural isolates?

Next we further characterized the eight strains from Figure 2.1(b) in order to determine if the observed variability in regulation could be accounted for in the model (see Section 2.5.2 for details on the 16S rRNA of this subset of strains). In particular, we measured the *in vivo* repressor copy number with quantitative immunoblots (see Material and Methods) and the growth rate. Table 2.1 shows the measured repressor copy number and the doubling time for these strains.

Using the thermodynamic model by taking into account the repressor copy number, the promoter

Strain	**Repressor/cell**	**Doubling time [min]**
Lab strain	21± 4	29.1± 0.2
Bat	12± 1	27.5± 0.2
Human-MA	20± 4	35.6± 0.6
Human-NY	23± 4	41.5± 0.4
Human-Sweden	28± 1	34.2± 0.3
Jaguar	21± 3	32.0± 0.2
Opossum	26± 2	33.5± 0.2
Perching bird	24± 4	30.2± 0.3

Table 2.1: Lac repressor copy number as measured with the immunodot blots and doubling time of the eight strains with measured repression level shown in Figure 2.1(b). The errors represent the standard error of three independent experiments.

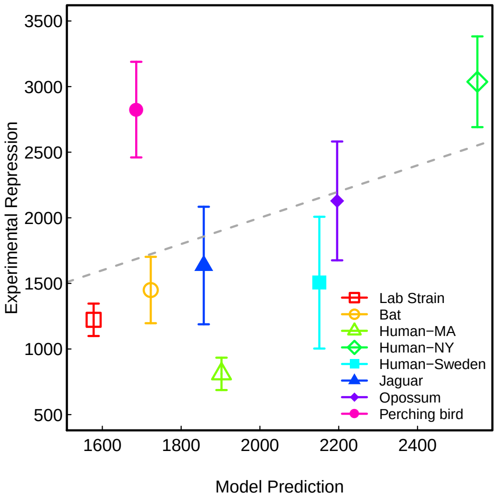


Figure 2.7: Comparison of model predictions with experimental measurements. Error bars represent the standard deviation of at least three independent measurements each with three replicates. The dotted line plots $x = y$.

sequence and the growth rate, we predict the repression level for each of the isolates measured in Figure 2.1(b). In Figure 2.7 we plot these predicted values vs. the experimental measurements. We find that the model accounts for the overall trends observed in the isolates, with the predictions for six of eight strains falling within two standard deviations of the measurements. A few of the measured repression values fall outside of the prediction, suggesting that the model may not capture the full set of control parameters operating in all of the strains.

2.2.6 Exploring the variability among different species

We extended our analysis to different microbial species with similar *lac* promoter architectures. After identifying bacterial species containing the *lac* repressor, we used the Sort-Seq derived energy matrices shown in Figure 2.4(a) to identify the positions of the transcription factor binding sites in each of these candidate strains. We identified a set of eight species whose *lac* promoter architecture was similar to *E. coli*. Figure 2.8 shows the 16S rRNA phylogenetic tree for these strains. The

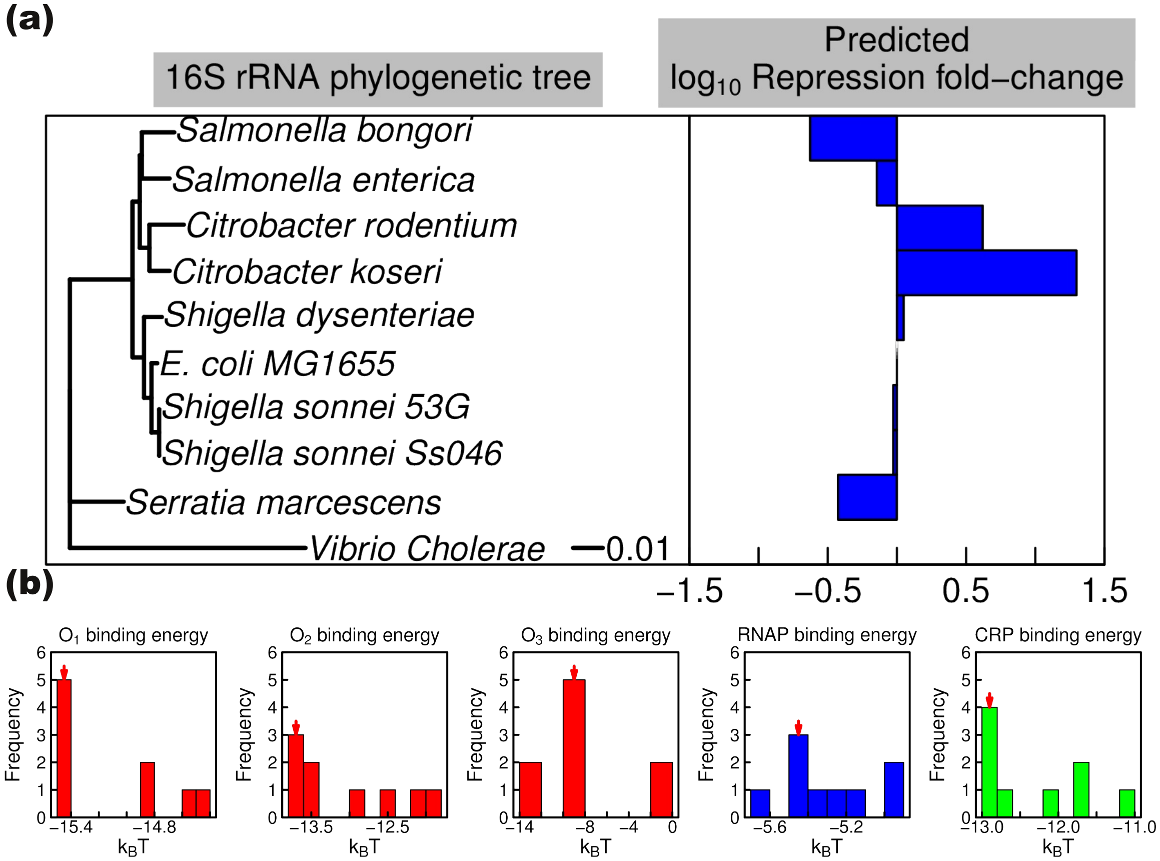


Figure 2.8: Predicted variability among different microbial species based on genome sequences and our model for regulation derived for *E. coli*. (a) On the left a 16S rRNA phylogenetic tree of diverse species with a similar *lac* promoter architecture done with the Neighbor-Joining algorithm. *Vibrio cholerae* was used as an outgroup species. The scale bar represents the relative number of substitutions per sequence. On the right the predicted $\log_{10}$ fold-change in repression with respect to *E. coli* MG1655 assuming the same growth rate and transcription factor copy numbers. The outgroup species fold-change was not calculated. (b) Parameter distribution calculated using the promoter region sequence and the energy matrices. The red arrow indicates the MG1655 reference value. Strains lacking a binding site were binned as zero.

predicted change in regulation was calculated for these strains using the model whose states are shown in Figure 2.2, the energy matrices in Figure 2.4(a), and assuming all strains have the same growth rate and transcription factor copy numbers as the lab strain MG1655. The repression level relative to *E. coli* among these species is predicted to increase as much as a factor of ≈ 20 and decrease as much as a factor of ≈ 4. Regulation of the operon seems to follow phylogenetic patterns in the 16S rRNA tree, with *E. coli* relatives having a similar predicted repression level, *Citrobacter* evolved to increase repression, and *Salmonella* evolved to decrease repression.

2.3 Discussion

The approach presented here combines thermodynamic models of gene regulation with energy matrices generated with Sort-Seq to produce a single-base pair resolution picture of the role that each position of the promoter region has in regulation. These types of models based on equilibrium statistical mechanics have been used previously for the *lac* operon [19, 25], here we expanded the model to account for important cellular parameters such as growth rate, the binding site strengths of all transcription factors, and the binding site strength of RNAP. Thermodynamic models are functions of the natural variables of the system as opposed to the widely used phenomenological Hill functions [49], where it is less straightforward to judge how changes to a promoter region translate to changes in regulatory parameters such as K_M, the half saturation constant, and n, the Hill coefficient. Currently our model assumes that protein-protein interactions and DNA looping energies are kept constant, but these variables could also be a function of the promoter sequence, affecting the positioning of the transcription factors and therefore their interactions with the other molecules involved.

The underlying framework developed here can be applied to any type of architecture. Here we use the *lac* operon because it is well characterized. There is no reason to believe that this approach could not be extended to other regulatory regions, however such an effort would require extensive quantitative characterization of the control parameters of each genetic circuit, such as protein copy numbers, interaction energies, and binding affinities. Although this level of characterization requires additional experimental effort, we believe that developing such predictive, single-base pair models of gene regulation can lead to significant insights into how genetic circuits function, interact with each other, and evolve.

The majority of the natural variability found among the sequenced promoters tended to fall in bases predicted to have low impact on overall regulation, as shown in Figure 2.6. As an example the highly conserved mutation in the CRP binding energy or the mutations along the RNAP binding site are predicted to change the binding energy by less than 1 k_BT, having a very low impact on the repression level. With respect to the repressor binding sites, among the sequenced natural isolates only one mutation was found in the O_2 binding site. Unlike the O_1 and O_3 operators, the evolution of O_2 may be constrained given that its sequence encodes both gene regulatory information and is part of the coding region of the β-galactosidase gene.

As shown in Figure 2.7, after taking into account the variability in the promoter sequence, changes in the repressor copy number, and changes in the growth rate, the model accounts for most of the variability in regulation for the majority of the isolates. Linear regression of the entire experimental dataset weighted by the inverse of their standard deviation gives a slope of 1.26 with an R^2 of 0.24. It can be seen that many of the points fall close to or on the x=y line, indicating that the poor fit is a result of a few outliers within the dataset. Removing the outliers (Perching bird, Human-MA, and Human-NY) results in a best fit line of slope 1.05 with R^2 0.74, reiterating that the model is consistent with the phenotype of five of eight isolates. It is interesting that the three isolates whose regulatory outputs were predicted poorly by the model (Perching bird, Human-MA, and Human-NY in Figure 2.7) all have identical promoter sequences, which is the consensus promoter sequence as shown in Figure 2.9. Although these three strains have identical sequences, two strains repressed more than predicted and the other strain repressed less. This indicates there are likely other cellular parameters that influence gene expression levels that are not included in the model. Currently the model cannot take into account variation in the protein structure of the transcription factors or the RNAP and its sigma factors. Changes in these proteins could account for some of the discrepancies between the model and the observed levels of regulation. It is likely that some global parameters that modulate transcriptional outputs which are not accounted for in the model also contribute to the disagreement with model predictions. We note that repression is a measurement of expression relative to expression in the absence of the repressor. This definition enables us to isolate the role of a particular transcription factor in regulation. Therefore, as discussed in Section 2.5.11, some global regulatory parameters such as ribosomal binding sites of the relevant genes and variables such as the ribosome copy number should not impact repression levels.

From an evolutionary perspective, it is interesting that the regulation seems to be more sensitive to changes in the activator binding energy than to the activator protein copy number, as shown in Figure 2.3. This result might be attributed to the nature of this transcription factor. CRP is known to be a "global" transcription factor that regulates >50% of the *E. coli* transcription units [50]. Given its important global role in the structure of the transcriptome, changing the copy number of CRP would have a global impact on expression whereas tuning its binding affinity at a particular regulatory region has a local impact on one promoter. The regulatory knob of CRP copy number not influencing expression at the *lac* operon indicates this regulatory region may have evolved to be robust against changes in this global regulatory parameter.

The fact that the O_3 operator has the possibility to change in both directions (greater or lower affinity) as reflected in Figure 2.4(b) suggests plasticity of the operon, allowing it to evolve according to environmental conditions. In fact this parameter changed the most among the related microbial species as shown in Figure 2.8(b), having species such as *Citrobacter koseri* with an operator predicted to be 5 k_BT stronger than the reference value, and other species such as *Salmonella bongori* that completely lost this binding site. Although we do not yet know whether these regulatory predictions will be borne out in experimental measurements, this analysis demonstrates the utility of our sequence-to-phenotype map in interpreting the consequences of variability within the regulatory regions of sequenced genomes.

To the best of our knowledge Figure 2.5 shows the first quantification of how easily regulation can change given one or two point mutations along the entire promoter region. Previous studies were limited to a subset of base pairs in the Lac repressor operators and two amino acid substitutions in the Lac repressor [51]. The distribution of predicted phenotypes is very sharp close to the reference value, and as a consequence the majority of the possible mutations would not be selected on. But given that regulation can change by an order of magnitude or more in both directions (increased or decreased repression) with only two mutations, changing the regulatory region of the gene could function as a fast response strategy of adaptation.

It is known from previous work that *lac* operon expression can have an impact on cell fitness [20–22, 24]. Under laboratory conditions, high expression of the *lac* operon resulted in loss of fitness due to expression of *lacY*, a transporter which imports lactose into the cell. This would suggest regulation is essential to avoid the negative consequences of *lacY* overexpression, and tight regulation would be selected. However it is possible that natural selection would act also to modulate the magnitude of the response. Strains exposed to environments with periodical bursts of lactose could trigger instantly a high gene dosage, resulting in a steeper slope on an induction curve, while strains rarely exposed to lactose would have a moderate response, i.e. a less steep induction curve. Our exploration and prediction of regulatory phenotypes in sequenced genomes shows that the biggest changes in regulation were found to increase repression (Figure 2.6(c)), suggesting that lactose might not be present regularly in the natural environment of some strains.

The combination of thermodynamic models with Sort-Seq-generated energy matrices presented here promises to be an useful tool with which to study the evolution of gene regulation. This theoretical framework allows us to explore the effect that the modification of control parameters can have on the expression levels, and to predict how point mutations in gene promoter regions enable cells to evolve their gene regulatory circuits.

2.4 Materials and methods

2.4.1 Growth conditions

Unless otherwise indicated, all experiments were started by inoculating the strains from frozen stocks kept at -80°C. Cultures were grown overnight in Luria Broth (EMD, Gibbstown, NJ) at 37°C with shaking at 250 rpm. In all of the experiments these cultures were used to inoculate three replicates for each of the relevant conditions, diluting them 1:3000 into 3 mL of M9 buffer (2 mM $MgSO_4$, 0.10 mM $CaCl_2$, 48 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.6 mM $NaCl$, 19 mM NH_4Cl) with 0.5% glucose and 0.2% casamino acids (here referred to as "supplemented M9"). Cells were cultured at 37°C with shaking at 250 rpm and harvested at the indicated OD_{600}.

2.4.2 Gene expression measurements

To perform the LacZ assay we followed the protocol used by Garcia and Phillips [26]. Strains were grown in supplemented M9 for approximately 10 generations and harvested at an OD_{600} around 0.4. A volume of the cells was added to Z-buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) for a total volume of 1 mL. For fully induced cells we used 50 μL and for uninduced cultures we concentrated the cells by spinning down 1 mL of culture and resuspending in Z-buffer. The cells were lysed by adding 25 μL of 0.1% SDS and 50 μL of chloroform and vortexing for 15 seconds. To obtain the readout, we added 200 μL of 4 mg/mL 2-nitrophenyl β-D-galactopiranoside (ONPG). Once the solution became noticeably yellow, we stopped the reaction by adding 200 μL of 2.5 M Na_2CO_3.

To remove cell debris we spun down the tubes at 13000 $\times$ g for 3 minutes. 200 μL of the supernatant were read at OD_{420} and OD_{550} on a microplate reader (Tecan Safire2). The absolute activity of LacZ was measured in Miller units as

$$MU = 1000 \times \frac{OD_{420} - 1.75 \times OD_{550}}{t \times v \times OD_{600}} \times 0.826, \tag{2.4}$$

where t is the time for which we let the reaction run and v is the volume of cells used in mL. The factor of 0.826 adjusts for the concentration of ONPG relative to the standard LacZ assay.

2.4.3 Measuring *in-vivo lac* repressor copy number

To measure the repressor copy number of the natural isolates we followed the same procedure reported by Garcia and Phillips [26]. Strains were grown in 3 mL of supplemented M9 until they reached an $OD_{600} \approx 0.4 - 0.6$. Then they were transferred into 47 mL of warm media and grown at 37°C to an OD_{600} of 0.4-0.6. 45 mL of culture were spun down at 6000×g and resuspended into 900 μL of breaking buffer (0.2 M Tris-HCl, 0.2 M KCl, 0.01 M Magnesium acetate, 5% glucose, 0.3 mM DTT, 50 mg/100 mL lysozyme, 50μg/L phenylmethanesulfonylfluoride (PMSF), pH 7.6).

Cells were lysed by performing four freeze-thaw cycles, adding 4 μL of a 2,000 Kunitz/mL DNase solution and 40 μL of a 1 M $MgCl_2$ solution and incubating at 4°C with mixing for 4 hours after the first cycle. After the final cycle, cells were spun down at 13,000×g for 45 min at 4°C. We then obtained the supernatant and measured its volume. The pellet was resuspended in 900 μL of breaking buffer and again spun down at 15,000×g for 45 min at 4°C. In order to review the quality of the lysing process, 2 μL of this resuspended pellet was used as a control to ensure the luminescent signal of the resuspension was <30% of the sample.

To perform the immuno-blot we prewetted a nitrocellulose membrane (0.2 μM, Bio-Rad) in TBS buffer (20 mM $Tris - HCl$, 500 mM $NaCl$) and left it to air dry. For the standard curve a purified stock of Lac repressor tetramer [46] was serially diluted into HG105 ($\Delta lacI$ strain) lysate. 2 μL were spotted for each of the references and each of the samples. After the samples were visibly dry the membrane was blocked using TBST (20 mM Tris Base, 140 mM NaCl, 0.1% Tween 20, pH 7.6) +2% BSA +5% dry milk for 1 h at room temperature with mixing. We then incubated the membrane in a 1:1000 dilution of anti-$LacI$ monoclonal antibody (from mouse; Millipore) in blocking solution for

1.5 h at room temperature with mixing. The membrane was gently washed with TBS ≈ 5 times. To obtain the luminescent signal the membrane was incubated in a 1:2000 dilution of HRP-linked anti-mouse secondary antibody (GE Healthcare) for 1.5 h at room temperature with mixing and washed again ≈ five times with TBS. The membrane was dried and developed with Thermo Scientific Super-Signal West Femto Substrate and imaged in a Bio-Rad VersaDoc 3000 system.

2.4.4 Constructing the *in-vivo lac* repressor energy matrix

The energy matrix was inferred from Sort-Seq data in a manner analogous to methods described in Kinney PNAS 2010 [12]. Briefly, a library of mutant *lac* promoters was constructed in which the region [-100:25] (where coordinates are with respect to the transcription start site) was mutagenized with a 3% mutation rate. The transcriptional activity of each mutant promoter was measured by flow cytometry using a GFP reporter. To fit the LacI energy matrix, we used a Markov chain Monte Carlo algorithm to fit an energy matrix to the LacI O_1 binding site by maximizing the mutual information between energies predicted by the matrix and flow cytometry measurements. The justification for maximizing mutual information is described in detail in [12, 52].

2.5 Supplementary information

2.5.1 Alignment of promoter sequences

Figure 2.9 shows the alignment of the promoter regions of the *E. coli* wild isolates sequenced.

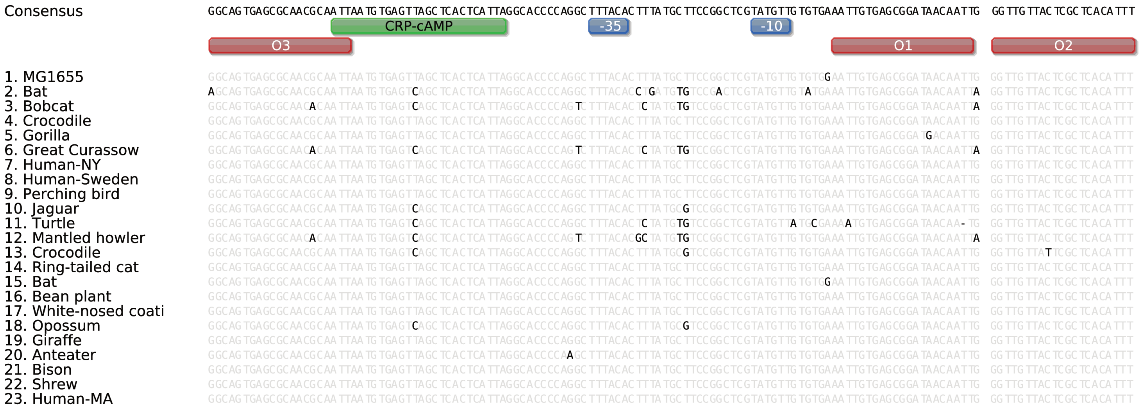


Figure 2.9: Promoter alignment of the sequenced strains. Highlighted bases differ from the consensus sequence on top. Colored boxes indicate the relevant binding sites for the Lac repressor (red), CRP (green) and RNAP (blue)

2.5.2 16S rRNA sequences

To confirm the identity of the strains we analyzed 490 bp of the 16S rRNA. Figure 2.10 shows a schematic representation of the sequences. Colored basepairs represent mutations with respect to the consensus sequence. All sequences were found to be ≥99% similar to the reference *E. coli MG1655* sequence.

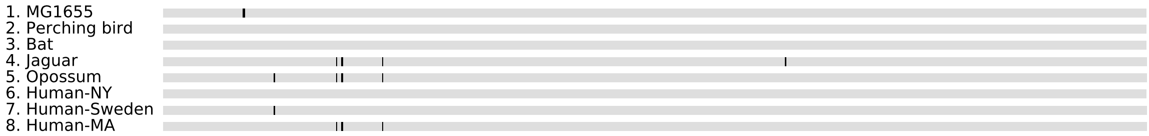


Figure 2.10: 16S sequence alignment. Black lines represent mutations with respect to the consensus sequence.

2.5.3 Model parameters

Table 2.2 shows the values of the reference parameters for MG1655 obtained from different sources.

Table 2.2: Reference parameters for the strain MG1655.

Parameter	Symbol	Value	Units	Reference
O_1 repressor operator binding energy	$\Delta\varepsilon_r^{O1}$	-15.3	k_BT	[26]
O_2 repressor operator binding energy	$\Delta\varepsilon_r^{O2}$	-13.9	k_BT	[26]
O_3 repressor operator binding energy	$\Delta\varepsilon_r^{O3}$	-9.7	k_BT	[26]
Repressor copy number	R	20	tetramer/cell	Measured
Activator binding energy	$\Delta\varepsilon_a$	-13	k_BT	[9, 19]
Number of active activators	A	55	active molecules/cell	[19]
RNAP binding energy for the *lac* promoter	$\Delta\varepsilon_p$	-5.35	k_BT	[15]
RNAP copy number	P	5500	active molecules/cell	[44]
Number of nonspecific binding sites	N_{NS}	4.6×10^6	-	GenBank: U00096.2
Looping free energy between $O_1 - O_2$	$\Delta F_{loop(l_{12})}$	4.7	k_BT	Fit to data from [42, 47]
Looping free energy between $O_1 - O_3$	$\Delta F_{loop(l_{13})}$	9	k_BT	[27]
Looping free energy between $O_2 - O_3$	$\Delta F_{loop(l_{23})}$	5.2	k_BT	Fit to data from [42, 47]
RNAP-CRP interaction energy	$\Delta\varepsilon_{ap}$	-5.3	k_BT	[12, 19]
Lac repressor - CRP interaction energy	$\Delta\varepsilon_{ar}$	-5.5	k_BT	Fit to data from [42, 47]

2.5.4 Derivation of the repression level equation

Thermodynamic models of gene regulation consider that the gene expression level is proportional to the probability of finding the RNAP bound to the promoter region [7–9, 11]. This biologically simplistic but powerful predictive tool allows us to study the effect of different transcription factors in different promoter architectures. In the case of the wild-type (WT) *lac* operon promoter architecture, where we have two different transcription factors involved in the regulation - the activator CRP and the Lac repressor.

The Lac repressor molecule, when bound to the main operator O_1, blocks the polymerase from binding to the promoter region, stopping the transcription of the operon. CRP plays a double role in the regulation of the operon, activating transcription by recruiting RNAP to the promoter region, and as several experiments have shown, enhancing repression by facilitating the formation of the upstream loop between the $O_1 - O_3$ operators [29, 45, 53]. Enhanced repression by CRP is due to pre-bending the DNA between 90° and 120° [54], thereby increasing the probability of looping by bringing the *lac* operators closer together. The model captures this effect by adding an interaction term $\Delta\varepsilon_{ar}$ in the states where CRP is bound and the Lac repressor forms a loop between operators O_1 and O_3.

Assuming quasi-equilibrium conditions for the relevant processes involved in transcription, we can use the Boltzmann distribution to compute the probability of finding the RNAP bound to the promoter region, obtaining

$$\begin{aligned}\mathrm{GE} \propto \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_p} \Bigg\{ 1 + \frac{2R}{N_{NS}} \left[e^{-\beta\Delta\varepsilon_r^{O2}} + e^{-\beta\Delta\varepsilon_r^{O3}} \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a} \right) \right] + \frac{4R\,(R-1)}{N_{NS}^2} e^{-\beta\left(\Delta\varepsilon_r^{O2} + \Delta\varepsilon_r^{O3}\right)} \\ \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a} \right) + \frac{A}{N_{NS}} e^{-\beta\left(\Delta\varepsilon_{a+\Delta\varepsilon_{ap}}\right)} \left(1 + \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_r^{O2}} \right) \Bigg\} / Z_{tot}, \quad (2.5)\end{aligned}$$

where GE stands for gene expression and Z_{tot} represents the partition function for the states shown in Figure 2.2 in the main text. The presence of CRP in the promoter region is not assumed to influence the kinetics of promoter escape, only the probability of RNAP binding. Tagami and Aiba [28] found that the role of CRP in the activation of the *lac* operon is restricted to the steps up to the formation of the open complex, in other words, the interaction between CRP and the RNAP are not essential for transcription after the formation of the open complex. In our model we capture this effect by including an interaction energy between CRP and the RNAP, $\Delta\varepsilon_{ap}$, that has been measured experimentally [12, 19].

In the activation mechanism proposed by Tagami and Aiba [28] CRP bends the DNA and RNAP recognizes the CRP-DNA bent complex. This model would imply that RNAP makes additional contacts with the upstream region of the promoter. Based on this model we assume that the presence of the Lac repressor bound on the O_3 operator and CRP bound on its binding site (without forming a DNA loop between $O_1 - O_3$) allows transcription to occur. Since the RNAP cannot contact the upstream region of the promoter because of the presence of the repressor, the interaction energy between CRP and RNAP is not taken into account in these states.

In order to quantify the influence of Lac repressor on expression levels, we measure repression, which is the fold change in gene expression as a result of the presence of the repressor. This metric has the benefit of normalizing to a strain with an identical genetic background, thus isolating the role of the repressor in regulation. This relative measurement is defined as

$$\text{repression} \equiv \frac{\text{gene expression}\,(R=0)}{\text{gene expression}\,(R \neq 0)}, \quad (2.6)$$

where R is the Lac repressor copy number. Computing this we obtain

$$\begin{aligned}
\text{repression} = &\left\{ \frac{\frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_p} \left[1 + \frac{A}{N_{NS}} e^{-\beta(\Delta\varepsilon_a + \Delta\varepsilon_{ap})}\right]}{1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a} + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_p} \left(1 + \frac{A}{N_{NS}} e^{-\beta(\Delta\varepsilon_a + \Delta\varepsilon_{ap})}\right)} \right\} \Big/ \\
&\left\{ \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_p} \left\{ 1 + \frac{2R}{N_{NS}} \left[e^{-\beta\Delta\varepsilon_r^{O2}} + e^{-\beta\Delta\varepsilon_r^{O3}} \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a}\right)\right]\right.\right. \\
&\left.+ \frac{4R\,(R-1)}{N_{NS}^2} e^{-\beta\left(\Delta\varepsilon_r^{O2} + \Delta\varepsilon_r^{O3}\right)} \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a}\right) + \frac{A}{N_{NS}} e^{-\beta\left(\Delta\varepsilon_{a+\Delta\varepsilon_{ap}}\right)} \left(1 + \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_r^{O2}}\right)\right\} \Big/ Z_{tot}.
\end{aligned} \tag{2.7}$$

This can be further simplified, resulting in

$$\begin{aligned}
\text{repression} = &\left\{ \frac{1 + \frac{A}{N_{NS}} e^{-\beta(\Delta\varepsilon_a + \Delta\varepsilon_{ap})}}{1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a} + \frac{P}{N_{NS}} e^{-\beta\Delta\varepsilon_p} \left(1 + \frac{A}{N_{NS}} e^{-\beta(\Delta\varepsilon_a + \Delta\varepsilon_{ap})}\right)} \right\} \Big/ \\
&\left\{ 1 + \frac{2R}{N_{NS}} \left[e^{-\beta\Delta\varepsilon_r^{O2}} + e^{-\beta\Delta\varepsilon_r^{O3}} \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a}\right)\right]\right. \\
&\left.+ \frac{4R\,(R-1)}{N_{NS}^2} e^{-\beta\left(\Delta\varepsilon_r^{O2} + \Delta\varepsilon_r^{O3}\right)} \left(1 + \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_a}\right) + \frac{A}{N_{NS}} e^{-\beta\left(\Delta\varepsilon_{a+\Delta\varepsilon_{ap}}\right)} \left(1 + \frac{2R}{N_{NS}} e^{-\beta\Delta\varepsilon_r^{O2}}\right)\right\} \Big/ Z_{tot},
\end{aligned} \tag{2.8}$$

the expression we use to predict the repression level of the natural isolates.

2.5.5 Estimating the number of active CRP molecules

The Catabolite Activator Protein, also known as cAMP-receptor protein (CRP) is a global transcriptional regulator in *E. coli* [50]. As it exists in two forms, the cAMP-CRP complex which is considered as the active state and the inactive state without cAMP bound, the number of active molecules is a function of the cAMP cellular concentration. From a thermodynamic perspective we can estimate this number as

$$[CRP - cAMP] = [CRP]\,\frac{[cAMP]}{K_{cAMP} + [cAMP]}, \tag{2.9}$$

where $[CRP - cAMP]$ is the concentration of active proteins, $[CRP]$ is the total concentration of this transcription factor, $[cAMP]$ is the cellular concentration of cAMP and K_{cAMP} is the *in vivo* dissociation constant of the cAMP-CRP complex.

Kuhlman et al. [19] reported the values for the CRP concentration ($[CRP] \approx 1500$ nM) and the dissociation constant ($K_{cAMP} = 10$ μM). Epstein et al. [55] measured the intracellular cAMP concentration in different media, including minimal media with glucose and casamino acids ($[cAMP] \approx$

0.38μM). Using these values we calculate the number of active CRP molecules as

$$A = 1500\left(\frac{0.38\mu M}{10\mu M + 0.38\mu M}\right) \approx 55\frac{molecules}{cell}, \tag{2.10}$$

where we used the rule of thumb that 1 nM$\approx 1\frac{molecule}{E.\ coli}$. This rule of thumb is enough for our predictions since the repression level is predicted to be largely insensitive to the activator copy number as shown in Figure 2.3 in the main text.

2.5.6 Estimating the number of available RNAP

In order to estimate the available number of RNAP molecules, we appeal to the work of Klumpp and Hwa [44] where they calculated the total number of RNAP molecules as well as the fraction of these molecules available for transcription as a function of the growth rate. Figure 2.11 shows the number of available RNAP as a function of the doubling cycles per hour.

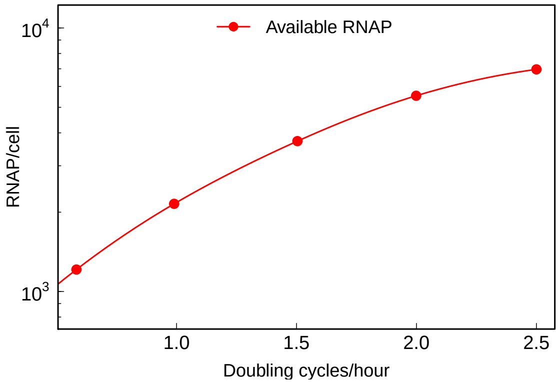


Figure 2.11: Adapted from Klumpp and Hwa [44]. RNAP available for transcription as a function of the number of doubling cycles per hour.

Using these results, we estimate 5500 $\frac{RNAP}{cell}$ for cells grown in 0.6% glucose + 0.2% casamino acids (with a doubling time of $\approx$ 30 min.). We interpolate between these data to obtain the RNAP copy number for each of the natural isolates.

2.5.7 Estimating CRP's binding energy

The activator binding energy was estimated as reported by Bintu et al. [9]. Using the reported dissociation constants from the specific binding site, K_{CRP}^{NS}, and nonspecific sequences, K_{CRP}^{S}, we can compute the binding energy as

$$\frac{\Delta\varepsilon_a}{k_BT} = \ln\left(\frac{K_{CRP}^{NS}}{K_{CRP}^{S}}\right). \tag{2.11}$$

Bintu et al. also reported the following values for both dissociation constants ($K_{CRP}^{NS} = 10^4$ nM and $K_{CRP}^{S} = 0.02$ nM), which gives us $\Delta\varepsilon_a \approx -13\ k_BT$.

2.5.8 Fitting parameters and testing the model

The three unknown parameters, the looping energies for the $O_1 - O_2$ and $O_3 - O_2$ loops and the decrease in the looping free energy when CRP and Lac repressor are bound at the same time, were inferred from the classic work of Oehler *et al.* [42, 47]. In these papers Oehler and collaborators measured the repression level of different *lac* operon constructs with either mutagenized or swapped Lac repressor binding sites while changing the repressor copy number. Because they reported the mutagenized sequences for the repressor binding sites we used the Sort-Seq derived energy matrix to calculate the residual energies of these modified binding sites. The three unknown parameters were fitted by minimizing the mean square error of the measurements,

$$f(\mathbf{x}^*) \leq f(\mathbf{x})\ \forall\, \mathbf{x} \in \mathbb{R}, \tag{2.12}$$

$$f(\mathbf{x}^*) = \left\{ \min \sum_{i=1}^{N} \frac{\left(Y_i\left(\mathbf{x}\right) - \bar{Y}_i\right)^2}{N} : \mathbf{x} \in \mathbb{R} \right\}, \tag{2.13}$$

where Y_i is the predicted value, $\bar{Y}_i$ is the experimental repression level for each of the constructs measured by Oehler *et al.* and $\mathbf{x}$ are the fitting parameters. Using this method we fit for the values of $\Delta F_{loop(l_{13})}$, $\Delta F_{loop(l_{23})}$, and $\Delta\varepsilon_{ar}$ using the data from references [42, 47]. The three parameter values are listed in Table 2.2.

2.5.9 Testing the model with different data

We used the model to predict the repression level of constructs reported by Oehler *et al.* [42, 47] and Müller *et al.* [56]. Figure 2.12 shows the comparison of the model predictions and the experimental results. The calculations were done using the model whose states are depicted in Figure 2.2, assuming a wild type repressor copy number of 10 repressors per cell, and calculating all the residual binding energies with the Lac repressor Sort-Seq derived energy matrix.

2.5.10 Error propagation

To calculate a confidence interval of the model, we used the *law of error propagation* [57] where we compute the contribution of the uncertainty in parameters to the uncertainty of the repression level as

$$\sigma_{\text{repression}} = \sqrt{\sum_i \left(\frac{\partial \text{repression}}{\partial x_i}\right)^2 \sigma_i^2}, \tag{2.14}$$

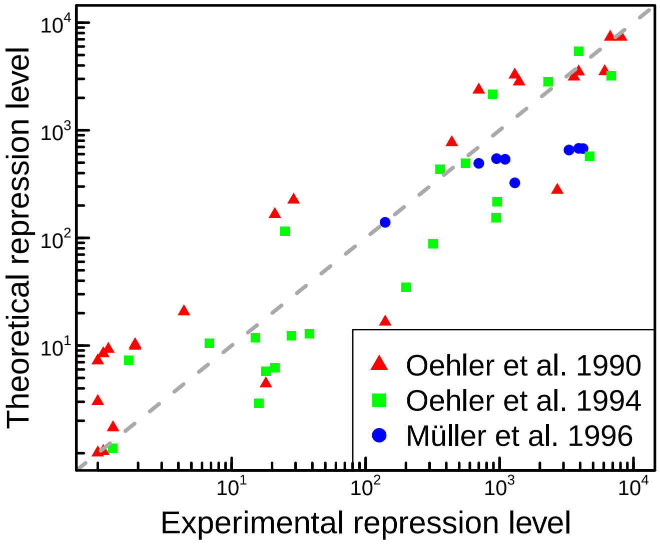


Figure 2.12: Comparing the experimental data from Oehler *et al.* [42, 47] and Müller *et al.* [56] with the model prediction.

where x_i represents each of the parameters of the model (binding energies, transcription factors copy number, looping energies, etc.) and σ_i represents the standard deviation of each of these parameters.

Paradoxically, calculating the contribution of each parameter to the uncertainty of the model requires "certainty" about the variability of these parameters. This means that we can only include the uncertainty of the parameters whose uncertainty measurements represent the natural variability in their values and not mostly error due to experimental methods. Table 2.3 lists the uncertainty of the parameters considered in this analysis given that the *in vivo* error was reported in the listed bibliography.

Table 2.3: Standard deviation of the parameters considered for the calculation of the confidence interval.

Parameter	Deviation	Units	Reference
R	Measured for each strain	LacI/cell	-
$\Delta\varepsilon_r^{O_1}$	±0.2	k_BT	[26]
$\Delta\varepsilon_r^{O_2}$	±0.2	k_BT	[26]
$\Delta\varepsilon_r^{O_3}$	±0.1	k_BT	[26]
$\Delta\varepsilon_a$	±1.1	k_BT	[19]

We used a customized *Mathematica* script (Wolfram Research, Champaign, IL) to calculate the partial derivatives. Figure 2.13 reproduces Figure 2.7 from the main text, including the predicted standard deviation.

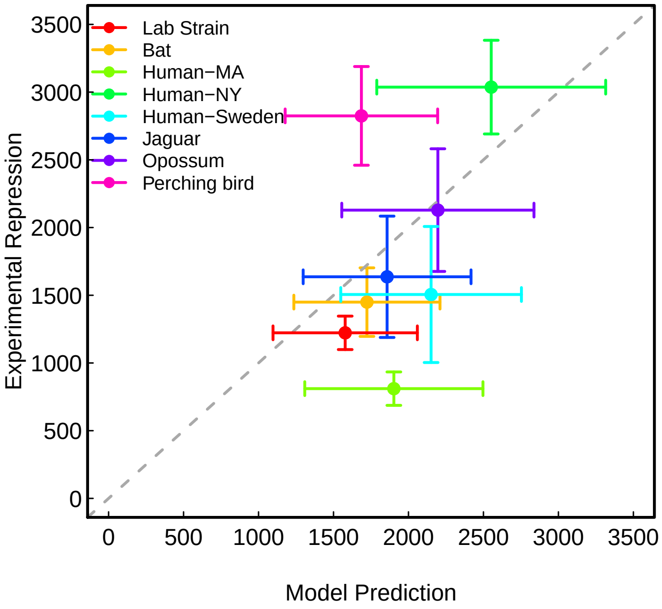


Figure 2.13: Comparison of the model prediction with the experimental measurement. Vertical error bars represent the standard deviation of at least three independent measurements each with three replicates. Horizontal error bars represent the 68% confidence interval of the model calculated by using the *law of error propagation* with the parameter uncertainties listed in Table 2.3.

2.5.11 Measuring repression level decouples growth rate effects in translation from effects in transcription

From previous work it was determined that one key regulatory parameter that is influenced by growth rate is the RNAP copy number [58]. However other cellular parameters such as ribosomal copy number and the dilution of mRNA concentration due to growth are also impacted. These parameters will influence protein copy number by influencing the efficiency of mRNA translation. In a very simple dynamical model of transcription, we can imagine that the change in the number of messenger RNA (mRNA) is proportional to the transcription rate and the degradation rate of the mRNA,

$$\frac{dmRNA}{dt} = k_t \cdot p_{bound} - \beta_{mRNA} \cdot mRNA, \tag{2.15}$$

where k_t is the maximum transcription rate when the operon is fully induced and p_{bound} is the probability of finding the RNAP bound to the relevant promoter, as derived using statistical mechanics; β_{mRNA} is the mRNA degradation rate and $mRNA$ is the number of transcripts of the gene per cell. This equation assumes that the most relevant effect for mRNA depletion is the degradation of the transcripts, compared with the dilution effect due to the growth rate. It is known that this degradation term is not strongly affected by the growth rate [58], so we assume that this term remains constant. In steady state, when cells are in the exponential growth phase, the concentration of mRNA is

$$mRNA = \frac{k_t \cdot p_{bound}}{\beta_{mRNA}}. \tag{2.16}$$

The Miller assay (LacZ assay) quantifies the level of LacZ expression, and we assume that the number of proteins is directly proportional to the mRNA copy number. Due to the relatively fast doubling time we assume that dilution is the relevant effect diminishing protein copy number, leading us to

$$\frac{dLacZ}{dt} = \gamma \cdot mRNA - \mu \cdot LacZ, \tag{2.17}$$

where γ is the proportionality constant of how many proteins per mRNA are produced, μ is the growth rate, and $LacZ$ is the β-galactosidase enzyme copy number. γ can be a function of the growth rate due to the changes in the number of available ribosomes, but still we argue that measuring the repression level should reduce the importance of these effects. If we substitute Equation 2.16 into 2.17 and assume steady state we obtain

$$LacZ = \frac{\gamma \cdot k_t \cdot p_{bound}}{\mu \cdot \beta_{mRNA}}. \tag{2.18}$$

By computing the repression level as measured in the LacZ assay we obtain

$$\text{repression} = \frac{LacZ(R=0)}{LacZ(R \neq 0)} = \frac{p_{bound}(R=0,P)}{p_{bound}(R \neq 0,P)}. \tag{2.19}$$

In this ratio γ, k_t, μ, and β_{mRNA} cancel each other, leaving only a ratio of p_{bound}s.

2.5.12 Related microbial species *lac* operon phylogenetic tree

See Figure 2.14.

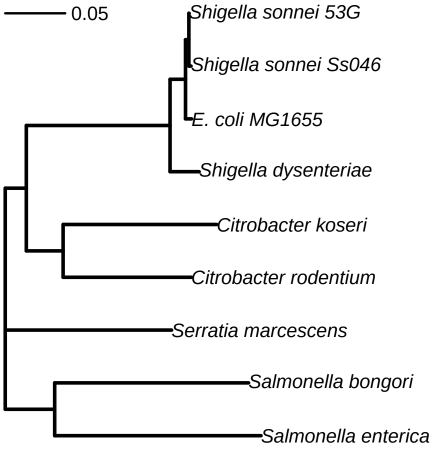


Figure 2.14: *lac* operon phylogenetic tree of diverse species with a similar *lac* promoter architecture done with the Neighbor-Joining algorithm. The scale bar represents the relative number of substitutions per sequence.

2.5.13 Epistasis analysis

Epistasis can be defined as the effect of mutations on the phenotypes caused by other mutations. Our theoretical model explicitly ignores possible interactions between mutations when calculating the transcription factor binding energies with the Sort-Seq energy matrices; but the same cannot be directly assumed for the phenotypic output. As shown in Figure 2.3 in the main text, the phenotypic response depends on the model parameters in a highly non-linear way. Given this non-linear relation we decided to perform an epistasis analysis on the data, where we defined epistasis as [59, 60]

$$\varepsilon = W_{xy} - W_x \cdot W_y, \tag{2.20}$$

where ε is the epistasis, W_{xy} is the repression value for the double mutant at positions x and y normalized to the reference MG1655 repression level, and W_x and W_y are the repression values for the single mutants in their respective positions also normalized to the same reference value. This multiplicative epistasis model indicates the type of interaction between mutations; $\varepsilon = 0$ indicates no epistasis, $\varepsilon < 0$ indicates antagonistic epistasis and $\varepsilon > 0$ indicates synergistic epistasis [59].

We calculated this epistasis metric for all the double mutants of the 134 base-pairs considered in the regulatory region of the *lac* operon including the O_2 downstream repressor binding site. For each pair of bases we calculated the epistasis for the two nucleotides with the biggest change with respect to our reference strain MG1655. Figure 2.15 shows the distribution of the epistasis values for the 8911 possible double mutants. As we initially assumed, most of the base-pairs do not interact with each other. Only 0.5% of the double mutants have an $\varepsilon < -0.5$, and 1% have an $\varepsilon > 0.5$.

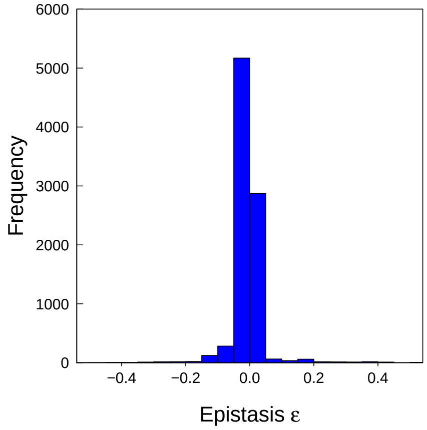


Figure 2.15: Epistasis level (Equation 2.20) distribution of all the possible double mutants of the *lac* operon regulatory region.

In order to find the base-pairs in the regulatory region predicted to have the biggest interactions Figure 2.16 shows the heat-map of the ε values. It is interesting to note that the few regions predicted to have significant epistasis fall mostly within a single binding site, i.e., basically no interaction is predicted between mutations located in different binding sites. The RNAP binding site is predicted to have antagonistic epistasis ($\varepsilon < 0$), while the CRP binding site is predicted to have strong synergistic epistasis ($\varepsilon > 0$). The O_3 binding site also presents synergistic interactions. This predicted epistasis can be attributed to the highly non-linear dependence of the repression level on these binding energies. Since, for example, the linear regime of the O_1 binding energy extends over a larger range of values (Figure 2.3 on the main text) two mutations are unable to move this parameter to the non-linear region and no epistasis would be expected at this binding site. Interestingly the only interactions between different binding sites are predicted to be between CRP and RNAP.

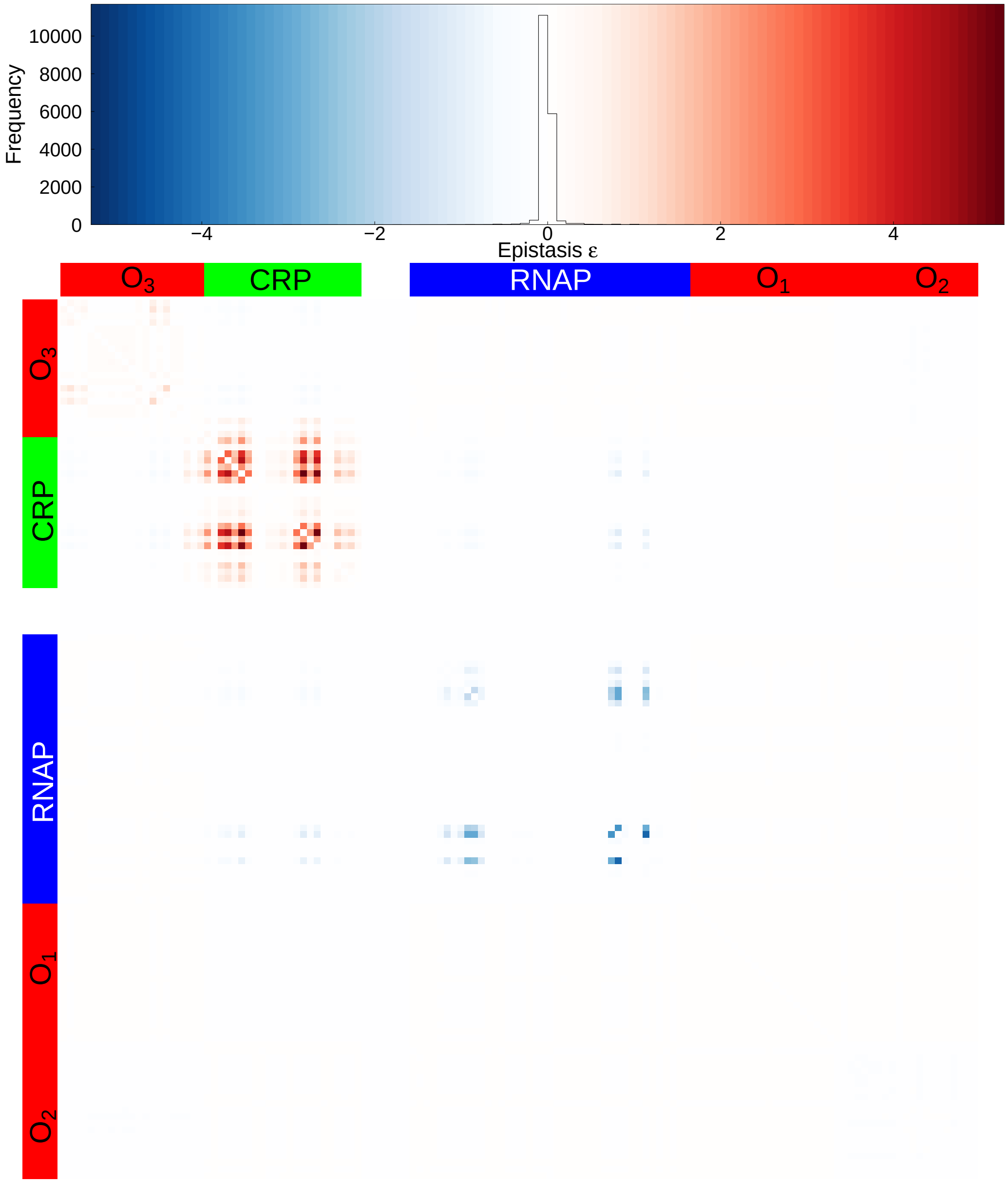


Figure 2.16: Epistasis level heat-map for all the possible double mutants. The binding sites positions are indicated with the lateral color bars.

Bibliography

[1] Janelle R Thompson, Sarah Pacocha, Chanathip Pharino, Vanja Klepac-Ceraj, Dana E Hunt, Jennifer Benoit, Ramahi Sarma-Rupavtarm, Daniel L Distel, and Martin F Polz. Genotypic diversity within a natural coastal bacterioplankton population. *Science (New York, N.Y.)*, 307(5713):1311–3, February 2005.

[2] Lior Zelcbuch, Niv Antonovsky, Arren Bar-Even, Ayelet Levin-Karp, Uri Barenholz, Michal Dayagi, Wolfram Liebermeister, Avi Flamholz, Elad Noor, Shira Amram, Alexander Brandis, Tasneem Bareia, Ido Yofe, Halim Jubran, and Ron Milo. Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Research*, 41(9):e98, May 2013.

[3] Eran Segal and Jonathan Widom. From DNA sequence to transcriptional behaviour: A quantitative approach. *Nature reviews. Genetics*, 10(7):443–56, July 2009.

[4] Thomas E Hansen. The evolution of genetic architecture. *Annual Reviews of Ecology, Evolution, and Systematics*, 37(May):123–157, 2006.

[5] Harley H McAdams, Balaji Srinivasan, and Adam P Arkin. The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics*, 5(3):169–78, March 2004.

[6] J Christian Perez and Eduardo a Groisman. Evolution of transcriptional regulatory circuits in bacteria. *Cell*, 138(2):233–44, July 2009.

[7] G K Ackers, A D Johnson, and A M Shea. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences of the United States of America*, 79(4):1129–33, February 1982.

[8] Nicolas E Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5136–41, April 2003.

[9] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics & Development*, 15(2):116–24, April 2005.

[10] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, Thomas Kuhlman, and Rob Phillips. Transcriptional regulation by the numbers: Applications. *Current Opinion in Genetics & Development*, 15(2):125–35, April 2005.

[11] Marc S Sherman and Barak A Cohen. Thermodynamic state ensemble models of cis-regulation. *PLoS Computational Biology*, 8(3):e1002407, January 2012.

[12] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *PNAS*, 107(20):9158–63, 2010.

[13] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–30, June 2012.

[14] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G Callan, Justin B Kinney, Manolis Kellis, Eric S Lander, and Tarjei S Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–7, March 2012.

[15] Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning promoter strength through rna polymerase binding site design in *Escherichia coli*. *PLoS Computational Biology*, 8(12):e1002811, December 2012.

[16] C J Wilson, H Zhan, L Swint-Kruse, and K S Matthews. The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cellular and Molecular Life Sciences*, 64(1):3–16, January 2007.

[17] W S Reznikoff. The lactose operon-controlling elements: A complex paradigm. *Molecular Microbiology*, 6(17):2419–22, September 1992.

[18] Y Setty, A E Mayo, M G Surette, and U Alon. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7702–7, June 2003.

[19] Thomas Kuhlman, Zhongge Zhang, Milton H Saier, and Terence Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6043–8, April 2007.

[20] Antony M Dean. Selection and neutrality in lactose operons of *Escherichia coli*. *Genetics*, 123:441–54, 1989.

[21] Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–92, July 2005.

[22] Lilia Perfeito, Stephane Ghozzi, Johannes Berg, Karin Schnetz, and Michael Lässig. Nonlinear fitness landscape of a molecular pathway. *PLoS genetics*, 7(7):1–10, 2011.

[23] Frank J Poelwijk, Philip D Heyning, Marjon G J de Vos, Daniel J Kiviet, and Sander J Tans. Optimality and evolution of transcriptionally regulated gene expression. *BMC Systems Biology*, 5(1):128, January 2011.

[24] Matt Eames and Tanja Kortemme. Cost-benefit tradeoffs in engineered *lac* operons. *Science (New York, N.Y.)*, 336(6083):911–5, May 2012.

[25] Jose M G Vilar. Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation. *Biophysical Journal*, 99(8):2408–13, October 2010.

[26] Hernan G Garcia and Rob Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12173–8, July 2011.

[27] James Q. Boedicker, Hernan G. Garcia, and Rob Phillips. Theoretical and experimental dissection of dna loop-mediated repression. *Physical Review Letters*, 110(1):018101, January 2013.

[28] H Tagami and H Aiba. Role of CRP in transcription activation at *Escherichia coli lac* promoter: CRP is dispensable after the formation of open complex. *Nucleic Acids Research*, 23(4):599–605, February 1995.

[29] J M Hudson and M G Fried. Co-operative interactions between the catabolite gene activator protein and the lac repressor at the lactose promoter. *Journal of Molecular Biology*, 214(2):381–96, July 1990.

[30] Valeria Souza, Martha Rocha, Aldo Valera, and Luis E Eguiarte. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Applied and Environmental Microbiology*, 65(8):3373–3385, 1999.

[31] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451(7178):535–40, January 2008.

[32] Leonor Saiz and Jose M G Vilar. *Ab initio* thermodynamic modeling of distal multisite transcription regulation. *Nucleic Acids Research*, 36(3):726–31, February 2008.

[33] Jose M G Vilar and Stanislas Leibler. DNA looping and physical constraints on transcription regulation. *Journal of Molecular Biology*, 331(5):981–989, August 2003.

[34] Jose M G Vilar and Leonor Saiz. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Current Opinion in Genetics & Development*, 15(2):136–44, April 2005.

[35] Leonor Saiz and Jose M G Vilar. Multilevel deconstruction of the *in vivo* behavior of looped DNA-protein complexes. *PloS One*, 2(4):e355, January 2007.

[36] Leonor Saiz and Jose M G Vilar. DNA looping: The consequences and its control. *Current Opinion in Structural Biology*, 16(3):344–50, June 2006.

[37] Leonor Saiz, J Miguel Rubi, and Jose M G Vilar. Inferring the *in vivo* looping properties of DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 102(49):17642–5, December 2005.

[38] Mattias Rydenfelt, Robert Cox, Hernan Garcia, and Rob Phillips. Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Physical Review E*, 89(1):012702, January 2014.

[39] H C Nelson and R T Sauer. Lambda repressor mutations that increase the affinity and specificity of operator binding. *Cell*, 42(2):549–58, September 1985.

[40] Johan Elf, Gene-Wei Li, and X. Sunney Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science (New York, N.Y.)*, 316(5828):1191–1194, 2007.

[41] Larry J Friedman and Jeff Gelles. Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. *Cell*, 148(4):679–89, February 2012.

[42] Stefan Oehler, Elisabeth R Eismann, Helmut Krämer, and Benno Müller-Hill. The three operators of the *lac* operon cooperate in repression. *EMBO Journal*, 9(4):973–979, 1990.

[43] H. Bremer and P. P. Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In Frederick C. Neidhardt et al., editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1553–1569. ASM Press, Washington DC, 1996.

[44] Stefan Klumpp and Terence Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20245–50, December 2008.

[45] K M Vossen, D F Stickle, and M G Fried. The mechanism of CAP-*lac* repressor binding cooperativity at the *E. coli* lactose promoter. *Journal of Molecular Biology*, 255(1):44–54, January 1996.

[46] Stephanie Johnson, Martin Lindén, and Rob Phillips. Sequence dependence of transcription factor-mediated DNA looping. *Nucleic Acids Research*, 40(16):7728–38, September 2012.

[47] Stefan Oehler, Michele Amouyal, Peter Kolkhof, Brigitte Von Wilcken-Bergmann, and Benno Müller-Hill. Quality and position of the three *lac* operators of *E. coli* define efficiency of repression. *EMBO Journal*, 13(14):3348–3355, 1994.

[48] Frank J Poelwijk, Daniel J Kiviet, and Sander J Tans. Evolutionary potential of a duplicated repressor-operator pair: simulating pathways using mutation data. *PLoS Computational Biology*, 2(5):e58, May 2006.

[49] Moisés Santillán. On the use of the hill functions in mathematical models of gene regulatory networks. *Mathematical Modelling of Natural Phenomena*, 3(2):85–97, 2008.

[50] Agustino Martínez-Antonio and Julio Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6(5):482–489, October 2003.

[51] Alexandre Dawid, Daniel J Kiviet, Manjunatha Kogenaru, Marjon de Vos, and Sander J Tans. Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos (Woodbury, N.Y.)*, 20(2):026105, June 2010.

[52] Justin B Kinney, Gasper Tkacik, and Curtis G Callan. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(2):501–6, January 2007.

[53] J Michael Hudson and M G Fried. DNA looping and lac repressor-CAP interaction. *Science*, 82:2–3, 1996.

[54] Steve C Schultz, George C Shields, Thomas A Steitz, and Thomas A Stejtz. Crystal structure of a CAP-DNA complex: The DNA is bent by 90 degrees. *Science (New York, N.Y.)*, 253(5023):1001–1007, 1991.

[55] W Epstein, L B Rothman-Denes, and J Hesse. Adenosine 3':5'-cyclic monophosphate as mediator of catabolite repression in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 72(6):2300–4, June 1975.

[56] J Müller, S Oehler, and B Müller-Hill. Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *Journal of molecular biology*, 257(1):21–9, March 1996.

[57] H H Ku. Notes on the Use of Propagation of Error Formulas. *Journal of Research of the National Bureau of Standards*, 70C(4):75–79, 1966.

[58] Stefan Klumpp, Zhongge Zhang, and Terence Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–75, December 2009.

[59] Daniel Segrè, Alexander Deluna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nature Genetics*, 37(1):77–83, January 2005.

[60] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3461–6, March 2008.

Chapter 3

Sort-Seq: High-throughput perturbation and characterization of regulatory DNA

3.1 Introduction

Much of the experimental work concerning transcriptional regulation done in the Phillips lab follows a similar paradigm: first, biophysical modeling is used to predict the effect on gene expression of changes in the regulatory DNA; and second, constructs incorporating such changes are cloned into *E. coli* and the level of gene expression is measured using mRNA FISH and/or a reporter gene (*e.g.*, LacZ, or a fluorescent protein). These quantitative measurements are made with high precision and have been carefully calibrated to yield results in terms of absolute numbers of mRNA or protein, allowing rigorous comparisons to be made between theory and experiment. At the same time, the throughput of this approach is somewhat limited by the need to generate and measure constructs one at a time. Moreover, its usefulness is to some extent limited to regulatory DNA that is already well characterized: for completely uncharacterized regulatory DNA, one would lack any basis for making model predictions or for choosing interesting mutations to measure experimentally.

An alternative approach, exemplified by recent work by Kinney *et al* [1], is to leverage high throughput techniques such as flow cytometry and DNA sequencing to measure gene expression from a large ($\approx 100,000$) library of mutants of some promoter region. As we will see, each individual measurement is relatively noisy and imprecise, but using techniques from information theory and machine learning, we can nonetheless characterize the function of regulatory DNA with quantitative accuracy. Specifically, we can identify regions of regulatory DNA where transcription factors (TFs) or RNA polymerase bind, fit quantitative models of TF-DNA and RNAP-DNA interaction, and fit full thermodynamic models (including protein-protein interactions) of gene regulation to sort-seq datasets. For the remainder of this introductory section, we will explore each of these points in more detail, largely following the work described in Reference [1].

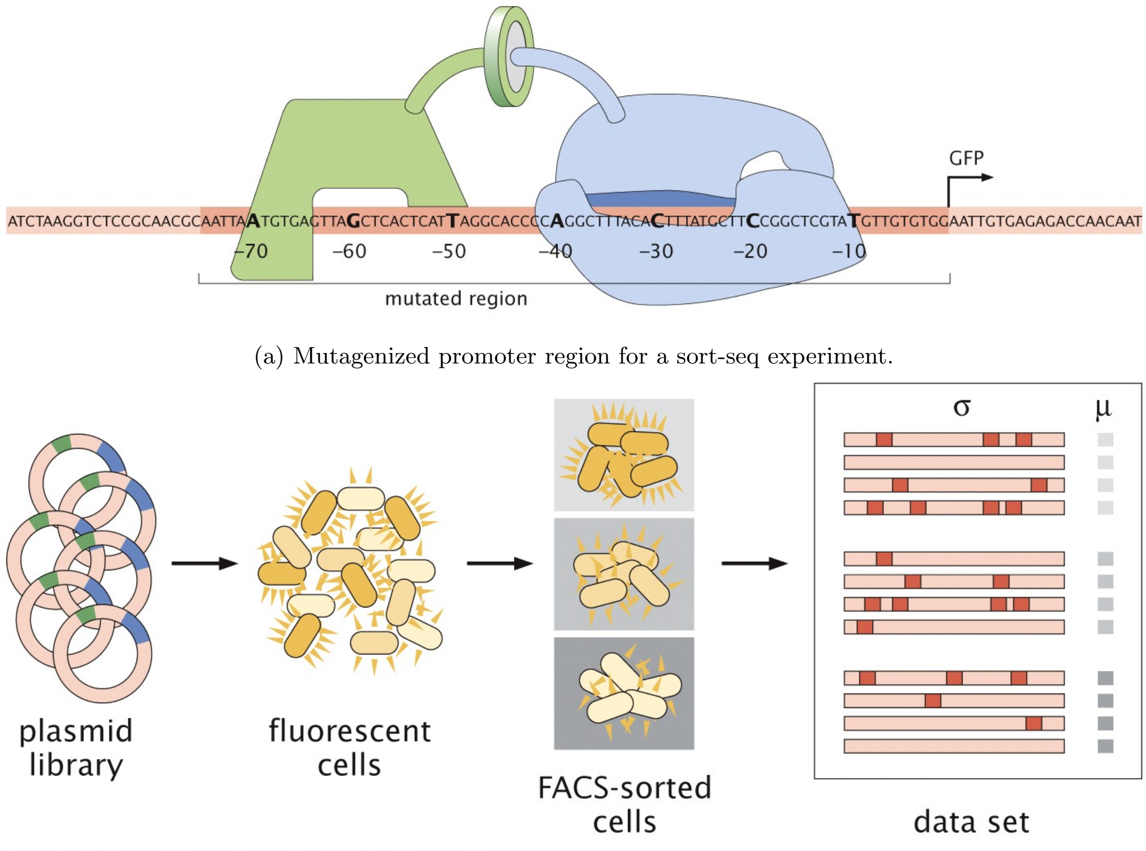
3.1.1 Materials and methods

3.1.1.1 Overview

The workflow for a sort-seq experiment is schematized in Figure 3.1. First, a plasmid library is constructed, in which each plasmid contains a mutated version of the promoter sequence for some gene of interest driving expression of GFP. The plasmids are transformed into *E. coli*, yielding a population of cells with a wide range of GFP expression levels. The population of cells is sorted into batches using a fluorescence activated cell sorting (FACS) machine. Each batch contains cells with similar levels of GFP fluorescence. The sorted cells are grown up, miniprepped, and PCR amplified with primers containing batch specific barcodes. When the resulting amplicons are sequenced using high throughput sequencing, the resulting sequence data comprises a list of mutant promoter sequences along with the batch that sequence was sorted into. Since the batch is in effect a coarse-grained measurement of gene expression, we end with a list of sequences and associated gene expression levels. Although the measurement of each sequence is relatively imprecise, we can use this data to fit high precision models of gene regulation, as will be discussed below.

3.1.1.2 Cloning and library construction

As shown in Figure 3.1, a sort-seq experiment starts with generating a plasmid library in which a promoter region of interest is mutated at a rate of around 10%. In reference [2], the authors report the construction of a set of approximately 2000 plasmids, one for each transcriptional unit in *E. coli.* Each plasmid in this set consists of a particular promoter driving expression of GFP. These plasmids will serve as the starting point for the construction of the mutant promoter libraries schematized in Figure 3.1B. Specifically, we start with the plasmid from the Zaslaver collection that contains the particular promoter we want to mutagenize. For instance, for the *lac* promoter, we start with the pUA66-lacZ plasmid from reference [2]. We next construct a cloning vector plasmid based on the pUA66-lacZ plasmid by replacing the promoter region to be mutagenized with an insert containing the *ccdB* gene flanked by BsmBI restriction sites. The ccdB toxin is fatal to *E. coli* that do not express the corresponding antitoxin ccdA, and thus the cloning vector plasmid can only be propagated in *E. coli* DB3.1 or other specialized strains containing the *ccdA* gene. At the same time, we order DNA oligos (Integrated DNA Technologies, Inc.) containing the promoter region of interest mutated at a given rate (around 10 % per base) flanked by BsaI restriction sites. We design the insert and vector sequences such that, when digested by BsmsBI and BsaI (respectively), the remaining sticky ends are complementary. Thus, to generate the plasmid library, we digest the insert oligos and vector plasmids with BsaI and BsmBI, and directly ligate the digestion product. Since the cloning vector plasmid contains the *ccdB* gene, and the plasmid libraries will be transformed into non-immune *E. coli* strains such as MG1655, we do not need to perform gel purification on the vector,


GFP
ATCTAAGGTCTCCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGAGACCAACAAT
−70
−60
−50
−40
−30
−20
−10
mutated region
σ
μ
plasmid library
fluorescent cells
FACS-sorted cells
data set

(a) Mutagenized promoter region for a sort-seq experiment.

(b) Workflow of a sort-seq experiment.

Figure 3.1: **Overview of Sort-Seq**. As described in the text, a mutant promoter library is constructed in which plasmids contain a mutagenized promoter region driving expression of GFP. This library is then transformed into *E. coli* cells. In general, some of of the promoter mutations will result in increased GFP expression, while some will result in decreased GFP expression. We use a FACS machine to sort cells into four batches based on their GFP fluorescence. Each batch is miniprepped and the mutagenized region is PCR amplified with batch-specific bar codes. Finally, the PCR products are sent for high-throughput sequencing. The resulting dataset is composed of a list of mutant promoter sequences σ, each associated with a batch number μ; the batch number for each sequence is obtained by simply reading off the batch specific barcode. Adapted from [1] via [3].

but can proceed directly to ligation of the BsmBI digested product. This cloning strategy yields typically on the order of 10 million transformants, which is more than sufficient library diversity for our purposes. See section 3.2.2 for a detailed account of library cloning (including DNA sequences) for sort-seq experiments performed on the *mscL* promoter.

Following the ligation reaction, we dialyze the ligation product for 1 hour (drop dialysis pads from Millipore, Inc) to ensure maximum transformation efficiency, then immediately transform into *E. coli* using electroporation. After 1 hour recovery in SOC medium, we plate a small fraction of the transformation reaction to estimate the number of transformants, and dilute the rest into 50 mL LB medium. This LB culture is allowed to grow overnight to saturation.

3.1.1.3 Growth and flow cytometry

Approximately 8 to 10 hours prior to sorting, the saturated library cultures are diluted 1:4000 into 25 mL M9 + 0.5% glucose minimal media. The M9 cultures are grown until mid log phase (OD600 ≈ 0.3), then diluted to OD600 ≈ 0.1 and placed on ice. This density usually yields a sorting rate of ≈ 2000 cells per second. The cells are sorted into 4 batches based on GFP fluorescence as shown in Figure 3.1. In reference [1], cells were sorted into 5 or 10 batches, but the authors found that sorting into 5 batches did not negatively affect subsequent data analysis as compared to sorting into 10 batches. This results from the fact that sorting into batches whose width is less than the fluorescence variability of a monoclonal cell population does not yield more information than sorting into broader fluorescence batches. Since four batches can be sorted at once using a FACSAria (BD Biosciences) machine, we elected to sort into four batches to avoid the need to switch sets of FACS tubes. One million cells are sorted into each batch. Following the sort, the cells are grown overnight in 10 mL LB medium. (During the sort, the cells are sorted directly into FACS tubes containing LB medium).

3.1.1.4 Library preparation

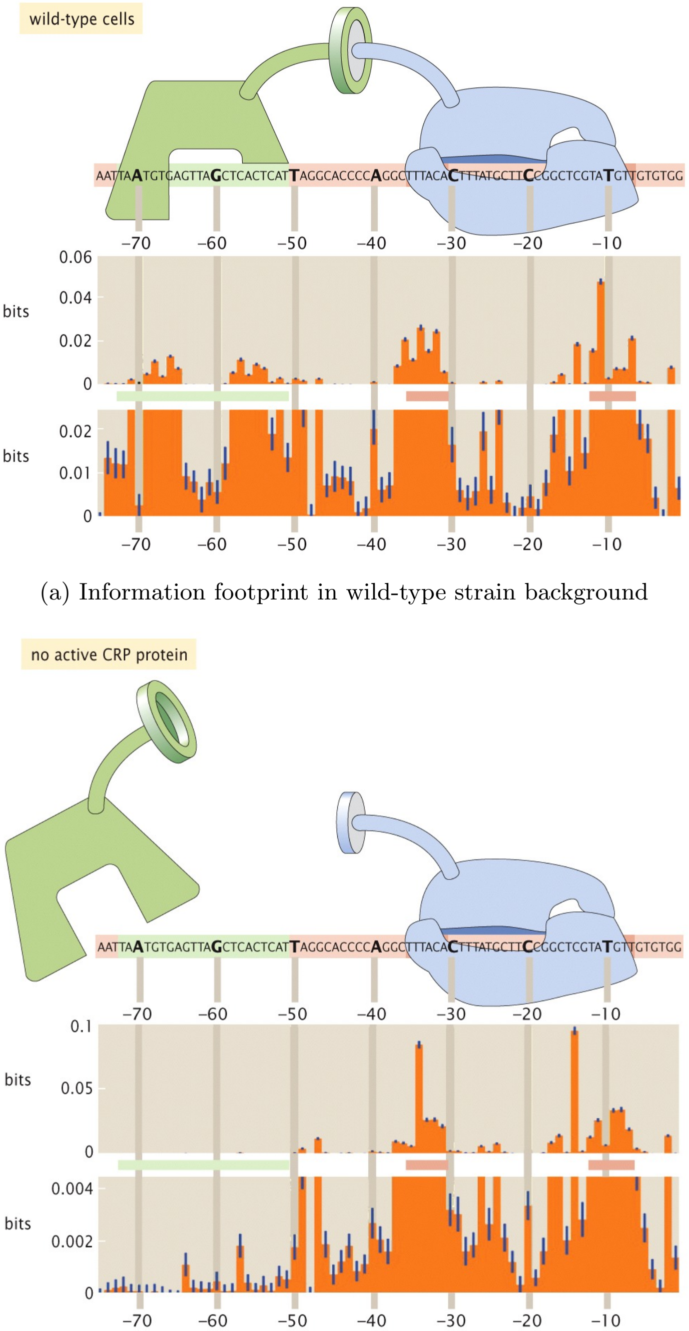
After overnight growth in LB, each batch is miniprepped separately, such that each miniprep corresponds to one FACS sorted batch. The product comprises a mix of plasmids with different promoter sequences. The miniprep product next serves as template for a PCR reaction to amplify the mutagenized promoter regions of the plasmids. The PCR primers used for this reaction include a four base pair barcode unique to the particular batch. In addition, the primers contain the adapter sequence required by the Illumina sequencing machine.

3.1.1.5 Sequencing

The PCR amplified products from each batch are mixed together and sequenced using an Illumina HiSeq or MiSeq machine. Since each batch was amplified with a unique barcode, the resulting sequencing data comprises a list of mutant promoter sequences along with a coarse-grained measurement of gene expression for that sequence (i.e.,, the batch that the particular mutant promoter was sorted into.) This data is schematized in Figure 3.1a.

3.1.2 Identifying TF and RNAP binding sites using "information footprints"

One of the simplest and most straightforward ways to analyze the resulting data is to construct an "information footprint" for the promoter region of interest. The underlying assumption is that mutating base pairs where TFs or RNAP bind will have a larger effect on gene expression than



(a) Information footprint in wild-type strain background

(b) Information footprint in CRP knockout strain.

Figure 3.2: **Information footprint for the *lac* promoter region**. The mutual information between base identity and batch number is computed for each position along the mutagenized region. As seen in the figure, regions of high mutual information correspond to transcription factor (in this case, CRP, green) and RNAP (blue) binding sites. If the sort-seq experiment is performed in a strain in which active CRP is not present, the footprint associated with CRP binding disappears, as shown in part b. Adapted from [1] via [3].

mutating base pairs where no proteins are bound. Put differently, the identity of a base pair where a TF or RNAP is bound is more informative about the resulting gene expression than the identity of a base pair where no DNA binding proteins are present. In Figure 3.2, we plot the informativeness of base pair identity at each position along a segment of the *lac* promoter. We see that regions of high informativeness correspond to positions where CRP or RNAP are bound.

We can formalize this notion of informativeness using the concept of mutual information. Specifically, let b_i be a random variable denoting the identity of the base pair (A,C,T,G) at the ith position along the promoter. Let μ denote the batch that a particular promoter sequence was sorted into. Then a sort-seq dataset allows us to define an empirical probability distribution $p_i(b,\mu)$, where for instance $p_i(\mathrm{A},2)$ denotes the observed frequency of sequences that have an A at position i and were sorted into batch 2, $p_i(\mathrm{C},2)$ denotes the observed frequency of sequences that have a C at position i and were sorted into batch 2, and so on. At the risk of being pedantic, this means that to compute $p_i(\mathrm{A},2)$ for a sort-seq dataset, we simply count the number of sequences in the dataset that were sorted into batch 2 and contain an A at position i, and divide this number by the total number N of sequences in the dataset. At a given base pair position i, then, the mutual information is defined as

$$\mathrm{I}_i = \sum_{b=A}^{T}\sum_{\mu=1}^{4}\frac{p_i(b,\mu)}{p_i(b)p_i(\mu)}, \tag{3.1}$$

where $p_i(b)$ and $p_i(\mu)$ are the marginal distributions of $p_i(b,\mu)$; *i.e.*, $p_i(b) = \sum_{\mu=1}^{4} p_i(b,\mu)$ and $p_i(\mu) = \sum_{b=A}^{T} p_i(b,\mu)$. The mutual information I_i can be understood as the amount of uncertainty in gene expression that is removed on average by knowing the identity of the base pair at position i, where uncertainty is quantified by the Shannon entropy H. For instance, say we have a dataset containing equal numbers of sequences sorted into each of four batches. If we pick a random sequence, it is equally likely to have come from any of the four batches; thus the entropy of the distribution over batches is $\log_2(4) = 2$ bits. Now assume that knowing the identify of the base pair at position i allows us to know exactly which batch the promoter is sorted into. Then the entirety of the uncertainty about gene expression is removed by knowing this base pair, and hence $I_i = 2$. This is an extreme example for illustration; in reality, no one basepair is nearly so informative about gene expression.

In reference [1], Kinney *et al* estimated the mutual information of each base pair in the mutagenized *lac* promoter region (shown in Figure 3.2), using the correction for finite sample sizes derived by Treves and Panzeri [4]. However, one drawback of using the mutual information as a metric of base pair importance is that it depends on the base substitution rates (A→C, A→G, etc) achieved for a particular experiment. Ideally these rates would be uniform (*e.g.* the rates of mutation from an A to C, G, or T would all be identical) but in reality, the DNA synthesis process tends to introduce nonuniformity in the substitution rates, because of variation in the efficiency with with different

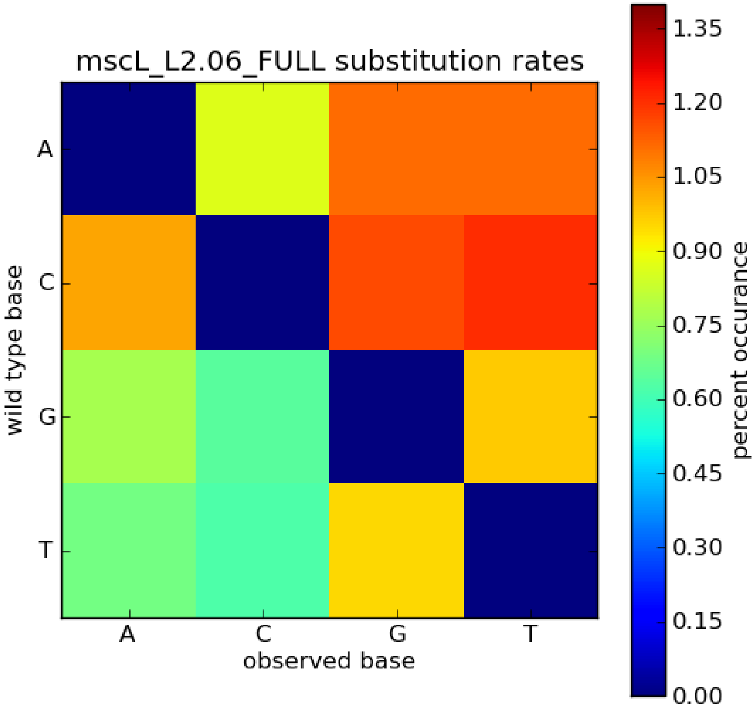


Figure 3.3: **Non-uniform base substitution rates.** Plot of observed base substitution rates for a sort-seq experiment performed on the *mscL* promoter. For clarity, diagonal elements have been set to zero. Looking at the bottom row (for instance), we see that the rate of T→G substitutions is substantially greater than the rates of T→A or T→C substitutions.

bases are incorporated during oligo synthesis. See Figure 3.3 for an illustration of this nonuniformity taken from sort-seq data. To see why the mutual information depends on these substitution rates, we rewrite equation 3.1 as follows (omitting the subscript i for convenience, and remembering that we are dealing implicitly with the mutual information at the ith base pair along the promoter):

$$I = \sum_{b=A}^{T} \sum_{\mu=1}^{4} p(b,\mu) \log_2 \left(\frac{p(b,\mu)}{p(b)p(\mu)} \right), \tag{3.2}$$

$$I = -\sum_{b=A}^{T} \sum_{\mu=1}^{4} p(b,\mu) \log_2(p(\mu)) + \sum_{b=A}^{T} \sum_{\mu=1}^{4} p(b,\mu) \log_2 \left(\frac{p(b,\mu)}{p(b)} \right), \tag{3.3}$$

$$I = -\sum_{\mu=1}^{4} p(\mu) \log_2(p(\mu)) + \sum_{b=A}^{T} \sum_{\mu=1}^{4} p(\mu|b)p(b) \log_2(p(\mu|b)), \tag{3.4}$$

$$I = H(\mu) + \sum_{b=A}^{T} p(b) \sum_{\mu=1}^{4} p(\mu|b) \log_2(p(\mu|b)), \tag{3.5}$$

$$I = H(\mu) - \sum_{b=A}^{T} p(b) H(\mu|b=b), \tag{3.6}$$

where $H(\mu)$ is the Shannon entropy of the distribution over batches $p(\mu)$, and (in a slight abuse of notation) $H(\mu|b=b)$ is the conditional Shannon entropy of the distribution over batches given a particular base pair b. From this last equation (3.6) we see that the mutual information depends on the individual base probabilities $p(b)$. To see why this is the case, consider a hypothetical scenario in which the presence of A, C, or T at a particular position has no effect on gene expression, but the presence of G is extremely favorable for gene expression. Thus, when A, C, or T are present, sequences are sorted with equal probability into any of the four batches, and hence $H(\mu|b=A) = H(\mu|b=C) = H(\mu|b=T) = 2$. But when a G is present, sequences are always sorted into the highest expression batch, and hence $H(\mu|b=G) = 0$. From equation 3.6, it is thus evident that the mutual information is highly dependent on the fraction $p(G)$ of sequences containing a G, with higher values of $p(G)$ leading to higher computed mutual information. This can be understood as resulting from the fact that the mutual information quantifies the average uncertainty that is removed from the distribution over batches by knowing the base pair b. Even if knowing that a G is present removes the entirety of the uncertainty in the batch number (as in this hypothetical example), if Gs are extremely rare, then knowing the base identity b still doesn't tell us very much about batch number on average (since most of the time, the base identity is not a G).

Since, as shown in Figure 3.3, we see that there is some nonuniformity in the base substitution rates, we would like to define a metric that independent of the base substitution rates. Specifically, we define a "renormalized" mutual information I_{renorm} as follows:

$$I_{\text{renorm}} = H(\mu) - \sum_{b=A}^{T} \frac{1}{4} H(\mu|b=b). \tag{3.7}$$

Comparing the preceding equation with equation 3.6, we see that the only difference is that $p(b)$ has been replaced with 1/4. The quantity I_{renorm} can be interpreted as an estimate of what the mutual

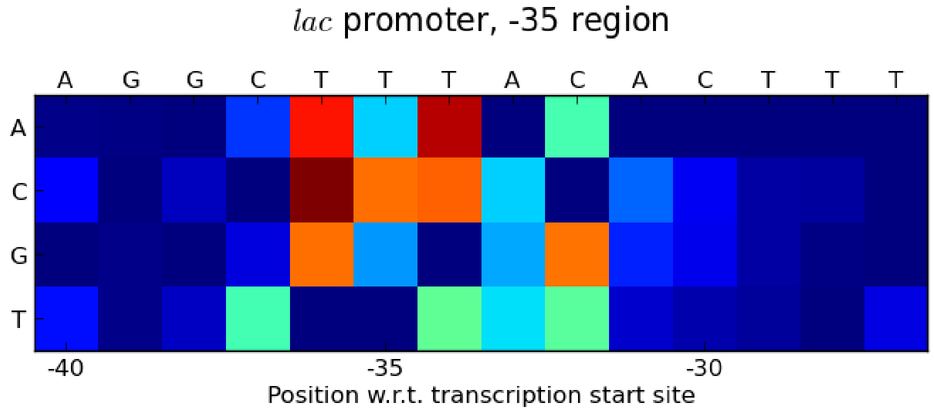


Figure 3.4: **Binding energy matrix for RNAP σ^{70} -35 recognition domain**. This matrix inferred from sort-seq experiments on the *lac* promoter, and covers the canonical -35 hexamer (positions -36 to -31) as well four flanking basepairs upstream and downstream. To compute the binding energy of a given sequence, sum the contributions from the appropriate base pairs at each position from left to right across the matrix. More negative matrix values (blue colors) correspond to more favorable binding, and more positive values (red colors) to unfavorable binding. The consensus -35 recognition sequence TTGACA is clearly visible in positions -36 to -31, corresponding to the lowest possible binding energy sequence for those positions.

information would be if all four base pairs were equally probable. For the remainder of this chapter, we will use this renormalized mutual information defined in equation 3.7 rather than the "naive" mutual information shown in equation 3.6.

3.1.3 Fitting models of protein-DNA interaction

In addition to determining where TFs and RNAP bind along a promoter region, we can use sort-seq data to fit high precision models of the sequence-dependent binding energies of DNA binding proteins. Such a model would take as input a particular DNA sequence, and output the binding energy of a given TF to that sequence. To do so, Kinney *et al* define a quantity called the error-model-averaged likelihood. We will next motivate this quantity and show how it can be used to fit models of sequence-dependent binding energies [5].

Let the symbol θ collectively denote a set of model parameters describing the sequence dependent binding energy of a DNA-binding protein. For instance, in the case of a linear binding energy model for a protein binding site of length L, θ denotes a binding energy matrix whose ijth element θ_{ij} is the energetic contribution of having base pair j present at position i along the binding site, where $i \in \{1..L\}$ and $j \in \{$A,C,T,G$\}$. An example of such a linear "energy matrix," describing binding of RNAP to the *lac* promoter -35 region, is shown in figure 3.4. In probability and statistics, the "likelihood" refers to the function

$$p(\{\mu_s\}|\theta), \tag{3.8}$$

which describes the probability of an observed dataset $\{\mu_s\}$ (where s runs from 1 to N, the number of sequences in the dataset) given a set of binding energy model parameters θ. In the case of sort-seq data, the dataset $\{\mu_s\}$ consists of a set of batch numbers associated with each promoter sequence as described above. We will make two additional assumptions about the form of the likelihood function: (1) the measurements $\{\mu_s\}$ are independent, so that $p(\{\mu_s\}|\theta) = \prod_{s=1}^{N} p(\mu_s|\theta)$; (2) the likelihood of the observed data depends on the model parameters through the model predictions $\{x_s\}$ only, where x_s is the predicted binding for the sth sequence, and hence $p(\mu_s|\theta) = p(\mu_s|x_s)$. Putting these assumptions together, we obtain

$$p(\{\mu_s\}|\theta) = \prod_{s=1}^{N} p(\mu_s|x_s). \tag{3.9}$$

A key element of equation 3.9 is the quantity $p(\mu_s|x_s)$, referred to as the "error model." In the case where we are trying to model the sequence dependent binding energy of a TF to DNA, x_s denotes the TF binding energy for the sth mutant promoter, and $p(\mu_s|x_s)$ denotes the probability that the mutant was sorted into batch μ_s given that the binding energy is x_s. A moment's reflection reveals that this quantity would be extremely difficult to calculate *a priori*, as it depends on the physical relationship between binding energy x_s and gene expression (which could, in principle, be calculated from a thermodynamic model), the distribution of mutations in other TF binding sites in the mutagenized region, the particular GFP fluorescence range chosen for each batch on the FACS machine, the noise in FACS GFP fluorescence measurements, and the rate of mis-sorting events.

The impracticability of calculating the error model $p(\mu_s|x_s)$ led Kinney *et al* to consider the effect of averaging over an ensemble of error models [5]. Specifically, they computed the form of the likelihood $p(\{\mu_s\}|\theta)$ if one averages over all possible error models $p(\mu_s|x_s)$ with a uniform prior on the space of error models, and found that in the large data limit $N >> 1$, the likelihood can be written as

$$p(\{\mu_s\}|\theta) = \text{const} \times 2^{N\mathrm{I}(\mu,x)}, \tag{3.10}$$

where N is the number of sequences, and $\mathrm{I}(\mu, x)$ is the mutual information between predicted binding energies $\{x_s\}$ and batch numbers $\{\mu_s\}$, estimated from the sort-seq dataset. If we assume a uniform prior on the space of model parameters θ, the probability distribution $p(\theta|\{\mu_s\})$ over values of θ given the observed data $\{\mu_s\}$ is directly proportional to the likelihood since

$$p(\theta|\{\mu_s\}) = \frac{p(\{\mu_s\}|\theta)p(\theta)}{p(\{\mu_s\})}, \tag{3.11}$$

$$p(\theta|\{\mu_s\}) \propto p(\{\mu_s\}|\theta), \tag{3.12}$$

and hence (from equation 3.10)

$$p(\theta|\{\mu_s\}) = \text{const} \times 2^{N\mathrm{I}(\mu,x)}. \tag{3.13}$$

Now that we have an expression for the posterior distribution of θ, we can use Markov Chain Monte

Carlo methods to sample from the distribution, and thus can compute the expected value (as well as other statistical quantities) of model parameters θ given the observed data. But in order to do so, we will need to estimate the mutual information $\mathrm{I}(\mu, x)$ between observed data and model predictions, which as we will see is not a trivial problem.

3.1.3.1 Estimating mutual information between observed data and model predictions

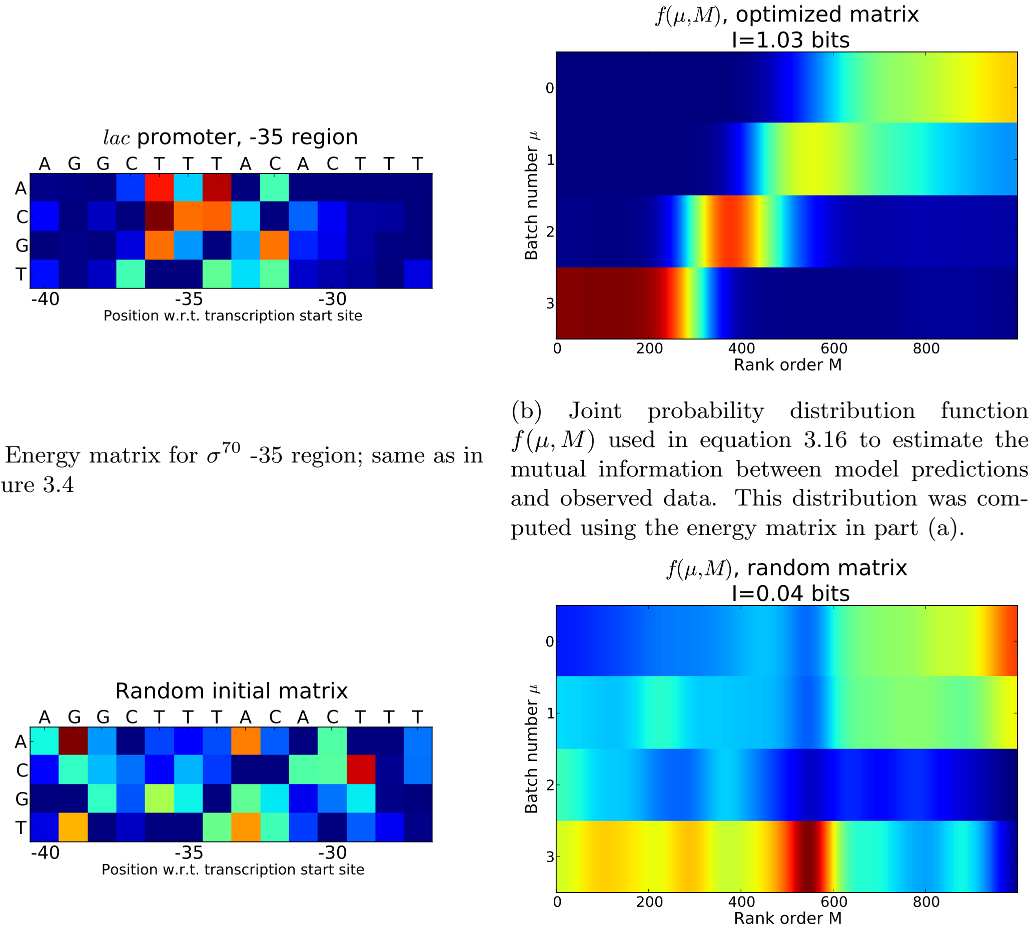
One difficulty with estimating $\mathrm{I}(\mu, x)$ is that unlike in the information footprint calculations above, where basepair identity {A,C,T,G} and batch number μ are both discrete variables, the batch number μ is discrete while the binding energy x is a continuous variable. Formally, the mutual information is given by

$$\mathrm{I}(\mu, x) = \int_{x=-\infty}^{x=+\infty} dx \sum_{\mu=\{1,2,3,4\}} p(x,\mu) \log_2 \left(\frac{p(x,\mu)}{p(x)p(\mu)} \right). \tag{3.14}$$

The problem is that we don't have direct access to the continuous probability distribution $p(x)$, but instead have only a set of values $\{x_s\}$ corresponding to predicted binding energies for each sequence in the dataset. In general, estimating continuous probability distributions based on discrete data is a highly nontrivial problem. To sidestep this issue, we will use the fact that for any transformation $z(x_s)$ that preserves the rank order of the x_s (for instance, adding a constant to each prediction x_s), the mutual information is unchanged; that is,

$$\mathrm{I}(\mu, x) = \mathrm{I}(\mu, z(x)). \tag{3.15}$$

In order to work with discrete quantities, we will define $z_i(x_s)$ as the rank order in binding energy of the sth sequence, and will estimate $\mathrm{I}(\mu, z)$. Again at the risk of being pedantic, to find the "rank order" we compute the predicted energy x_s for each sequence, then sort the sequences from lowest to highest according to their predicted binding energy. Then the sequence with rank order 1 is the sequence with the lowest predicted binding energy, the sequence with rank order 2 has the second lowest predicted energy, and so on. Thus, z runs from 1 to N where N is the total number of sequences under consideration. To estimate $\mathrm{I}(\mu, z)$, Kinney et al used a procedure that, while not strictly mathematically rigorous, appears to work well in practice [1]. We bin the rank orders z_i into 1000 bins and define an empirical frequency matrix $F(\mu, M)$ as the number of sequences sorted into batch μ that are in the Mth rank order bin. To make this a bit more concrete, let's consider a scaled down scenario in which 20 sequences are sorted into 2 batches, and the rank orders are binned into 4 bins. Then $F(1, 1)$ is the number of sequences sorted into batch one whose energy rank orders are between 1 and 5 inclusive, $F(1, 2)$ is the number of sequences sorted into batch 1 whose energy rank orders are between 6 and 10 inclusive, $F(2, 3)$ is the number of sequences sorted into batch 2 whose energy rank orders are between 11 and 15 inclusive, and so on. Finally, so that sum of all matrix elements is one, we define a normalized version of F as $\tilde{f}(\mu, M) = \frac{1}{N} F(\mu, M)$.



(a) Energy matrix for σ^{70} -35 region; same as in Figure 3.4

(b) Joint probability distribution function $f(\mu, M)$ used in equation 3.16 to estimate the mutual information between model predictions and observed data. This distribution was computed using the energy matrix in part (a).

(c) Random energy matrix, used to initialize a MCMC run.

(d) Joint probability distribution function $f(\mu, M)$ used in equation 3.16 to estimate the mutual information between model predictions and observed data. This distribution was computed using the energy matrix in part (c).

Figure 3.5: **Energy matrices and regularized distributions $\mathbf{f(\mu, M)}$.** For the distributions $f(\mu, M)$, shown in parts (b) and (d), red indicates regions of higher probability while blue indicates regions of lower probability. The x axis corresponds to the rank order of predicted binding energy, and the y axis to observed batch number. For the optimized energy matrix (part (a)), we see sequences whose rank ordered binding energies are low (i.e., strong binding sequences) are clustered in the highest expression batch (batch 3), while sequences whose rank ordered binding energies are high (poor binding sequences) are clustered in the lowest expression batch. The estimated mutual information value is relatively high, at 1.03 bits. For the random energy matrix, shown in part (c), there is no clear association between predicted energy and batch number, which is reflected in the low estimate mutual information, at 0.04 bits.

Empirically, it turns out that using this matrix $\tilde{f}$ to directly estimate mutual information (and in turn, likelihood) yields landscapes that are excessively rough in θ space. To ameliorate this issue, we define finally a regularized matrix $f(\mu, M)$ that is simply $\tilde{f}$ convolved along the binned rank order axis with a Gaussian with standard deviation equal to 40 (or 4% of the total number of sequences). (It should be noted that there is no principled or rigorous justification for this procedure: it is simply something that seems to work well in practice. See reference [1] for additional discussion of these issues.) We can now use $f(\mu, M)$ to estimate mutual information using the formula

$$\mathrm{I}_{\mathrm{smooth}}(\mu, z) = \sum_{M=1}^{1000} \sum_{\mu=1}^{4} f(\mu, M) \log_2 \left(\frac{f(\mu, M)}{f(\mu) f(M)} \right). \tag{3.16}$$

Hence, to compute the posterior probability of model parameters θ, we simply plug the estimated mutual information $\mathrm{I}_{\mathrm{smooth}}(\mu, z)$ into equation 3.13, yielding

$$p(\theta | \{\mu_s\}) = \mathrm{const} \times 2^{N \mathrm{I}_{\mathrm{smooth}}(\mu, z)}. \tag{3.17}$$

3.1.3.2 Markov Chain Monte Carlo (MCMC) sampling of model parameters θ

A full discussion of Markov Chain Monte Carlo (MCMC) methods is beyond the scope of this work, but a brief introduction will be provided here for clarity. For this author, the technical report by Neal [6] was a valuable resource, and many textbooks and other resources exist as well. Markov Chain Monte Carlo methods are frequently used to compute properties of probability distribution functions that are not amenable to analytic calculations. To compute statistical properties such as the expected value $\langle \theta \rangle = \int \theta p(\theta) d\theta$ of a distribution $p(\theta)$, it is necessary to perform various integrals over the distribution; if $p(\theta)$ takes a complicated functional form (as it does in many "real life" probabilistic models), it is often infeasible to perform these integrals analytically. For the current case of inferring models of protein-DNA interaction, equation 3.17 allows us to compute the probability $p(\theta | \{\mu_s\})$ of a particular set of model parameters θ given a set of experimental measurements $\{\mu_s\}$. The expected value of θ is given by

$$\langle \theta \rangle = \frac{\int \theta \times \mathrm{const} \times 2^{N I_{\mathrm{smooth}}(\mu, z)} d\theta}{\int \mathrm{const} \times 2^{N I_{\mathrm{smooth}}(\mu, z)} d\theta}, \tag{3.18}$$

where the integrals run over all possible values of theta (and the denominator accounts for the unknown constant in equation 3.17). Unfortunately, it is not possible to perform the integrals analytically.

The idea behind MCMC methods is that even if we can't directly compute these integrals, as long we can draw samples according to the distribution, we can still use these samples to estimate properties of the distribution. For instance, to estimate the expected value $\langle \theta \rangle$, we draw (say) 100 samples $\{\theta_1, \theta_2, \ldots, \theta_{100}\}$ according to $p(\theta)$, and then compute the sample mean:

$$\langle\hat{\theta}\rangle = \frac{\sum_{l=1}^{100} \theta_l}{100}. \tag{3.19}$$

We want to construct a Markov chain whose stationary distribution converges to the distribution of interest $p(\theta)$. A Markov chain is a sequence of values $\{\theta_1, \theta_2, \ldots, \theta_{100}\}$ that has no memory; that is, the probability that the lth value in the chain takes a value θ_l depends **only** on the $(l-1)$th value in the chain. To make things a bit more concrete, let's leave aside θ for the time being. Imagine that we have a light switch, and we know the switch is "on" 25% of the time, and "off" 75% of the time, so that $p(\text{on}) = 0.25$ and $p(\text{off}) = 0.75$. Once a second, we can change the state of the switch; if the switch is on, we turn it off with probability k_{off}, and if the switch is off, we turn it on with probability k_{on}. The sequence of states {on, off, off, on, off, off, off, off} of the light switch constitutes a Markov chain (this is just one possible sequence of states for illustration). To find the stationary distribution of this Markov chain, we would let it run for a long time, then compute the fraction $p_s(\text{on})$ of states in the chain that are "on", and the fraction $p_s(\text{off})$ of states that are "off". Our goal here is to choose values for k_{on} and k_{off} such that $p_s(\text{on}) = 0.25$ and $p_s(\text{off}) = 0.75$; or, in other words, to construct a Markov chain whose stationary distribution p_s is equal to the distribution p of on and off values.

A Markov chain is stationary if (but not only if) detailed balance is satisfied between its states. The condition of detailed balance obtains if the total rate of transitions from on to off is the same as the total rate of transitions from off to on. Mathematically, this condition can be written as

$$k_{on} \times p(\text{off}) = k_{off} \times p(\text{on}). \tag{3.20}$$

This equation means that if we choose values for k_{on} and k_{off} such that equation 3.20 is satisfied, we will have constructed a Markov chain whose stationary distribution is given by $p(\text{on})$ and $p(\text{off})$. Rearranging equation 3.20, we obtain

$$\frac{k_{on}}{k_{off}} = \frac{p(\text{on})}{p(\text{off})}, \tag{3.21}$$

$$\frac{k_{on}}{k_{off}} = \frac{0.25}{0.75} = \frac{1}{3}, \tag{3.22}$$

where we have substituted the values of $p(\text{on})$ and $p(\text{off})$ given above. This equation tells us the ratio between the transition probabilities k_{on} and k_{off}, but not their absolute scale. For convenience, we set $k_{off} = 1$, and hence $k_{on} = 1/3$. Thus, to construct a Markov chain with the desired stationary distribution, we start in either the on or the off state. Every second, we can undergo a transition to the other state; if we are on, we transition to the off state with probability 1, and if we are off, we transition to the on state with probability 1/3 and remain in the off state with probability 2/3. If we let the chain run for a sufficiently long time (in general, there is no simple definition of what constitutes a sufficiently long time, so various empirical convergence tests are used), the distribution of on and off states will converge to the desired distribution $p(\text{on}) = 0.25$ and $p(\text{off}) = 0.75$.

This is obviously a highly simplified example, but it nonetheless effectively illustrates several general points. Even for more complicated scenarios where more than two states are possible, it remains the case that detailed balance is a pairwise condition. As long as the transition rates $k_{i\to j}$ $k_{j\to i}$ for any pair of states i and j are chosen such that detailed balance is satisfied, *i.e.*,

$$k_{i\to j} \times p(i) = k_{j\to i} \times p(j), \tag{3.23}$$

then the resulting Markov chain will have a stationary distribution given by $p(i)$ and $p(j)$. The same is also true for continuous variables.

Returning to the case of sampling model parameters from the distribution $p(\theta)$ defined in equation 3.17, we define the following algorithm. It is nothing more than the standard Metropolis-Hastings algorithm applied to our particular scenario. (Recall that θ denotes a binding energy matrix whose ijth element θ_{ij} is the energetic contribution of having base pair j present at position i along the binding site, where $i \in \{1..L\}$ and $j \in \{\text{A,C,T,G}\}$. However, the analysis described here could equally well apply to more complicated models incorporating dinucleotide interactions.)

1. Start with a random energy matrix θ_0.

2. Make a random perturbation $d\theta$ to θ_0.

3. Compute the probabilities $p(\theta_0)$ and $p(\theta_0 + d\theta)$ using equation 3.17. (Recall that $p(\theta)$ is proportional to 2 raised to the power of N times the mutuppal information between model predictions and observed data).

4. If $p(\theta_0+d\theta) > p(\theta_0)$, accept $\theta_0+d\theta$ as the next element θ_1 of the Markov chain. Otherwise, accept $\theta_0+d\theta$ with probability $p(\theta_0+d\theta)/p(\theta_0)$, and reject with probability $1-p(\theta_0+d\theta)/p(\theta_0)$. Rejection means that the next element θ_1 in the Markov chain is again θ_0. These acceptance/rejection probabilities mean that detailed balance is satisfied between the states θ_0 and $\theta_0 + d\theta$.

5. Go to step 2 (replacing θ_0 and θ_1 with the appropriate elements θ_l and θ_{l+1}) and repeat until the chain converges to the stationary distribution. In practice, convergence can be monitored by tracking the mutual information as shown in Figure 3.6.

The end result of these model-fitting efforts is an optimized linear binding energy matrix like the one shown in Figure 3.4 or 3.5a. One question that has been neglected thus far is that of units,

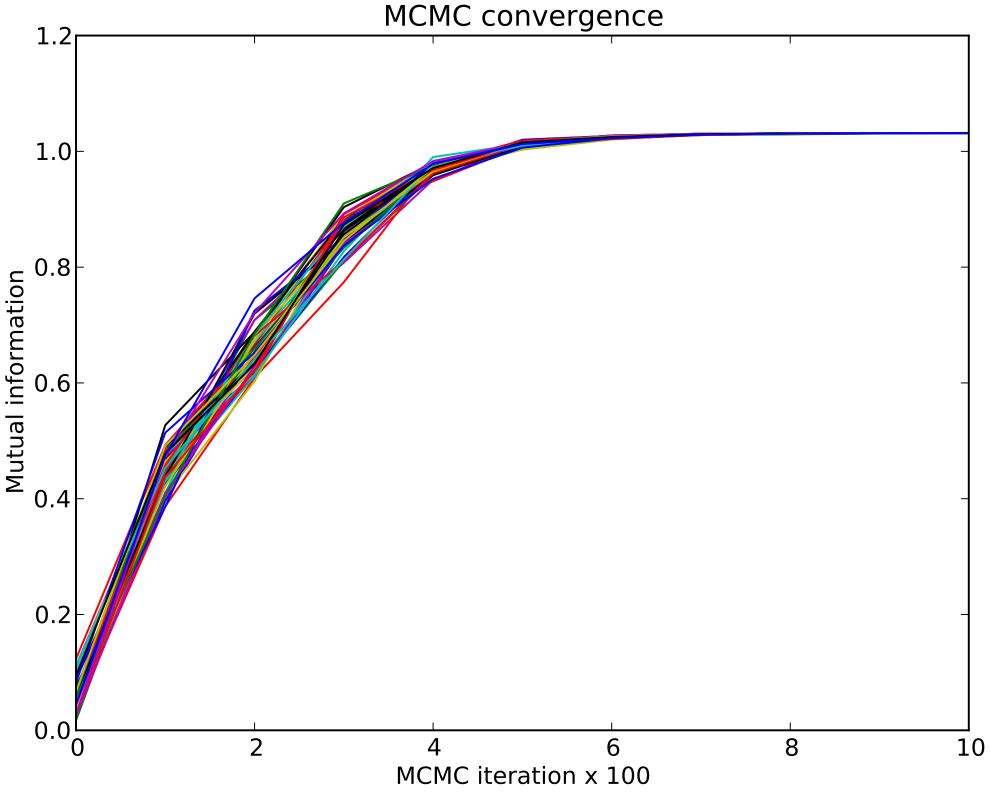


Figure 3.6: **MCMC convergence.** Plot of estimated mutual information vs. MCMC iteration number. The Markov chain is initialized with a random matrix as in Figure 3.5c, and converges to the "correct" (or at the very least, locally optimal) energy matrix of Figure 3.5a by about the 600th iteration. To ensure that we aren't stuck in a local minimum, we initialize multiple (≈ 100) chains from random starting points and check that all chains converge to the same matrix.

which we will now address. The short answer to this question is that the units are arbitrary, and can only be calibrated to physical units (*e.g.* k_BT or kcal) using external information, such as two sequences with known binding energies.

The more involved answer can perhaps be safely skipped, but will briefly be presented here. We begin by noting that some "degrees of freedom" are unconstrained by this model-fitting procedure. In particular, adding a constant to all matrix elements or to all elements in a particular column will not change the energy **difference** between two particular DNA sequences. In a similar vein, multiplying all matrix elements by a constant will change the energy difference between two sequences, but not which sequence has a greater predicted energy than the other. Thus, neither of these transformations to an energy matrix will affect the mutual information between predicted energies and observed batch number for a sort-seq dataset, since the rank orders of the predicted binding energies will be unchanged. Consequently, we (somewhat) arbitrarily impose the gauge conditions that (1) the sum of matrix elements in each column is zero (2) the sum of the squares of all matrix elements is one. These choices have two convenient properties; namely, that (1) the average energy across all possible random DNA sequences is zero (2) the standard deviation in energy across all possible random sequences is one. This means that the predicted binding for a particular sequence can instantly be interpreted as a sort of "z-score," in the sense that *e.g.* a sequence with a predicted energy of -1 has a binding energy one standard deviation less than the average binding energy of random DNA. To convert from these arbitrary energy units to physical units, we can perform a calibration using two sequences of known energy. The details of how to do so are described in the following chapter, in section 4.5.3. Finally, we note that in Figure 3.4 and all subsequent figures showing energy matrices, the smallest matrix element has been subtracted from each column. This is simply to make the matrices easier to interpret visually, as the optimal base pair at each position (*i.e.*, in each column) will always be dark blue, and the other elements in each column can be interpreted as the difference in energy between a particular choice of base pair and the optimal base pair. However, this has no effect on the physical information conveyed by the matrix, and is simply for convenience.

3.2 Exploring uncharacterized regulatory DNA: the *mscL* promoter as a case study

3.2.1 Introduction

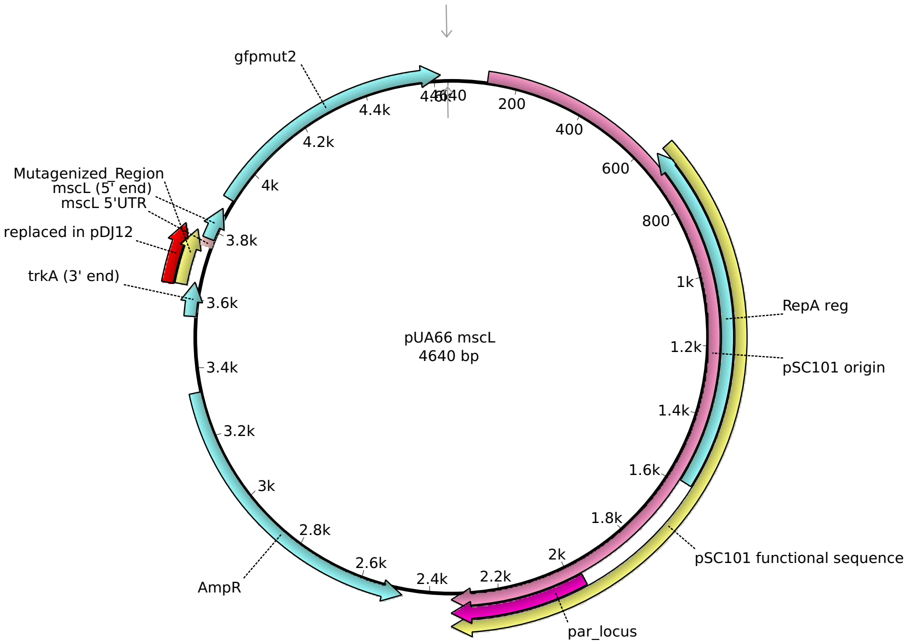
Mechanosensitive channels - ion channels embedded in cell membranes that gate based on membrane tension - are ubiquitous in living organisms. Homologs of the *E. coli* mechanosensitive channels of large and small conductance (*mscL* and *mscS*, respectively) are found in organisms ranging from *E. coli* to *Arabidopsis thaliana*, while *E. coli* alone contains genes encoding no less than seven distinct mechanosensitive channels. Yet the function and physiology of these channels remains unclear. The conventional view is that these channels protect the integrity of the cell membrane in the event of osmotic downshock (*i.e.*, the transfer of cells from an environment of high osmolarity to an environment of low osmolarity) [7]. If the resulting osmotic pressure threatens to rupture the cell membrane, the mechanosensitive channels will gate in response to increased membrane tension, thereby alleviating osmotic pressure. This view appears to be at least partially correct, in the sense that physiological experiments have demonstrated that deleting the mechanosensitive channels MscS and MscL from *E. coli* decreases survivability by at least tenfold in osmotic downshock assays [8]. Conversely, the presence of either mscS or mscL alone is sufficient to confer the same level of osmoprotection as all seven channels in osmotic downshock assays [8]. Moreover, the number of MscL proteins present in *E. coli* appears to be as many as 100 times greater than the number required to confer osmoprotection [9–11].

These uncertainties about channel physiology and function are mirrored in the scant information available about their regulation. In fact, the *mscL* gene is completely un-annotated in the transcriptional regulatory database RegulonDB [12]. The only information directly known about its transcriptional regulation is that its expression is upregulated by the stress response sigma factor RpoS (also referred to as σ^{S} and σ^{38}) [9]. For these reasons, we decided to use sort-seq methods described above to attempt to illuminate the transcriptional regulation of *mscL*. We hoped that learning about its transcriptional regulation could help to shed light on its physiology and function, and furthermore, that these efforts could serve as a case study in the use of biophysical methods to understand transcriptional regulation *de novo* in a poorly characterized system.

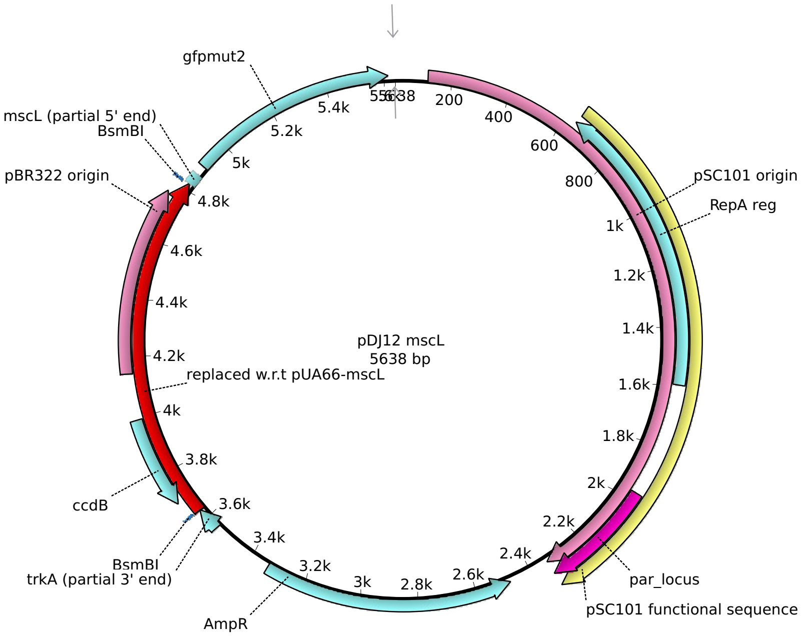
3.2.2 Materials and methods.

We took as a starting point the pUA66-mscL plasmid from the Zaslaver collection (Figure 3.7a) [2]. This plasmid contains the intergenic region between *mscL* and its upstream neighbor *trkA*, extending 97 base pairs upstream into the *trkA* open reading frame (ORF) and 94 base pairs into the *mscL* ORF, driving expression of GFP, as seen in Figure 3.7a. We mutagenized the region indicated in the schematic; this mutagenized region extends from 118 bp upstream of the transcription start site (12 bp into the *trkA* ORF) to 36 bp downstream of the transcription start site (13 bp into the *mscL* ORF). Our target mutation rate for this region was 12%, but unfortunately a miscommunication with IDT caused the actual mutation rate to be 3%.

In order to construct the mutant promoter library, we first created a cloning vector plasmid (pDJ12) based on pUA66-mscL (Figure 3.7b). pDJ12 differs from pUA66-mscL in that the region of pUA66-mscL to be mutagenized has been replaced with an insert (annotated in red in Figure 3.7b) containing the *ccdB* gene, which encodes a toxin that is fatal to *E. coli* which do not carry the corresponding antitoxin *ccdA*. The insert also carries the high-copy pBR322 origin of replication for



(a) **Map of plasmid pUA66-mscL.** This plasmid contains the *mscL* promoter driving expression of GFP. The mutagenized region and the region replaced in the cloning vector pDJ12 are indicated (between 3.6k and 3.8k). The red region (replaced in pDJ12) contains the mutagenized region with an additional 3 bp flanking each side.
.



(b) **Map of plasmid pDJ12.** This plasmid is essentially a cloning vector version of pUA66-mscL. The region indicated in red in part (a) is replaced by an insert (red, ≈3.6k - 4.8k) containing the *ccdB* gene, which is toxic in standard laboratory *E. coli* strains. The insert is flanked by dual BsmBI restriction sites (detail in Figure 3.9).

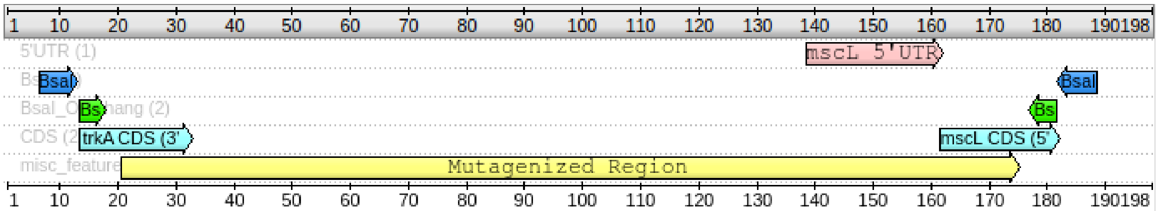
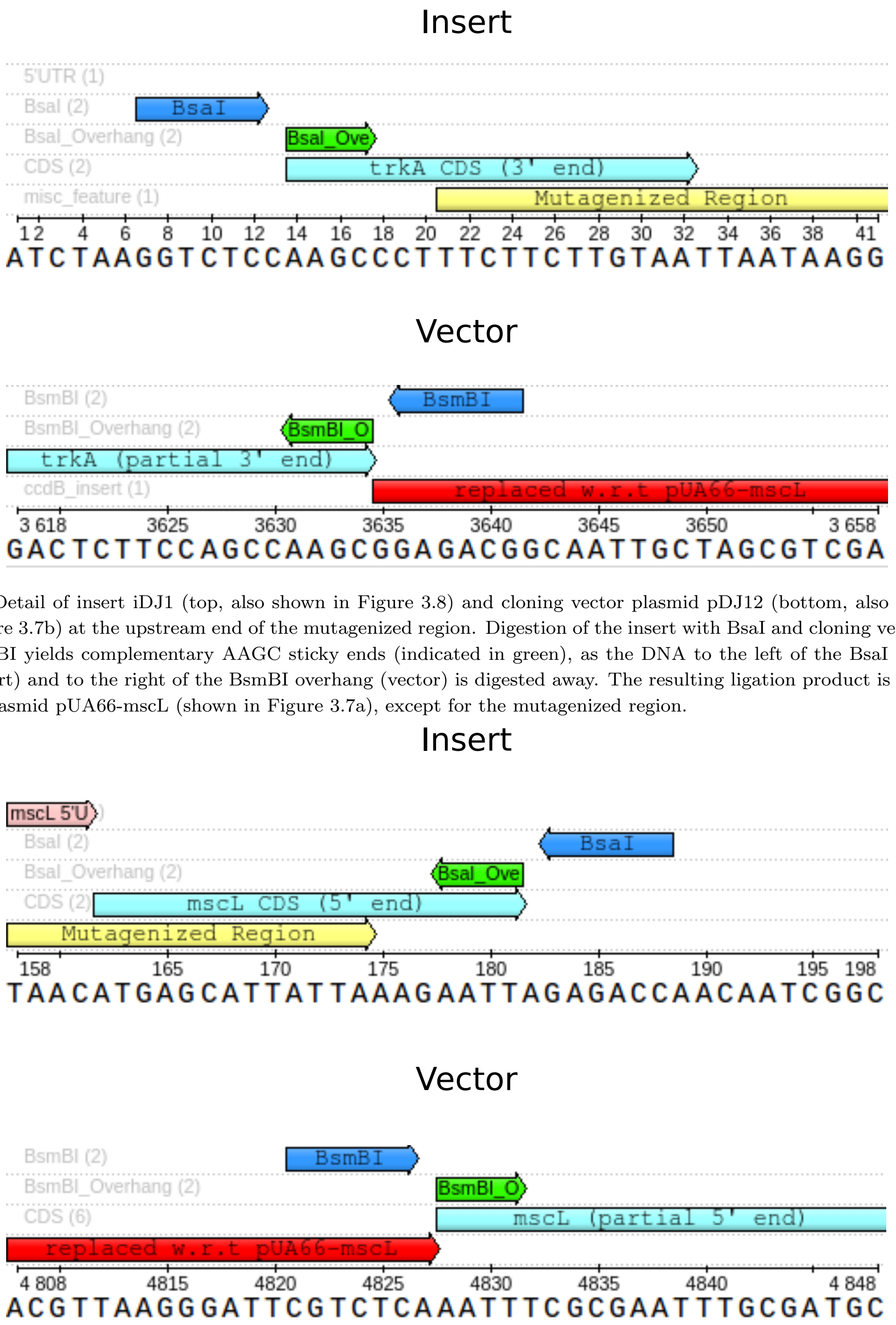


Figure 3.8: **Schematic of insert iDJ1 ordered for library cloning.** The insert consists of the mutagenized region flanked by BsaI binding sites. BsaI is a type II restriction enzyme and thus upon digestion leaves an overhang (green annotations) downstream of its binding site (blue annotations), as indicated in the schematic. The sequence is designed such that the overhangs are complementary to the overhangs resulting from digestion of pDJ12 with BsmBI, enabling convenient ligation. See Figure 3.9 for a detailed schematic of the ligation.

convenience (in the sense that fewer cells are needed to obtain a given yield of plasmid DNA in minipreps). The insert region is flanked by BsmBI restriction enzyme binding sites. Aside from the insert region indicated in red in Figure 3.7b, pDJ12 is identical to pUA66-mscL.

To obtain the actual mutant promoter library, we ordered DNA oligos from Integrated DNA Technologies, Inc. A schematic of the insert DNA oligo iDJ1 is shown in Figure 3.8. The insert oligo consists of the mutagenized region flanked on either side by 7 bp of constant DNA, flanked in turn by BsaI restriction enzyme binding sites. The insert oligos are synthesized as single stranded and are thus made double stranded by PCR prior to ligation. These insert oligos are of course comprised of a pool of millions of variants of the *mscL* promoter flanked by constant regions as shown in the schematic (Figure 3.8).

To perform the ligation, we first perform a miniprep to obtain purified pDJ12 plasmid. 30 μL of 100 ng/μL plasmid is a typical yield for a single miniprep. Because of the presence of the *ccdB* gene, pDJ12 must be propagated in a specialized *E. coli* strain such as DB3.1 containing the gene for the corresponding antitoxin *ccdA*. pDJ12 miniprep product is then digested with BsmBI, cutting out the region annotated in red in Figure 3.7b. At the same time, the double stranded insert iDJ1 is digested with BsaI. The products of the two digestion reactions are then mixed together (10 fmol of pDJ12 digestion product and 30 fmol of iDJ1 digestion product) and ligated with T4 DNA Ligase (Invitrogen, Inc.). Figure 3.9 shows a detailed view of the restriction and ligation reactions. The sequences were designed such that digestion of pDJ12 with BsmBI and iDJ1 with BsaI yields complementary sticky ends, as shown in Figure 3.9. The net result of the ligation reaction is a pool of plasmids identical to pUA66-mscL except that each contains a slightly different version of the mutagenized promoter sequence. One notable feature of this cloning strategy is that the presence of *ccdB* in plasmid pDJ12 means that gel purification of the vector backbone after BsmBI digestion is not necessary. Although in some cases the red insert region in Figure 3.7b will be re-ligated into the vector backbone (instead of the desired iDJ1 insert containing the mutagenized promoter region), because the ligation product is being transformed into a strain that is not immune to the



(a) Detail of insert iDJ1 (top, also shown in Figure 3.8) and cloning vector plasmid pDJ12 (bottom, also shown in Figure 3.7b) at the upstream end of the mutagenized region. Digestion of the insert with BsaI and cloning vector with BsmBI yields complementary AAGC sticky ends (indicated in green), as the DNA to the left of the BsaI overhang (insert) and to the right of the BsmBI overhang (vector) is digested away. The resulting ligation product is the same as plasmid pUA66-mscL (shown in Figure 3.7a), except for the mutagenized region.

(b) Detail of insert iDJ1 (top, also shown in Figure 3.8) and cloning vector plasmid pDJ12 (bottom, also shown in Figure 3.7b) at the downstream end of the mutagenized region. Digestion of the insert with BsaI and cloning vector with BsmBI yields complementary AATT sticky ends (indicated in green), as the DNA to the right of the BsaI overhang (insert) and to the left of the BsmBI overhang (vector) is digested away. The resulting ligation product is the same as plasmid pUA66-mscL (shown in Figure 3.7a), except for the mutagenized region.

Figure 3.9: Detail of ligation reactions at upstream and downstream ends of the mutagenized region.

ccdB toxin, cells containing these plasmids will simply fail to grow. The resulting slight decrease in ligation efficiency is more than compensated by the fact that gel purification is unnecessary, and in any case is ameliorated by adding digested iDJ1 product in molar excess.

3.2.3 Gene expression assays under various environmental conditions.

In preparation for performing sort-seq experiments, gene expression from the wild-type pUA66-mscL plasmid (which contains the wild-type *mscL* promoter driving GFP expression) was measured under various genetic backgrounds and environmental conditions. Expression was measured by measuring median GFP fluorescence across a population of 50,000 cells using a BD Biosciences LSRII flow cytometer. As shown in Figure 3.10, we measured expression in strains MG1655, MG1655 with *rpoS* deleted, MG1655, with *mscL* deleted, and strain MG1655, in which the *mscS* (mechanosensitive channel of small conductance), *mscK* (mechanosensitive channel of medium conductance), and *mscL* genes are deleted. The *rpoS* gene encodes for the *E. coli* stationary phase/stress response sigma factor RpoS, also known as σ^{S} and σ^{38}. Previous work has shown that MscL expression is upregulated by RpoS [9], and thus deletion of the *rpoS* gene is expected to reduce expression of MscL. In Figure 3.10 we see that this is indeed the case, as expression from the *mscL* promoter is reduced by about two thirds in the *rpoS* deletion strain.

Many genes in *E. coli* exhibit autoregulatory function, and though there is no direct evidence that *mscL* does so, we were thus motivated to measure the effect of deletion of *mscL* on gene expression. However, as seen in Figure 3.10, we found that deletion of *mscL* has no observable effect on expression from the *mscL* promoter. We thus elected not to continue with sort-seq experiments in the Δ*mscL* genetic background.

We also examined transcription from the *mscL* promoter in strain MJF465, in which the *mscS*, *mscK*, and *mscL* genes are deleted. We observed an approximately 4.5 fold increase in expression relative to strain MG1655. One possible explanation lies in the observation that RpoS expression levels are significantly elevated in the MJF465 strain (M. Bialecka and H.J. Lee, unpublished data). Of course, this explanation merely shifts the phenomenon to be explained up a level in the gene regulatory network. It is possible that deletion of all three channels negatively impacts the cell's ability to maintain turgor pressure homeostasis, thereby placing the cell under stress and upregulating RpoS expression. However, the specific pathways by which this could occur are unclear. An important caveat to these MFJ465 results is that the mechanosensitive channel gene deletions in MJF465 are in the context of a different strain background, strain Frag1, than the deletions of *rpoS* and *mscL*, which were performed in strain MG1655. While Frag1 and MG1655 differ only at four genetic loci, none of which seem likely to affect mechanosensitive channel transcriptional regulation, the possibility must nonetheless be admitted [13, 14].

Finally, as seen in Figure 3.10, we performed gene expression assays in each of the varying genetic

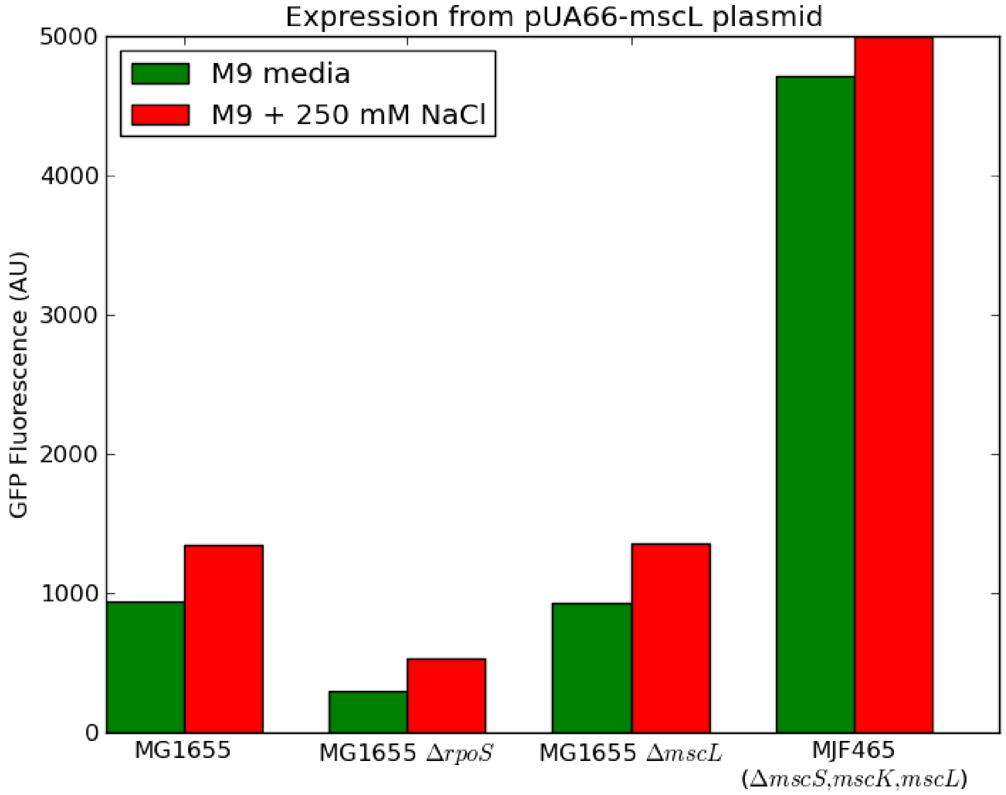
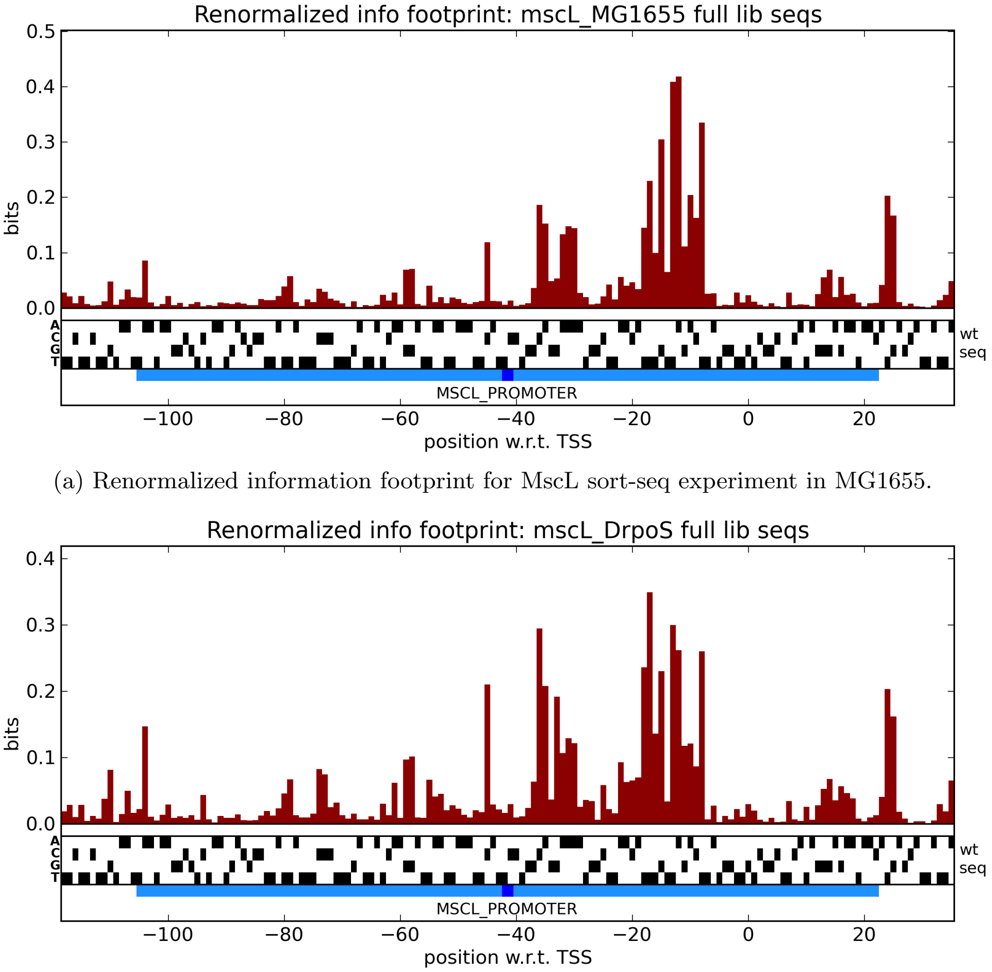


Figure 3.10: **MscL gene expression assays.** GFP expression from the pUA66-mscL plasmid was measured under various genetic backgrounds and environmental conditions. Measurements were obtained using a BD Biosciences LSRII flow cytometer; 50,000 cells were measured for each sample, and the median values are reported on this plot. Cells were grown in either M9 minimal media + 0.5% glucose or M9 minimal media + 0.5% glucose + 250 mM supplemental NacL, as indicated in the legend. See the text for a detailed discussion of these results.

backgrounds in both normal M9 + 0.5% glucose minimal media, and M9 + 0.5% glucose minimal media supplemented with 250 mM NaCl. Previous work by members of the Phillips lab has indicated that expression of MscL is upregulated in cells grown in high salt conditions [11]. We found that growth in supplemental NaCl increased expression from the *mscL* promoter by 43% in MG1655, 82% in MG1655 $\Delta rpoS$, 46% in strain MG1655 $\Delta mscL$, and 6% in strain MJF465. The 82% increase in MG1655 $\Delta rpoS$ is somewhat surprising, since RpoS expression is known to be upregulated in high salt conditions, and thus deletion of *rpoS* might be expected to impair upregulation of MscL expression in high salt conditions [15, 16]. Also of note is the fact that in strain MJF465, differential expression in normal and high salt conditions is almost completely abolished, suggesting again that deletion of *mscL*, *mscS*, and *mscK* introduces some kind of substantial perturbation into *mscL* regulation. Finally, an important caveat to these results is the fact that growth in high salt conditions causes a slight decrease in the growth rate, and thus caution must be exercised in comparing the



(a) Renormalized information footprint for MscL sort-seq experiment in MG1655.

(b) Renormalized information footprint for MscL sort-seq experiment in strain MG1655 $\Delta rpoS$.

results from cells grown in normal and high salt media.

3.2.4 Analysis of putative binding sites.

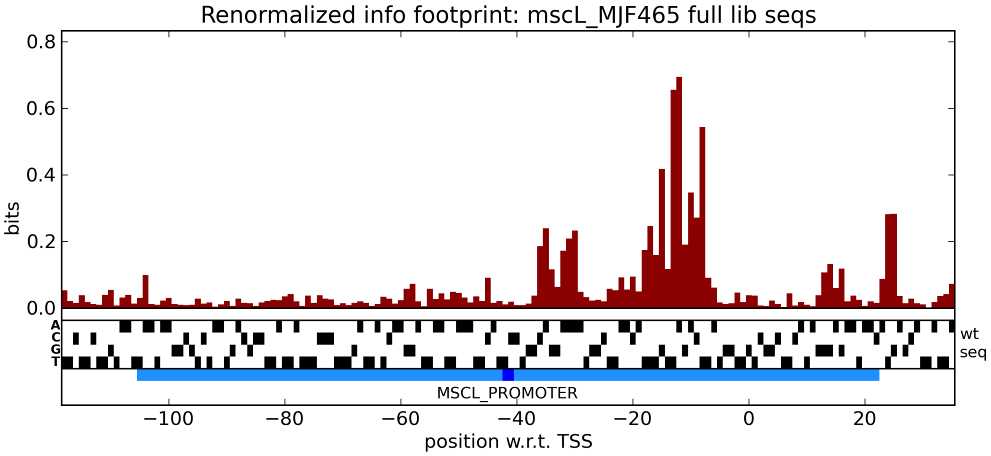
As described above and in equation 3.7, we can construct an "information footprint" from sort-seq data to gain an understanding of where transcription factors and RNAP are binding. We performed sort-seq experiments in the following strain backgrounds and environmental conditions (all strains were grown in M9+0.5% glucose minimal media with supplemental NaCl where indicated):

1. MG1655

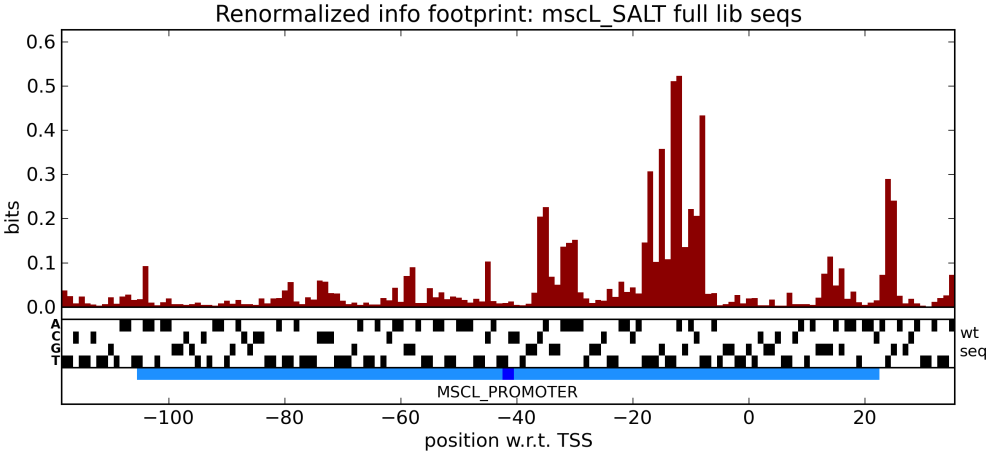
2. MG1655 $\Delta rpoS$

3. MJF465

4. MG1655 grown in M9+0.5% glucose + 250 mM NaCl.



(c) Renormalized information footprint for MscL sort-seq experiment in strain MJF465 ($\Delta mscL, mscS, mscK$).



(d) Renormalized information footprint for MscL sort-seq experiment in strain MG1655, grown in M9 media supplemented by 250 mM NaCl.

Figure 3.11: **Renormalized *mscL* promoter information footprints**. Renormalized information footprints (as defined in equation 3.7) in various strain background and growth conditions. The features of these footprints are described in detail in the main text. Briefly, the RNAP binding site is clearly visible around positions -40 to -20. The ribosomal binding site is visible around positions 15 to 20. Farther upstream, we identify three potential TF binding sites, centered around approximately -65, -75, and -110.

Information footprints computed using equation 3.7 for each of these conditions are shown in Figure 3.11. Before examining the features of these footprints in detail, we will begin by noting the fact that despite the differing conditions between sort-seq experiments, the footprints are remarkably consistent from experiment to experiment. This seems particularly interesting when comparing the cases of background strains MG1655 and MJF465, which as shown above differ by nearly a factor of five in expression from the wild-type *mscL* promoter. If this change in gene expression were mediated by differential binding of some transcription factor in the two strain backgrounds, we would expect the information footprints to be different. For instance, if a repressor were binding the promoter region in MG1655, but this repressor was not binding the promoter in MJF465, the "footprint" of the repressor present in MG1655 would no longer be visible in MJF465. The fact that the overall footprints are so similar for these two strains suggests that the increase in gene expression is mediated by something other than differential transcription factor binding.

Since the information footprints appear qualitatively the same for the four sort-seq experiments, we will focus here on the MG1655 experiment seen in Figure 3.11a. In addition to the RNAP binding site (clearly visible between roughly -40 and -3), several informative regions that look like potential transcription factor binding sites are immediately apparent. This is promising, since as mentioned above, the current state of transcriptional regulatory annotation for the *mscL* gene comprises only the transcription start site. We will start at the downstream end of the mutagenized region and work our way upstream, considering each potential binding site in turn. For each potential binding site, we fit an energy matrix to model the sequence dependent binding energy of the putative TF. It should be noted that the coordinates corresponding to each putative binding site are based on visual inspection of the information footprints in figure 3.11, and are thus unavoidably somewhat subjective.

The first informative region appears in the region [10:29] (coordinates are with respect to the transcription start site at 0). The energy matrix for this region is shown in Figure 3.12. Examination of the energy matrix reveals that this region actually corresponds to the ribosomal binding site and start codon. The optimal binding sequence for positions 12 to 18 can be read off the energy matrix as AGGAGG which corresponds to the canonical Shine-Dalgarno ribosome binding site sequence. The start codon ATG is also clearly visible in the energy matrix at positions 23 to 25. The red matrix elements in positions 24 and 25 mean that any sequence other than TG at those positions is very detrimental for gene expression. The initial A is also clearly optimal, but a G (light blue matrix element) can also be substituted at this position without an excessive penalty. This is consistent with the fact that GTG is an alternative start codon in *E. coli* and is the second most abundant start codon after ATG. Although this informative region does not correspond to a TF binding site, it is encouraging that our analysis recapitulates known biology of translation initiation. However, this analysis also serves to inject a note of caution in that it illustrates that sort-seq information footprints

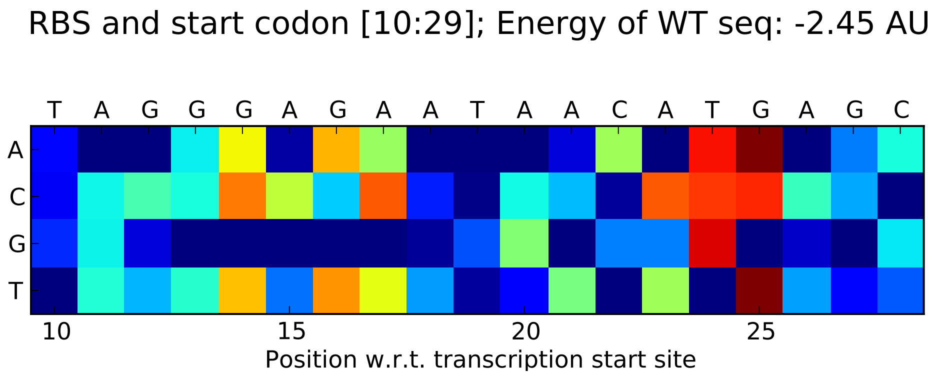
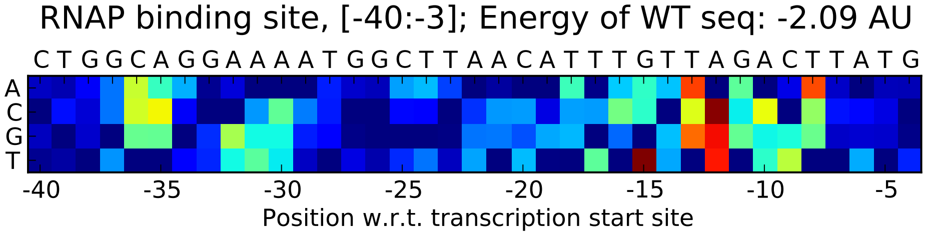


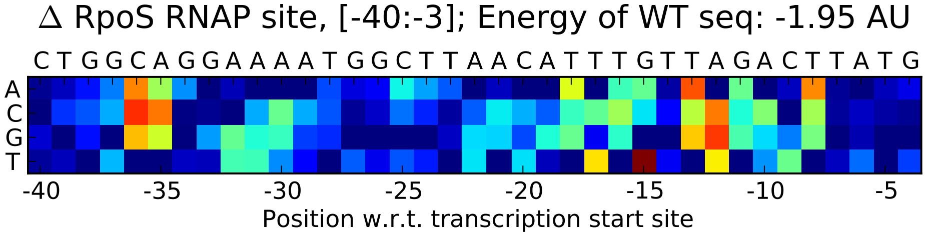
Figure 3.12: Energy matrix for RBS region. The Shine-Dalgarno sequence AGGAGG can be seen as the optimal sequence for positions 12 to 18. The start codon ATG is also clearly visible as the optimal sequence for positions 23 to 25. The alternative start codon GTG is also visible as the second most favorable sequence at positions 23 to 25.

and "energy matrices" will pick up any basepair changes that affect the downstream measurement of gene expression (in this case, GFP expression), without prejudice as to the mechanism (*i.e.*, transcriptional or translational regulation) whereby those changes affect the downstream readout of expression. In a hypothetical sort-seq experiment in which gene expression is measured at the mRNA level, we would not expect to observe these translational effects.

The next informative region is found at the coordinates [-40:-3] and corresponds to the RNAP binding site. We will begin by examining the -10 region. The consensus -10 hexamer sequence for RNAP σ^{S} has been reported as (T/C)ATA(C/A)T; this sequence is very similar to the RNAP σ^{70} (the *E. coli* "housekeeping" sigma factor) consensus sequence TATAAT [16]. In the energy matrix shown in Figure 3.13a, this consensus sequence is clearly visible at positions -13 to -8. In addition, there appears to be at least one σ^{S}-specific promoter element present in the RNAP energy matrix shown in Figure 3.13a: namely, a strong preference for C immediately upstream of the -10 hexamer, at position -14. Becker and Hengge-Aronis have reported that this C is directly contacted by the residue Lys-173 of σ^{S} region 2.5 [17]. Moreover, changing this residue to a glutamate, which is the residue found at the corresponding position in σ^{70}, changes the sequence preference to a G, which is the sequence preference of σ^{70} at this position [17]. These effects are clearly visible in the energy matrices shown in Figure 3.13. In Figure 3.13a, the energy matrix shows a clear preference for C at position -14, while in Figure 3.13b, inferred in the MG1655 Δ *rpoS* strain, there is a slight preference for G at position -14. Finally, the presence of a TG at positions -16 to -15 has been termed the "extended -10" promoter element [18]. This element generally increases the strength of the promoter. It has been proposed that RNAP σ^{70} can recognize the extended -10 element strictly at positions -16 to -15, whereas RNAP σ^{S} can also recognize the element at positions -17 or -18 [16, 19]. In any case we clearly see the extended -10 element at position -16 in both energy matrices in Figure 3.13, so this cannot be considered a σ^{S}-specific promoter element.



(a) Inferred binding energy matrix for RNAP in strain MG1655.



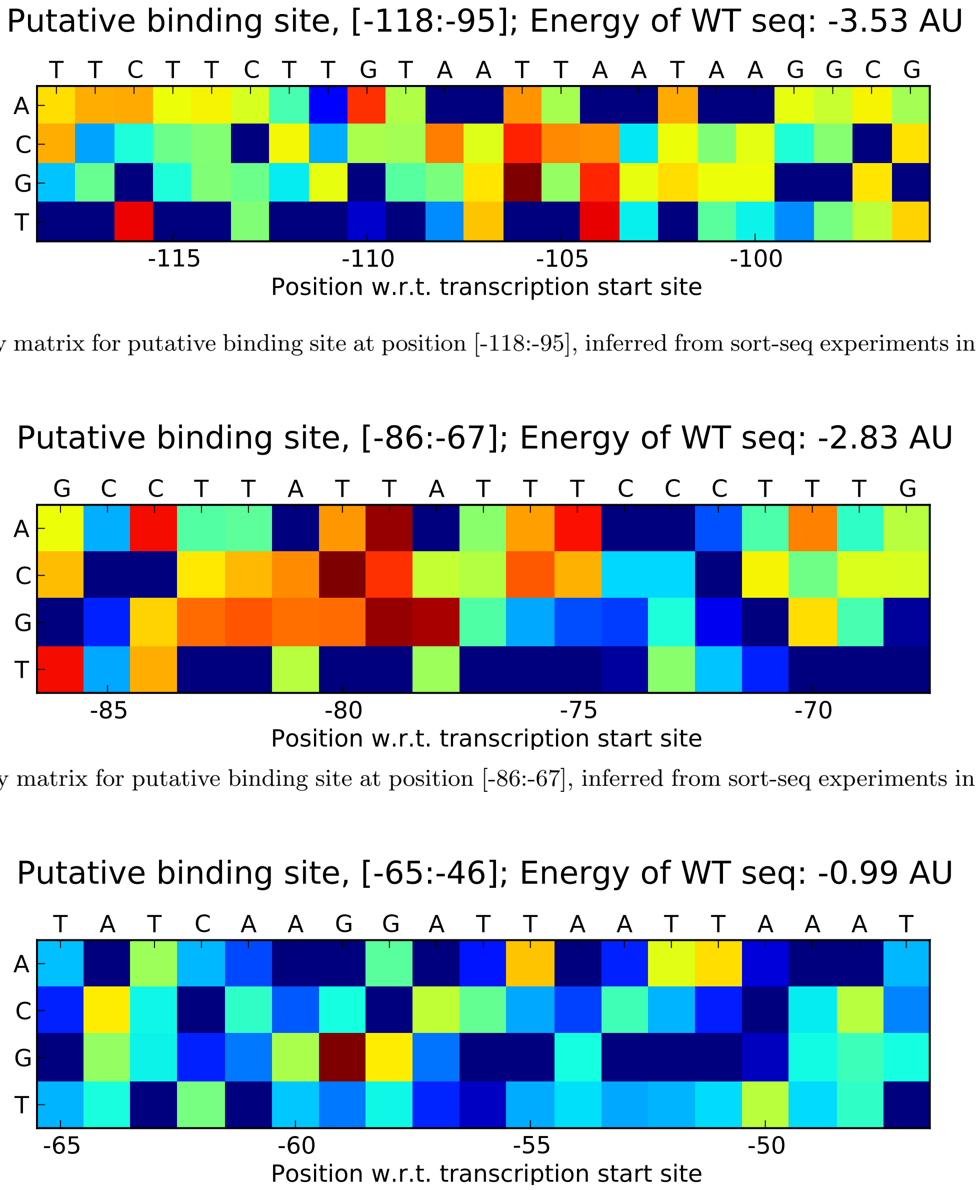
(b) Inferred binding energy matrix for RNAP in strain MG1655 Δ*rpoS*.

Figure 3.13: RNAP binding energy matrices in MG1655 and MG1655 Δ *rpoS* strain backgrounds. Cooler (blue) colors indicate favorable binding interactions, while hotter (red) colors indicate unfavorable interactions. The -35 hexamer corresponds to positions -36 to -31, and the -10 hexamer corresponds to positions -13 to -8.

The -35 region exhibits less sequence specificity than the -10 region, as evidenced graphically by the generally cooler colors in the heat map in this region. The optimal binding sequence for the -35 region is given by the matrix as TTGCCA, which is nearly same as the -35 consensus sequence TTGACA for σ^{70}. There is considerably variability in the -35 regions of σ^{S} promoters [20], so it is somewhat difficult to interpret the observed optimal sequence of TTGCCA in light of σ^{S} sequence specificity, but this variability in σ^{S} -35 recognition sequence does seem to be consistent with the overall reduced specificity in the -35 region of the energy matrix compared with the -10 region. In conclusion, the RNAP binding energy matrix seems broadly consistent with transcription by both σ^{70} and σ^{S}, in agreement with reports that transcription from the same promoter can be initiated by either sigma factor [21, 22].

Based on the information footprint, we identity three additional putative transcription factor binding sites, found at the following coordinates: [-65:-46], [-86:-67], and [-118:-95]. The strength of the "signature" of these binding sites in the information footprint is less than for the RNAP binding site and RBS, and the choice of starting and ending coordinates for each of these sites is inevitably somewhat arbitrary. Moreover, the optimized mutual information values for these three binding sites are substantially lower: 0.01, 0.02, and 0.02 bits, respectively, vs 0.14 bits for the RNAP binding site. This means that the energy matrices for these binding sites are about tenfold less informative about gene expression than the energy matrix for the RNAP binding sites. So it will not be possible to analyze these energy matrices in the same level of detail as the RNAP binding site, not least because we have no idea which proteins (if any) are actually binding there. Nonetheless, we will attempt to provide what insight we can, and will offer some speculation as to the possible identities of the proteins in play.

The energy matrix for the binding site in the region [-65:-46] is shown in Figure 3.14c. We identify this binding site as a repressor, based on the fact that across the sort-seq dataset the binding energies of this sequence are positively correlated with gene expression. (In other words, lower (better) binding energies are associated with lower levels of gene expression, and vice versa; which is what one would expect for a repressor binding site). Interestingly, the sequence of this region in the wild-type *mscL* promoter differs from the predicted optimal binding sequence (according to the energy matrix) at 10 of 19 positions, suggesting that this is not a particularly strong binding site. Along these lines, it is interesting to note that the position with the largest difference by a substantial margin between best and worst base identity is found at -59, and the wild-type sequence indeed contains the worst possible base pair, a G. Pull-down assays by graduate student Nathan Belliveau of the Phillips lab have suggested that the glycine cleavage system transcriptional activator GcvA binds to this region of DNA. However, more work needs to be done to verify whether this is the case. To that end, efforts are currently underway to perform sort-seq experiments on the *mscL* promoter in a strain in which *gcvA* is deleted; if GcvA is indeed binding to this region, we would



(a) Energy matrix for putative binding site at position [-118:-95], inferred from sort-seq experiments in strain MG1655.

(b) Energy matrix for putative binding site at position [-86:-67], inferred from sort-seq experiments in strain MG1655.

(c) Energy matrix for putative binding site at position [-65:-46], inferred from sort-seq experiments in strain MG1655.

Figure 3.14: **Energy matrices for putative binding sites.** See section 3.1.3 for details of model fitting procedures. Putative binding sites were identified based on analysis of the information footprint shown in Figure 3.11a. Blue colors indicate favorable binding, while red colors indicate unfavorable binding.

expect the region of high information in the information footprint to disappear, as in Figure 3.2. Pull-down assays also indicate significant enrichment of the histone-like protein H-NS in this region.

The energy matrix for the binding site in the region [-86:-67] is shown in Figure 3.14b. Again, based on the positive correlation between binding energies and gene expression across the sort-seq dataset, we identify this as a repressor binding site. The wild-type binding sequence differs from the predicted optimal sequence at 4 of 19 positions (-74,-73,-71, and -68), although a ‘G’ at position -68 is virtually the same as the optimal ‘T’. Pull-down assays indicate enhanced binding of the probable helix-turn-helix (HTH) type transcriptional regulator LrhA in this region (N. Belliveau, unpublished data). An alternative approach to determining TF identity is to compare the binding energy matrix with known transcription factor binding motifs. A software program called TOMTOM has been developed to perform precisely this function [23]. Comparing the binding energy matrix with known TF binding motifs from the RegulonDB database, we find that the transcription factor *fis* is the top hit. A comparison between the two binding motifs is shown in Figure 3.15. By eye, the agreement looks reasonably convincing, although on the other hand Fis was not identified in the pull-down binding assays. In any case, it is clear that more work is required to unearth the identity of this putative binding site.

Finally, the energy matrix for the binding site in the region [-118:-95] is shown in Figure 3.14a. This binding site is also identified as a repressor. The wild-type sequence differs from the inferred optimal sequence at only one position (-116) out of 23. This fact is curious in light of the fact that wild-type transcription factor binding sites rarely (if every) correspond exactly to the optimal or “consensus” sequence for the particular TF in play. Another interesting feature of the matrix is the preponderance of ‘A’s and ‘T’s in the optimal binding sequence: the optimal sequence is 65% AT, or 70% if the optimal base at position 110 is taken to be ‘T’. (The same is true, incidentally, of the optimal sequence for the region [-86:-67], which has 74% AT content). Pull-down assays indicate enhanced binding of H-NS and StpA to this region, both of which are known to bind relatively non-specifically to AT rich regions.

Much work remains to be done in identifying the molecular players involved in transcriptional regulation of *mscL*. Nonetheless, we find that our analysis successfully reproduces known aspects of *mscL* regulation, including the role of RpoS in promoting transcription of *mscL*. It also seems highly likely that H-NS and/or StpA are involved, on the basis of both pull-down mass spectroscopy analysis, and the fact that both the observed wild-type promoter sequence and the inferred optimal binding sequences are AT rich for the binding sites identified at [-118:-95] and [-86:-67]. Moreover, though the inferred energy matrix for the region [-65:-46] shows an optimal sequence with only 53% (10/19) AT content, the wild-type sequence for this region has a whopping 84% AT content. A role for H-NS is also corroborated by CHIP-Seq data that indicates H-NS binding in the *mscL* promoter region [24], as well as the fact that H-NS plays a role in regulation of many RpoS-dependent

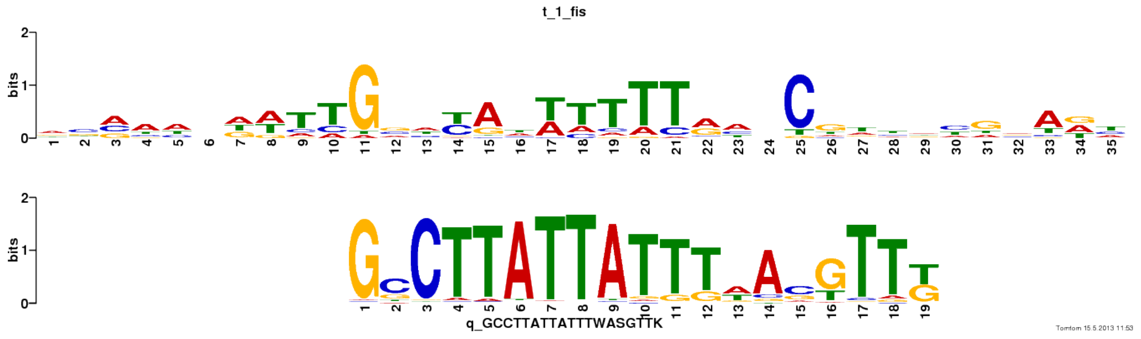


Figure 3.15: **Output of binding motif comparison tool.** TOMTOM output comparing inferred binding energy matrix for the region [-86:67] (bottom) with known Fis binding motif. TOMTOM is a software program that compares a particular binding motif with a database of motifs to search for the best match. There appears to be reasonably good similarity between the motifs, but not good enough to draw definitive conclusions in the absence of additional corroboration.

genes [25], as well in regulation of RpoS itself [15]. Efforts are currently underway to perform sort-seq experiments in strains in which the various candidate TFs identified above are knocked out, in order to identify whether knocking out those particular genes abolishes any of the putative binding sites identified here. However, it also seems worth noting that the putative TF binding sites appear to have a substantially smaller effect on gene expression than the RNAP binding site. This can be seen in the information footprints (Figure 3.11) where the mutual information values in the RNAP binding region are higher than in the upstream binding sites, and can also be seen in the fact that the informativeness of the RNAP energy matrix is about tenfold higher than the informativeness of the putative TF energy matrices (0.14 bits vs 0.01 or 0.02 bits). This disproportionate importance of the RNAP binding site is borne out in the next section.

3.2.5 Testing the energy matrices using designed variants of the *mscL* promoter.

Finally, although we do not know the identities of the molecular players, we do know whether the putative binding sites increase or decrease gene expression. We can thus use the inferred binding energy matrices to design sequences predicted to increase or decrease gene expression. In Figure 3.16, we show measured gene expression from the wild type promoter, from a promoter in which the three unknown putative binding sites have been mutated to enhance binding, and a promoter in which the RNAP binding site has been enhanced in addition to the putative binding sites. We find that enhancing the unknown repressor binding sites decreases gene expression by 30%. Enhancing the RNAP binding increases gene expression by approximately 10 fold. These results are consistent with the results in the previous section that the unknown binding sites are associated with repression of gene expression, yet their effect on gene expression seems to be substantially less than the RNAP binding site. (Data courtesy of Nathan Belliveau.)

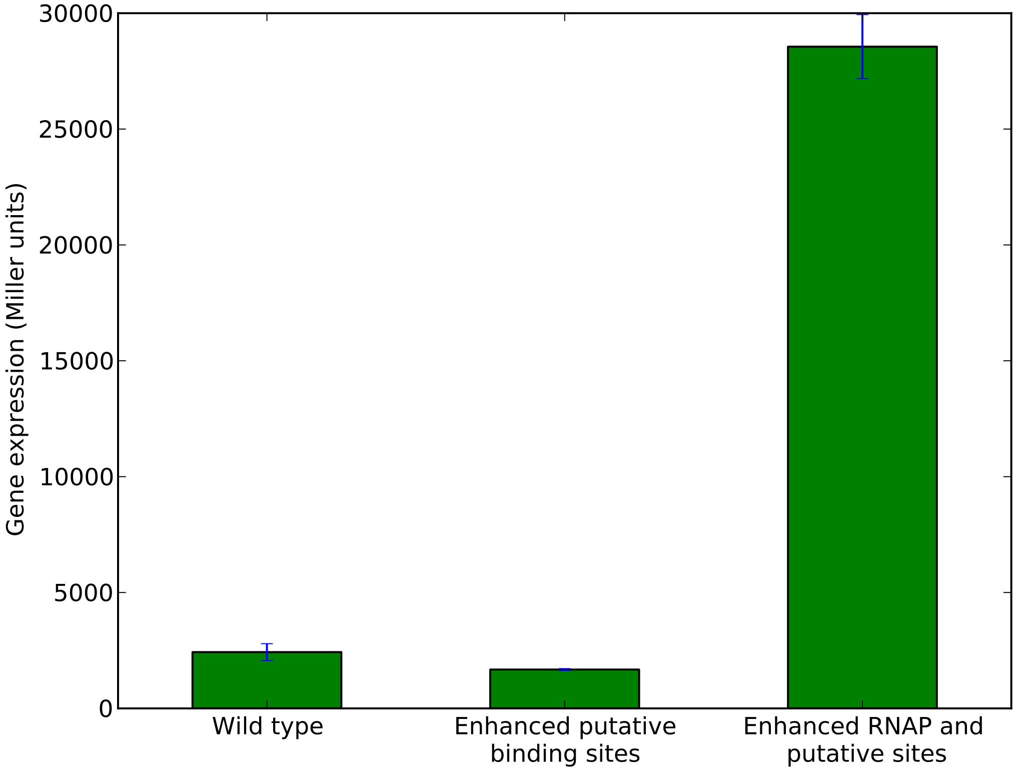


Figure 3.16: **Gene expression from "designed" mscL promoters.** We used the energy matrices for putative TF binding sites and for RNAP to design promoter variants. Although we do not know the identities of the putative TFs, we do know that they are predicted to function as repressors. As seen in the plot (center column), increasing the strength of these binding sites decreases gene expression by 20%. However, if in addition the RNAP binding site is optimized, gene expression increases by about tenfold. This result agrees qualitatively with the idea that the putative TF binding sites have a smaller effect on gene expression than the RNAP binding site.

Bibliography

[1] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20):9158–63, May 2010.

[2] A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. G. Surette, and U. Alon. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, 3(8):623–8, 2006.

[3] Rob Phillips, Jane Kondev, Julie Theriot, and Hernan Garcia. *Physical Biology of the Cell*. Garland Science, New York, 2 edition.

[4] Alessandro Treves and Stefano Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, 1995.

[5] Justin B. Kinney, Gaper Tkaik, and Curtis G. Callan. Precise physical models of proteindna interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 2007.

[6] Radford M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto, Ontario, 1993.

[7] Ian R. Booth, Michelle D. Edwards, Susan Black, Ulrike Schumann, and Samantha Miller. Mechanosensitive channels in bacteria: Signs of closure? *Nature Reviews Microbiology*, 5(6):431–440, Jun 2007.

[8] Natalia Levina, Sabine Ttemeyer, Neil R. Stokes, Petra Louis, Michael A. Jones, and Ian R. Booth. Protection of *Escherichia coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: Identification of genes required for MscS activity. *The EMBO Journal*, 18(7):1730–1737, 1999.

[9] Neil R. Stokes, Heath D. Murray, Chandrasekaran Subramaniam, Richard L. Gourse, Petra Louis, Wendy Bartlett, Samantha Miller, and Ian R. Booth. A role for mechanosensitive channels in survival of stationary phase: Regulation of channel expression by RpoS. *Proceedings of the National Academy of Sciences*, 100(26):15959–15964, 2003.

[10] Ching Kung, Boris Martinac, and Sergei Sukharev. Mechanosensitive channels in microbes. *Annual Review of Microbiology*, 64(1):313–329, 2010. PMID: 20825352.

[11] Maja Bialecka-Fornal, Heun Jin Lee, Hannah A DeBerg, Chris S Gandhi, and Rob Phillips. Single-cell census of mechanosensitive channels in living bacteria. *PloS one*, 7(3):e33077, 2012.

[12] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muiz-Rascado, Jair S. Garca-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martnez-Flores, Alejandra Medina-Rivera, Gerardo Salgado-Osorio, Shirley Alquicira-Hernndez, Kevin Alquicira-Hernndez, Alejandra Lpez-Fuentes, Liliana Porrn-Sotelo, Araceli M. Huerta, Csar Bonavides-Martnez, Yalbi I. Balderas-Martnez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Vernica Jimnez-Jacinto, Leticia Vega-Alvarado, Victor del Moral-Chvez, Alfredo Hernndez-Alvarez, Enrique Morett, and Julio Collado-Vides. Regulondb v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[13] Wolfgang Epstein and Byung S Kim. Potassium transport loci in escherichia coli k-12. *Journal of Bacteriology*, 108(2):639–644, 1971.

[14] Frederick R. Blattner, Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, 1997.

[15] Regine Hengge-Aronis. Back to log phase: σs as a global regulator in the osmotic control of gene expression in *Escherichia coli*. *Molecular Microbiology*, 21(5):887–893, 1996.

[16] S Lacour and P Landini. σS-dependent gene expression at the onset of stationary phase in Escherichia coli: Function of σS-dependent genes and identification of their promoter sequences. *Journal of Bacteriology*, 186(21):7186–7195, 2004.

[17] Gisela Becker and Regine Hengge-Aronis. What makes an *Escherichia coli* promoter σS dependent? role of the 13/14 nucleotide promoter positions and region 2.5 of σS. *Molecular Microbiology*, 39(5):1153–1165, 2001.

[18] Jennie E. Mitchell, Dongling Zheng, Stephen J. W. Busby, and Stephen D. Minchin. Identification and analysis of "extended -10" promoters in *Escherichia coli*. *Nucleic Acids Research*, 31(16):4689–4695, 2003.

[19] Stephan Lacour, Annie Kolb, and Paolo Landini. Nucleotides from -16 to -12 determine specific promoter recognition by bacterial σS-RNA polymerase. *The Journal of Biological Chemistry*, 278(39):37160–8, September 2003.

[20] A Wise, R Brems, V Ramakrishnan, and M Villarejo. Sequences in the -35 region of *Escherichia coli* rpos-dependent genes promote transcription by EσS. *Journal of Bacteriology*, 178(10):2785–93, 1996.

[21] O L Lomovskaya, J P Kidwell, and a Matin. Characterization of the sigma 38-dependent expression of a core Escherichia coli starvation gene, pexB. *Journal of bacteriology*, 176(13):3928–35, July 1994.

[22] K Tanaka, Y Takayanagi, N Fujita, A Ishihama, and H Takahashi. Heterogeneity of the principal sigma factor in Escherichia coli: the rpoS gene product, sigma 38, is a second principal sigma factor of RNA polymerase in stationary-phase Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):3511–3515, September 1993.

[23] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biol*, 8(2):R24, 2007.

[24] Christina Kahramanoglou, Aswin S. N. Seshasayee, Ana I. Prieto, David Ibberson, Sabine Schmidt, Jurgen Zimmermann, Vladimir Benes, Gillian M. Fraser, and Nicholas M. Luscombe. Direct and indirect effects of h-ns and fis on global gene expression control in escherichia coli. *Nucleic Acids Research*, 39(6):2073–2091, 2011.

[25] M Barth, C Marschall, A Muffler, D Fischer, and R Hengge-Aronis. Role for the histone-like protein h-ns in growth phase-dependent and osmotic regulation of sigma s and many sigma s-dependent genes in escherichia coli. *Journal of Bacteriology*, 177(12):3455–64, 1995.

Chapter 4

Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*

A version of this chapter originally appeared as: Brewster RC*, Jones DL*, Phillips R (2012) Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. PLoS Comput Biol 8(12): e1002811. doi:10.1371/journal.pcbi.100281

4.1 Introduction

The regulation of gene expression is one of the primary ways that cells respond to their environments. The quantitative dissection of the networks that control such expression as well as the construction of designed networks has been a central preoccupation of regulatory biology. As sketched in Figure 4.1, the level of gene expression exhibited by a cell can be targeted at multiple levels along the path from DNA to protein. Key biological tuning variables include the copy number of the transcription factors that act on a gene of interest, the strength of their binding sites, the strength of RNA polymerase binding, the strength of ribosomal binding sites and the degradation rates of the protein products of the gene of interest. Many of these tuning parameters have been studied in quantitative detail. For instance, Salis *et al.* [1] developed a model to describe the interaction energy between the ribosomal binding site (RBS) of an mRNA transcript and the 30S ribosomal subunit, which they relate to translation initiation rate using statistical thermodynamics. Using this model, gene expression can be predictively tuned over five orders of magnitude by modulating translation efficiency for a given gene [1, 2]. Translation initiation (and hence protein expression) is thus tuned by choosing an RBS sequence with the desired interaction energy. The rate of protein degradation is another key determinant of intracellular protein concentration. Protein degradation can be modulated by the use of degradation tags appended to the C-terminal domain of a given protein. The ssrA tag [3], for instance, targets proteins for destruction by the *E. coli* degradation machinery, which includes

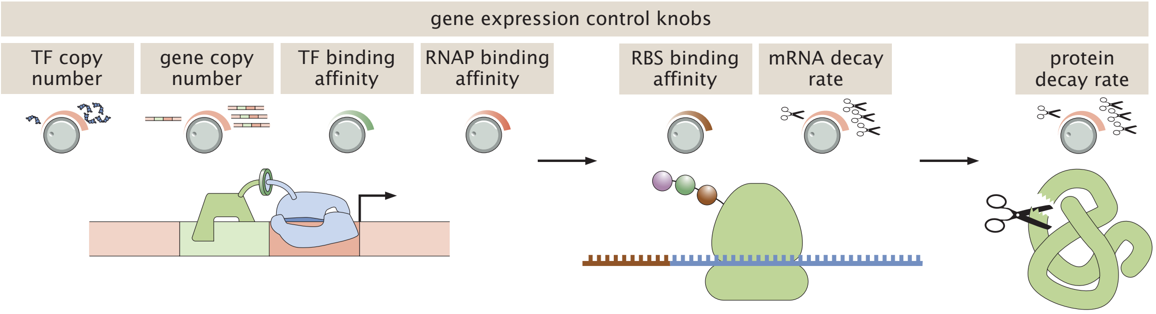


Figure 4.1: **Regulatory control knobs.** A schematic view of the available knobs which can be systematically tuned to change the mRNA and protein distributions. In this work we begin by studying constitutive expression, eliminating the extra layer of complexity associated with transcription factors, and systematically control the RNAP binding affinity through control of the promoter sequence. These results are then generalized to the case in which these same promoters are subjected to regulation by repressor binding, with the level of repressor (i.e. TF copy number) controlled systematically.

proteases ClpXP, ClpAP and SspB [4]. This degradation system has been artificially implemented in yeast, where ClpXP is expressed from an inducible promoter, and degradation rates of ssrA-tagged proteins can be tuned over a factor of ≈ five by controlling the ClpXP concentration in the cell [5]. Similarly, manipulating the decay rate of the protein's transcript allows for modulation of the steady-state protein copy number [6, 7].

In this paper, we focus on two sets of these transcriptional parameters: namely, the strength with which polymerase binds the promoter, and the number of transcription factors present when that promoter is controlled by simple repression. We begin by focusing on the simplest case where there are no repressor proteins present in the cell. Our interest in such "constitutive" promoters (those not regulated by transcription factors) stems from the goal of creating a set of promoters in which we can systematically vary both the mean and the noise to test recent models of transcriptional kinetics [8]. These experiments are further motivated by measurements which question our understanding of how the mean and noise in transcription depend on the architecture of the promoter [9]. To test these ideas on noise in transcription, we must know how to predictively tune the binding strength of RNAP to the promoter.

Precise physical modeling of protein-DNA interaction energies is a difficult problem involving many degrees of freedom. Such binding energies are at the heart of the molecular interactions which result in (or, in the case of repressor transcription factors, prevent) transcription events. Hence, precise control of protein-DNA binding is an essential prerequisite for quantitative control of transcription. Despite the complexity of protein-DNA interactions and numerous molecular mechanisms involved in transcription initiation [10–14], simple linear models of sequence-dependent binding energies are often sufficient to describe the interactions of transcription factors (TFs) or RNAP with DNA [15–20]. A "linear model" treats each base along the binding site as independently contributing a defined amount to the total binding energy. The total binding energy is then the sum of the contributions from each base along the binding site. In one recent study, the authors inferred the 4×41 parameters describing the interaction of RNAP σ^{70} holoenzyme with DNA [20]. This matrix is shown pictorially in Figure 4.2 and the numerical values are provided in Supplementary Information (SI) Text S2. Mathematically, the binding energy of RNAP to a specific sequence is calculated using a matrix $M_{i,j}$ of 4×41 energy values where i represents the base identity (A,C,T,G), and j represents the base pair position along the binding site. For instance, $M_{2,8}$ represents the contribution from having a C present at position 8 along the binding site. We represent a particular promoter sequence by a 41×4 matrix $S_{j,i}$ which is unity if the j^{th} base pair has identity i and zero otherwise. The total energy of the sequence in question is the inner product of these matrices, namely,

$$E(S) = \sum_{ij} M_{i,j} S_{j,i}. \tag{4.1}$$

For convenience, we have added a constant offset to the matrix such that the average value of $E(S)$ across the *E. coli* genome is zero (see SI Text S1 for the original matrix from reference [20], SI Text S2 for the adapted matrix, and SI Text S3 for the Python source code to perform the adaptation). Since only differences in energy (such as between two different promoter sequences) are physically meaningful, we can add the same constant value to each element of the matrix without affecting its physical interpretation.

We use this correspondence between promoter sequence and RNAP binding affinity to generate a suite of promoters with a wide range of binding affinities. We then show how a simple thermodynamic model of transcription, which postulates that transcriptional activity is proportional to the probability of finding the RNAP bound at the promoter, accurately predicts the scaling of the expression with RNAP binding energy. In addition, these measurements allow us to determine the proportionality between RNAP binding probability and transcriptional output for our gene. With this information, we can make absolute predictions for the transcriptional output of our designed promoters under other regulatory conditions. We test and confirm these predictions by measuring the transcriptional output of some of our promoters in the architectural context of simple repression (similar to reference [2]) and show we are able to make accurate, absolute predictions of the transcription as a function of average repressor copy number.

4.2 Results

We set out to design sets of unique RNAP sites with specific binding energies separated by $\approx 0.5\ k_BT$ steps. Taking as a starting point the wild-type *lac* and *lac*UV5 promoters, we used the RNAP binding energy model in Figure 4.2 to choose appropriate base pair mutations (concentrated in the

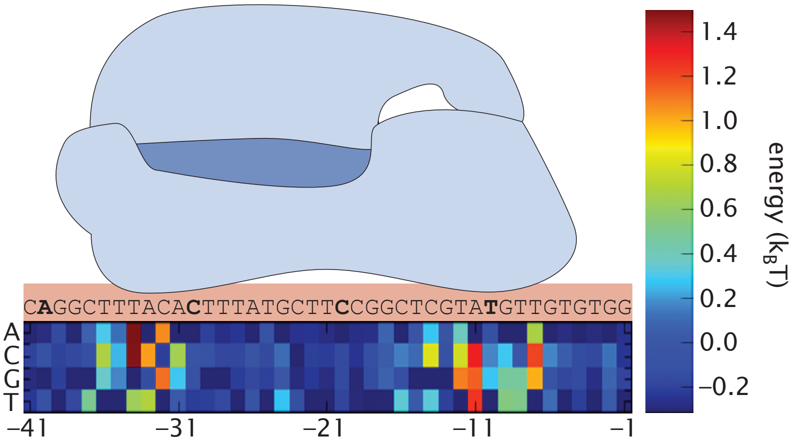
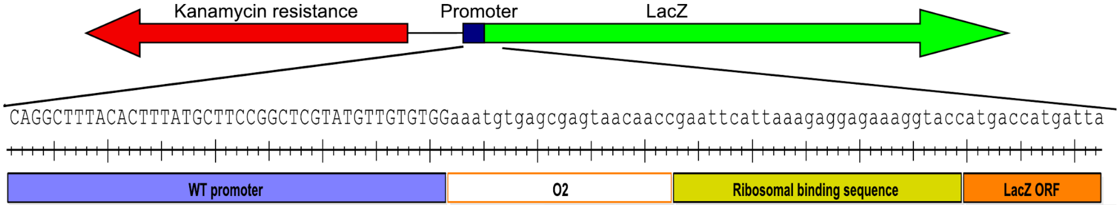


Figure 4.2: **Energy matrix for RNAP binding.** Figure adapted from Kinney *et al* [20]. The contribution of each basepair to the total binding energy is represented by color. The total binding energy of a particular sequence can be calculated by summing the contribution from each base pair. Positive values indicate disfavorable contributions to binding energy. As expected, the most influential base pairs are those in the −10 and −35 region which interact directly with the binding domains of RNAP σ^{70}. Numeric matrix entries are available in SI Text S2. The sequence displayed above the energy matrix corresponds to the wild-type *lac* promoter; the bold bases mark 10 base pair increments. x-axis coordinates are with respect to the transcription start site.

-10 and -35 boxes, where mutations carry the most weight) which result in our desired energy levels. The 18 strains designed by this process have binding energies spanning roughly 6 k_BT and levels of constitutive gene expression from roughly 50 times less to 10 times greater than that of the wild-type *lac* operon. The specific sequences of these 18 promoters are listed in the table shown in Figure 4.3 along with their predicted "model" RNAP binding energy for that sequence. Four promoters are marked with a colored dot; this color coding will be preserved throughout every figure. While the *lac*O2 site is present in our reporter construct, the strain used to measure constitutive expression does not produce LacI, the repressor which specifically binds to this site (see Methods). In addition, the CRP binding site which would otherwise serve to activate the *lac* promoter has been removed. Based on intuition from thermodynamic models of transcription regulation [21–25], we expect that the expression level of a given promoter will scale with the probability that RNAP is bound at that promoter. A derivation of this probability as a function of RNAP binding energy for our promoter architecture is shown below. To test the predictive power of our design process in conjunction with the thermodynamic model, we used single-cell mRNA fluorescence in-situ hybridization (mRNA FISH) and a colorimetric enzymatic assay to measure, for each construct, the average mRNA and protein copy number per cell of LacZ reporter. We then compared these results with those predicted by the calculated RNAP binding energy of that promoter. Finally, we use this same strategy to examine simple repression in the context of our designed promoters.


Kanamycin resistance
Promoter
LacZ
CAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGaaatgtgagcgagtaacaaccgaattcattaaagaggagaaaggtaccatgaccatgatta
WT promoter
O2
Ribosomal binding sequence
LacZ ORF

Promoter sequence (-41 through -1)	$-\Delta\epsilon(k_BT)$			Promoter sequence (-41 through -1)	$-\Delta\epsilon(k_BT)$		
	model	LacZ	mRNA		model	LacZ	mRNA
● TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG	6.99	7.93	7.53	TCGAGTTAAGACTTTATGCTTCCGGCTCGTATAATGTGTGG	4.79	5.05	4.86
TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGGGTGG	6.87	6.69	7.55	CAGGCATTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG	4.73	3.88	3.80
TCGCGTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG	6.86	7.45	7.10	TCGAGTTTACACTTTATGCTTCCGGCTCGGATAACGTGTGG	4.06	3.40	3.22
TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGGGAGG	6.84	7.93	7.53	TCGAGCTCAGACTTTATGCTTCCGGCTCGTATAATGTGTGG	3.81	4.28	4.23
● TCGAGATTACACTTTATGCTTCCGGCTCGTATAATGTGTGG	6.37	5.73	6.15	CAGGCTTAAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG	3.15	1.44	2.82
TCGAGTTTACACTTTATGCTTCCGGCTCGAATAATGTGTGG	6.29	3.93	3.99	TCGAGTTTACACTTTATGCTTCCGGCTCGGTTAAAGTGTGG	3.03	3.84	3.66
TCGTGTTTACCCTTTATGCTTCCGGCTCGTATAATGTGTGG	5.89	6.79	6.50	CAGGCCTTAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG	3.02	3.19	2.97
● TCGAGTTTACACTTTATGCTTCCGGCTCGGATAATGTGTGG	5.58	4.53	4.18	CAGGCCTCAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG	2.16	2.70	2.68
● CAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG	5.35	5.19	5.25	CAGGCTTTACACTTTATGCTTCCGGCTCGGTTGTAGTGTGG	1.39	3.24	3.24
-35 -10				-35 -10			

Figure 4.3: **Schematic of DNA construct inserted in the *galK* region.** The area between the promoter and the LacZ start codon is shown in more detail below along with a table displaying the specific RNAP binding sites (promoters) listed in order of descending binding affinity. The wild-type binding sequence is shown in red text, the *lac*UV5 sequence is shown in magenta text, and two additional promoters are marked by blue text and green text. The data points involving these four promoters will maintain this color coding throughout every figure. The −35 and −10 RNAP recognition sequences are highlighted in a green box and a red box, respectively. The bases in these regions carry the most weight in the energy matrix. Sequences are available in text format in SI Text S4.

4.2.1 Thermodynamic model for constitutive expression

To construct promoters with a targeted level of gene expression, we compute the RNAP binding probability using a simple thermodynamic model based upon the RNAP binding energy matrix from the work of Kinney *et al* [20] (shown in Figure 4.2). A schematic of the allowed microscopic states of the promoter in the constitutive expression system, along with their thermodynamic weights, is shown in Figure 4.4. This model treats all non-specific binding sites (i.e., binding sites other than the promoter of interest) as binding RNAP with a fixed energy ϵ_{NS}. More nuanced treatments of the non-specific background can be found in Refs. [19, 26, 27], for example. Consider a cell with P RNAP molecules which can bind non-specifically with energy ϵ_{NS} to N_{NS} non-specific RNAP binding sites and with energy ϵ_S to the promoter of interest [21–25]. The energy of the state in which the promoter is unoccupied is $P\epsilon_{NS}$ which can occur in $\frac{N_{NS}!}{P!(N_{NS}-P)!}$ unique configurations. Similarly, the energy of the state in which RNAP is specifically bound is given by $\epsilon_S + (P-1)\epsilon_{NS}$, and its multiplicity is given by $\frac{N_{NS}!}{(P-1)!(N_{NS}-P-1)!}$. The probability that RNAP is bound is the Boltzmann

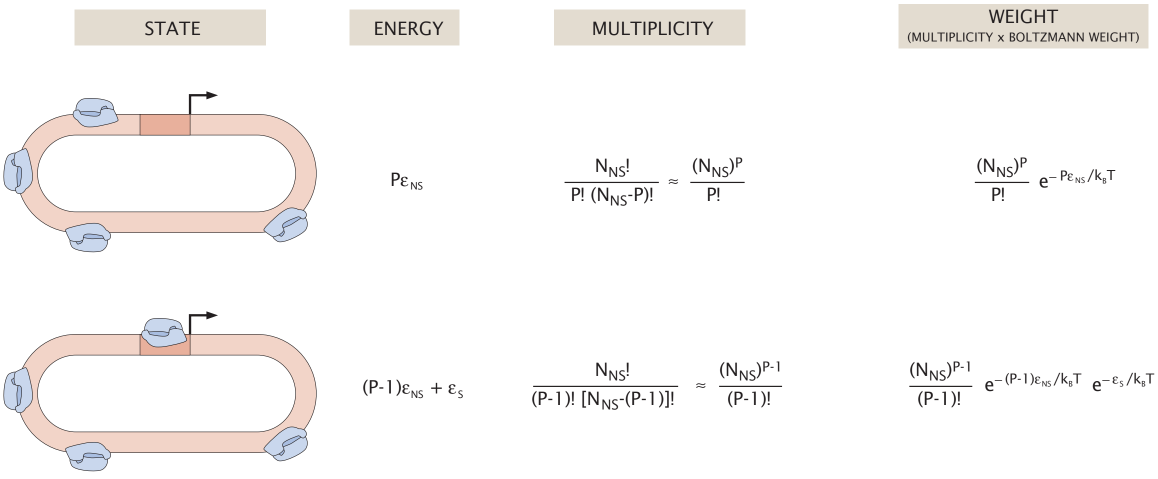

STATE
ENERGY
MULTIPLICITY
WEIGHT
(MULTIPLICITY x BOLTZMANN WEIGHT)
$P\varepsilon_{NS}$
$\frac{N_{NS}!}{P!\,(N_{NS}-P)!} \approx \frac{(N_{NS})^P}{P!}$
$\frac{(N_{NS})^P}{P!}\, e^{-P\varepsilon_{NS}/k_BT}$
$(P-1)\varepsilon_{NS} + \varepsilon_S$
$\frac{N_{NS}!}{(P-1)!\,[N_{NS}-(P-1)]!} \approx \frac{(N_{NS})^{P-1}}{(P-1)!}$
$\frac{(N_{NS})^{P-1}}{(P-1)!}\, e^{-(P-1)\varepsilon_{NS}/k_BT}\, e^{-\varepsilon_S/k_BT}$

Figure 4.4: **States and weights of the unregulated promoter.** In the thermodynamic model, the promoter can be in one of two configurations: unoccupied by RNA polymerase (top) or occupied by RNA polymerase (bottom). The remaining polymerases are bound nonspecifically on the *E. coli* genome. The total energy is the sum of all the nonspecific binding energies and the specific energy of binding at the promoter (when occupied). The multiplicity factor accounts for the number of different ways of arranging polymerases on the genome.

factor of the bound state normalized by the partition function of the system, which simplifies to

$$P_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\Delta\epsilon/k_BT}}{1 + \frac{P}{N_{NS}} e^{-\Delta\epsilon/k_BT}}, \tag{4.2}$$

where $\Delta\epsilon = (\epsilon_S - \epsilon_{NS})$ and where we have used the fact that $\frac{N_{NS}!}{(N_{NS}-P)!} \approx N_{NS}^P$ for $N_{NS} >> P$. In the simplifying case of a "weak promoter", where $\frac{P}{N_{NS}} e^{-\Delta\epsilon/k_BT} << 1$, this expression reduces to

$$P_{\text{bound}} = \frac{P}{N_{NS}} e^{-\Delta\epsilon/k_BT}. \tag{4.3}$$

Note that the microscopic language used to make these derivations is convenient for interpreting binding energies and the dependence on number of polymerases. However, all of these results can be naturally derived and written in the alternative language of dissociation constants without ever making reference to the nonspecific background [23]. For example, we can write

$$P_{\text{bound}} = \frac{\frac{[P]}{K_d}}{1 + \frac{[P]}{K_d}}, \tag{4.4}$$

where K_d is the *in vivo* dissociation constant for RNAP from the promoter of interest.

With these results, we can now explore the connection between the measured and the corresponding predicted level of expression. Since gene expression is (by assumption) proportional to

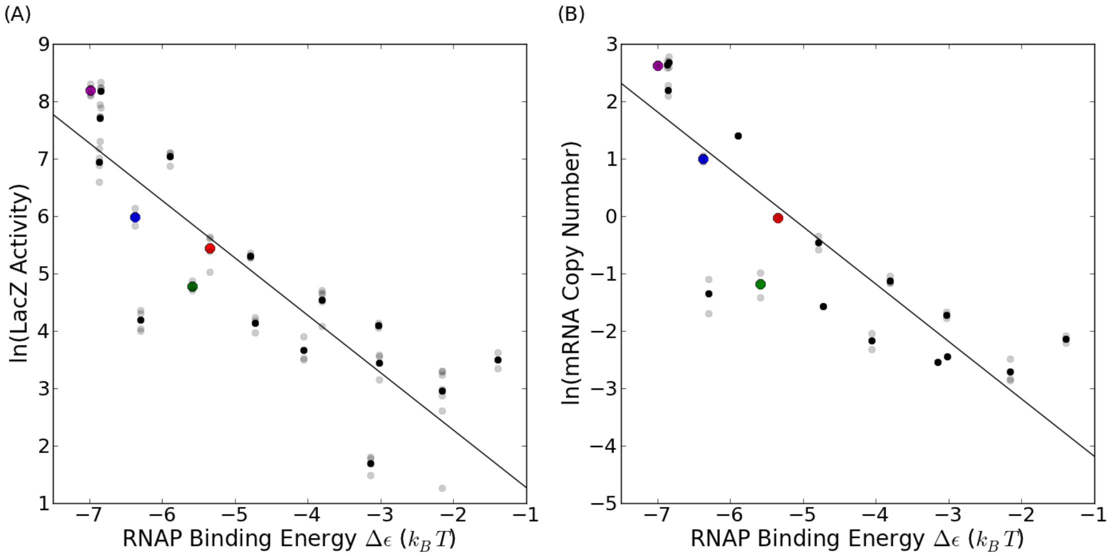


Figure 4.5: **Gene expression as a function of RNAP binding energy.** (A) LacZ activity measured in Miller units and (B) average mRNA per cell vs. promoter binding energy in units of $k_{\mathrm{B}}T$ (with the zero of energy set to be the average interaction energy between RNAP and the the entire *E. coli* chromosome). To illustrate the reproducibility of our measurements, the translucent points represent individual measurements and the solid points represent the averaged value over repeated experiments. The solid black line in each plot is the Boltzmann factor scaling, $\propto e^{(-\Delta\epsilon/k_{\mathrm{B}}T)}$. The red data points correspond to the wild-type *lac* promoter, which was used to calibrate the arbitrary units of our energy matrix to (physical) $k_{\mathrm{B}}T$ units. The magenta, red, blue, and green data points represent promoters which we examine in the context of simple repression.

P_{bound}, we can use equation 4.3 to conclude that

$$\log\left(\text{Gene Expression}\right) = \log(n_0) - \frac{\Delta\epsilon}{k_BT}, \tag{4.5}$$

where n_0 is an unknown constant of proportionality related to the number of mRNA or proteins expected from a promoter with $\Delta\epsilon = 0$. With this relation in hand, we are now equipped to take the predicted energy for each RNAP binding site and compare the resulting expression to that predicted from equation 4.5.

4.2.2 Constitutive gene expression measurements: mRNA and protein

To test the predictive power of the binding energy model, we measured protein expression and mRNA copy numbers for constitutive expression from each of our unique promoters. Based on equation 4.5, a semi-log plot of these data against their respective predicted binding energies in units of k_BT should fall along a straight line with slope equal to -1, consistent with Boltzmann scaling. Indeed, with the unknown constant n_0 as our single fit parameter, we find that gene expression follows the exponential relation predicted from the thermodynamic model in equation 4.5, as seen in Figure 4.5. In this figure, we have taken the zero of energy to be the average energy of RNAP binding across the whole *E. coli* genome calculated from the energy matrix of Figure 4.2, as detailed in the Methods section below. The root-mean-square deviations of our fits are 1.02 for mRNA and 1.06 for protein. Since these values are the deviations of the natural logarithm of gene expression, we must exponentiate them to get a sense of the deviation in physical units. We conclude that our design process accurately predicts expression to within a factor of $e^1 \approx 3$ over nearly three orders of magnitude. In addition, the table in Figure 4.3 shows the predicted energy for each promoter (the column labeled "Model"), calculated using the matrix in Figure 4.2, as well as the experimentally measured energies of each promoter. To compute these measured energies, we solve equation 4.5 for $\Delta\epsilon$, yielding $\Delta\epsilon = \log(n_0/\text{Gene Expression}) \times k_\text{B}T$. We then plug in the measured expression for each promoter and the inferred value for n_0 (the y-intercept of the black line in Figure 4.5) to compute $\Delta\epsilon$ for each promoter. The measured values for the RNAP binding energies for the LacZ and mRNA data are listed in Figure 4.3. The promoters with colored entries will be further examined in the context of simple repression later in this work. The direct correlation between these two measurements of gene expression are shown in Figure 4.8 where protein expression is plotted vs. average mRNA copy number for every promoter strength, exhibiting an excellent linear relation between these two readouts of expression.

Fitting the data in Figure 4.5 to the full form for P_bound in equation 4.2, allowing both P/N_{NS} and the unknown proportionality constant between P_bound to vary, we find $P/N_{NS} \approx 10^{-4}$ for both the mRNA and the protein data. This is consistent with typical values for RNA polymerase copy number and the length of the *E. coli* genome ($1 - 3 \times 10^3$ [28–31] and 10^7, respectively), and thus the weak promoter limit appears to hold over the range of promoter strengths tested.

4.2.3 Protein burst size

Since mRNA and protein are linked by translation, their levels for a given promoter should be related. Individual mRNAs can be translated multiple times and it has been shown that the number of translations per mRNA is well described by an exponential distribution with mean b, known as the protein burst size, which is the average number of proteins produced per mRNA [8, 32, 33]. Using the data described above, we can extract the burst size, defined as the ratio of protein production rate and the mRNA production rate, $b =< r_\text{protein} > / < r_\text{mRNA} >$ [8, 34]. The quantity we measure, however, is the steady-state copy number $n =< r > /\gamma$, where $< r >$ is the average rate of mRNA or protein production and γ is the associated decay rate. Figures 4.5A and B demonstrate that the copy number n is well described by Boltzmann scaling with $n = n_0 \exp(-\Delta\epsilon/k_BT)$. Using this knowledge, we rewrite the burst size as

$$b = (n_0^\text{LacZ}/n_0^\text{mRNA})(\gamma_\text{LacZ}/\gamma_\text{mRNA}), \tag{4.6}$$

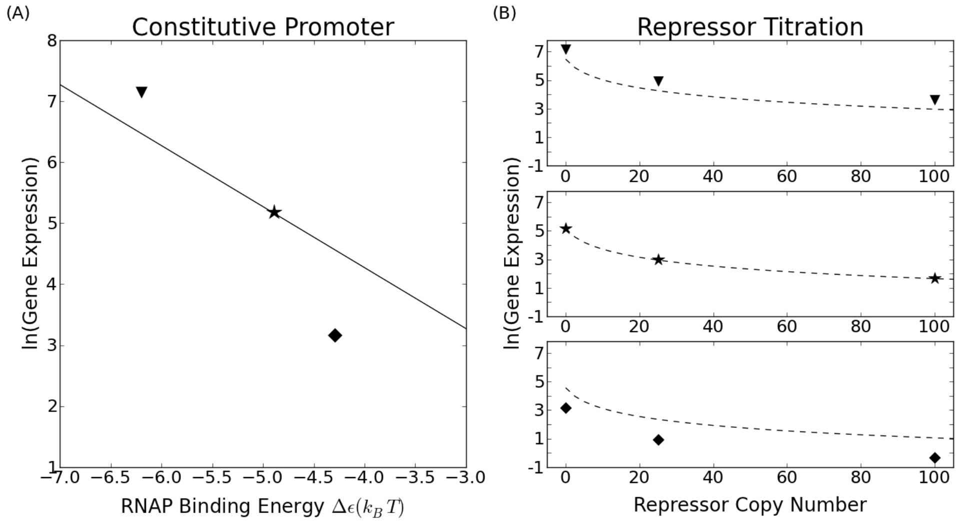


Figure 4.6: **Expected relation between predictions and measurement for simple repressor titration.** Figure (A) shows three hypothetical promoters for which the predictions of the promoter design are either numerically correct ($\star$), underestimated ($\bigtriangledown$) or overestimated ($\diamond$). The three smaller figures in (B) show the expected result as repressors are added in a simple repression architecture. The predicted theory line and the data points differ on average by the same percent as they do at $R = 0$.

with $\gamma_{\text{mRNA}} = 1/1.5$ minutes^{-1} [35] and $\gamma_{\text{LacZ}} = 1/60$ minutes^{-1} (equal to the inverse of the cell division time). This gives us a measurement of the LacZ activity (measured in Miller units, described in the methods section) per mRNA; from available biochemical data we convert from Miller units to number of LacZ tetramers [36–39] (1 Miller unit $\approx$ 0.5 LacZ tetramers/cell [39]). Plugging these values into equation 4.6 we find the protein burst size, b, for the particular RBS we have used, is roughly $5 - 6$ LacZ tetramers or $20 - 24$ individual LacZ proteins per mRNA.

4.2.4 Thermodynamic model for simple repression

Our discussion so far has focused on the behavior of the designed promoters in the absence of any regulatory interventions. We were interested in examining the portability of these promoters to other contexts such as when they are regulated by transcription factor binding. In the *E. coli* genome, there are hundreds of genes that are regulated by motifs involving simple repression [40]. For these architectures, there is a single binding site for a repressor protein which reduces the expression from the gene of interest.

Addition of a repressor which binds to a proximal binding site necessitates the addition of a term to the partition function of the RNAP binding probability given by equation 4.2. This additional term corresponds to the probability of repressor binding and making the promoter unavailable to polymerase. The resulting expression level in the context of thermodynamic models is then given by

$$P_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\Delta\epsilon/k_B T}}{1 + \frac{P}{N_{NS}} e^{-\Delta\epsilon/k_B T} + \frac{2R}{N_{NS}} e^{-\Delta\epsilon_R/k_B T}}, \tag{4.7}$$

where R is the number of repressors (the factor of two originates from the fact that LacI has two binding heads) and $\Delta\epsilon_R$ is the binding strength of that repressor to the specific binding site [2, 25]. In the weak promoter limit the expression can be simplified to,

$$\text{LacZ expression} = n_0^{\text{LacZ}} e^{-\Delta\epsilon/k_B T} (1 + \frac{2R}{N_{NS}} e^{-\Delta\epsilon_R/k_B T})^{-1}, \tag{4.8}$$

where n_0^{LacZ} was determined in the previous section by fitting equation 4.5 to the constitutive expression data in Fig. 4.5A. We therefore have an absolute prediction for the level of gene expression in our LacZ measurements. The prefactor $n_0^{\text{LacZ}} \exp(-\Delta\epsilon/k_{\text{B}}T)$ is the constitutive (R=0) prediction for expression. It is a constant prefactor for all values of R (at a given promoter strength) and thus the model predicts that any discrepancies between predicted and measured RNAP binding energies will be inherited through all repressor concentrations. This point is illustrated in Figure 4.6 where we show how the repressor titration predictions depend upon how well the original constitutive promoters follow the simple Boltzmann scaling. In particular, we show the level of expression for three hypothetical promoters, one whose constitutive properties are underestimated, one whose constitutive properties are overestimated and one for which the Boltzmann scaling is obeyed precisely. What we see is that the repressor titration (Figure 4.6B) inherits the error already present in the constitutive promoters from incorrectly predicting the RNAP binding energy.

4.2.5 Gene expression in simple repression

In each of our strains, the LacI O2 binding site is present near the promoter (see Figure 4.3). We reintroduce the repressor into our strains by integrating a cassette into the genome which expresses LacI. Specific LacI concentrations are obtained through modulation of the ribosomal binding sequence of the LacI gene. Using this process we create five unique strains with average LacI copy numbers between 10 and 140 repressors per cell. Using equation 4.8, we can make parameter-free predictions for the overall level of gene expression as a function of promoter strength, repressor binding strength and repressor copy number for the simple repression architecture. In Figure 4.7A, we show a comparison between predicted and measured protein expression in the case of simple repression, as a function of repressor copy number and of predicted promoter binding strength (using $\Delta\epsilon$ from the “model” column of Figure 4.3, and $\Delta\epsilon_R = -14.3\ k_{\text{B}}T$ as found in reference [2]). Our measurements (using the same LacZ assay as for the constitutive data above) for three distinct promoters along with data from the *lac*UV5 promoter (from reference [2]) are shown as points color

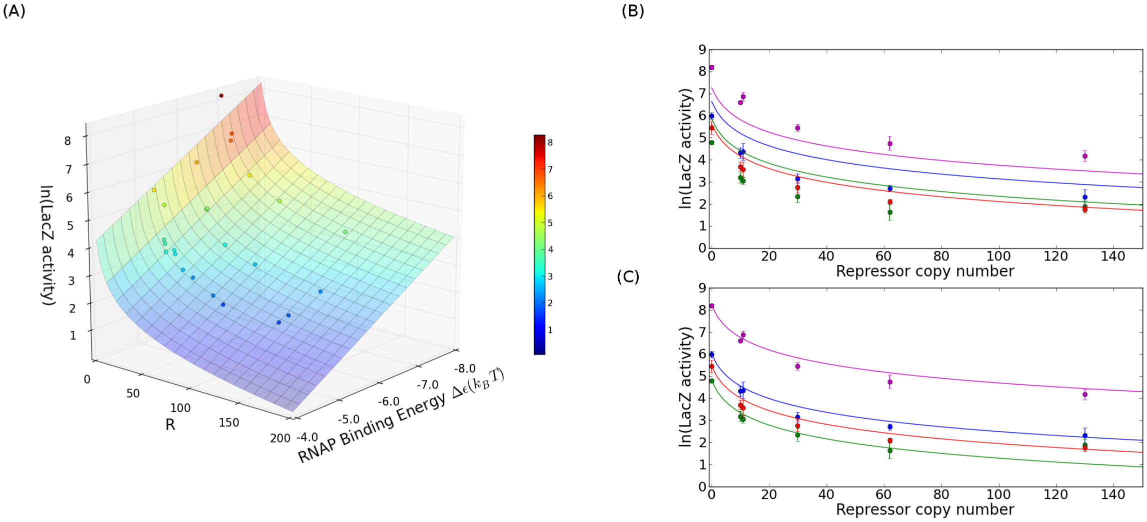


Figure 4.7: **Gene expression in the simple repression case.** (A) Solid surface: predicted gene expression of equation 4.7 as a function of repressor copy number R and RNAP binding energy $\Delta\epsilon$. Data points represent measurements of gene expression in a strain with a given promoter and repressor copy number. (B) Data from part (A) collapsed onto the RNAP binding energy axis. The solid lines are the zero parameter predictions from the theory in equation 4.7 using $\Delta\epsilon$ predicted from the position-weight matrix in Figure 4.2 (numerical values listed in Figure 4.3 under "model"). There is a systematic deviation between the theory and the experimental data which is inherited from the imperfect prediction of $\Delta\epsilon$ by the RNAP binding strength model (illustrated schematically in Figure 4.6. In (c) the same data are shown after we have corrected $\Delta\epsilon$ to fall on the theory fit line based on the constitutive expression (numerical values listed in Figure 4.3 under "LacZ"). Here we see that by correcting for the initial uncertainty in the binding energy prediction we observe good agreement between the theory and experimental data which indicates that our designed promoters function as expected even in a different regulatory context.

coded by expression level; Figure 4.7B shows the same comparison between theory and experiment collapsed along the promoter-strength axis. Each color represents a different promoter strength, with points representing measurements and the solid line representing the theoretical prediction for that promoter.

The data in Figure 4.7B show a clear trend, for any one promoter, to either over or under predict the expression as was sketched in Figure 4.6. We attribute this to imperfect predictive powers of the RNAP binding energy model from Kinney *et al* (shown in Figure 4.2) [20]: if the thermodynamic theory underpredicts the measured expression at R=0 using the model value for the RNAP binding energy (for instance, the magenta point in Figure 4.5A), the theory will continue to underpredict the measured expression as repressors are added (as seen for the magenta points in Figure 4.7B). In Figure 4.7(C) we show the result of using the measured RNAP binding energies (from the column labeled "LacZ" in Fig. 4.3) for the promoter binding strength and the accordance between theory and experimental data is evident. It is clear from these measurements that our promoter library exhibits the kind of "transferability" required in order to use them in different regulatory contexts.

In particular, the comparison between theory and experiment is very favorable even for the repressed architectures and the imperfect agreement is actually primarily an inheritance of the imperfect accord between theory and experiment for the unregulated promoters themselves.

4.3 Discussion

In this paper, we have shown how high throughput data obtained from experiments like those in reference [20] provide a foundation that, together with quantitative predictions from simple thermodynamic models [21–25], can be used to *predictively* tune protein-DNA interactions to produce a desired output from a gene with high precision. This approach contrasts with previous promoter engineering efforts, which have typically relied upon generating promoter libraries using random mutagenesis, followed by selection for mutants with desired expression levels [41–43]. We believe that predictive, model-based engineering of promoters represents a significant technical improvement over random mutagenesis, and moreover points the way to simultaneously engineering multiple aspects of promoter function (such as repressor or activator binding strengths) in a scalable way. We demonstrate the validity of our approach by simultaneously varying RNAP-promoter binding strength and the copy number of a transcription factor that represses these promoters. In this case, we can predict the absolute level of gene expression (once the conversion constant between binding probability and expression units, n_0, is known) as a function of transcription factor concentration.

While the binding site design procedure described here focused on alterations to the -10 and -35 region of promoters, we have made preliminary studies in which promoters are subjected to more severe perturbations, which indicate that the energy function does not describe these situations nearly so well. It is clear that changes in the linker region can have subtle effects on the twist registry and absolute spacing of the -10 and -35 binding sites that are not well accounted for by a linear weight matrix, which ignores correlations in multiple basepair changes [44]. Despite these challenges, constitutive expression from promoters designed in this study agrees well with the scaling predicted from the simple thermodynamic model presented here, and we have shown that our knowledge of simple repression can be applied on top of our understanding of constitutive expression to accurately predict the absolute expression from a gene when repression is introduced.

4.4 Methods

4.4.1 Energy matrix

The energy matrix from [20] is given in arbitrary energy units (AU). To calibrate these arbitrary units to physical units, we need two known reference energies, since only differences in energy are physically significant. From [45], we know that RNAP binds the wild-type (WT) *lac* promoter with a binding energy 5.35 $k_\mathrm{B}T$ more favorable than the non-specific background. Using the matrix from [20], we find that the wild-type *lac* promoter has a binding energy of 53.4 AU, while the average binding energy of all 41 bp segments in the *E. coli* strain MG1655 is 91.3 AU (recall that the more positive the energy value, the less favorable the binding interaction). To obtain this value, we began at the chromosomal origin of replication and applied the matrix sequentially to each 41 bp segment (both forward and reverse strands) around the chromosome, and computed the mean of the resulting $\sim 10^7$ energy values. Thus, we find that a difference of $91.3 - 53.4 = 37.9$ AU is equivalent to a difference of 5.35 $k_\mathrm{B}T$, providing us with a conversion factor of $37.9/5.35 = 7.08$ AU per $k_\mathrm{B}T$.

To see how this plays out in practice, consider a hypothetical sequence whose binding energy is computed to be 60.0 AU. The number we are actually interested in is $\Delta\epsilon = (\epsilon_S - \epsilon_{NS})$. For this promoter sequence, we find that $\Delta\epsilon = (60.0 - 91.3)/7.08 = -4.42$ $k_\mathrm{B}T$. We used the same approach to convert from AU to the $k_\mathrm{B}T$ units on the x-axis of Figure 4.5 for each of our distinct promoter sequences.

4.4.2 Strains

All strains used are wild-type *E.coli* (MG1655) with a complete deletion of the *lacIZYA* genes [39]. Modified promoters are created through site-directed mutagenesis of plasmid pZS2502+11-lacz [2, 46], which has the *lac*UV5 promoter expressing LacZ (our reporter gene). These constructs are then integrated into the *galK* region using recombineering [47]. A schematic of the integrated region is shown in Figure 4.3. The end result is a strain with a desired, multi-basepair change to the *lac*UV5 promoter which expresses LacZ and a complete deletion of the LacI protein. Our designed promoters span roughly three orders of magnitude in constitutive expression and vary from the wild-type promoter by as few as one or as many as nine individual basepair changes. The site labeled “O2” is a binding site for the LacI repressor protein.

For the strains involving simple repression, we took our constitutive expression strains and created as many as eight different strains with the LacI cassettes from reference [2] integrated at the *ybcN* site. The cassettes contain LacI expressed from an unregulated *tet* promoter with unique ribosomal binding sequences to produce varying LacI copy numbers. The exception is the data point at an average LacI copy number of 11, which corresponds to the native wild-type LacI gene. The measurements for repressors per cell are from quantitative immunoblots in Ref [2]. One of our strains, the one with 10 repressors/cell, has not been characterized this way, but instead the repressors/cell has been inferred from the measured expression of the *lac*UV5 promoter.

4.4.3 Growth

Cultures were grown overnight (at least 8 hours) in LB and diluted 1:4000 into 30 mL of M9 minimal media supplemented with 0.5% glucose in a 125mL baffled flask. Cells were grown approximately 8 hours and harvested in exponential phase when OD600= $0.3 - 0.5$ was reached.

4.4.4 LacZ assay

Our assay for measuring LacZ activity is the same as described in reference [2], which is a slightly modified version of that described in Ref [36]. A volume of cells from each sample between 5 μL and 200 μL was added to Z-buffer (60mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) to reach a total of 1 mL. This volume is chosen to minimize the uncertainty in measuring the time of reaction ($\sim 1 - 10$'s of hours) and the yellow color is easily distinguishable from a blank sample of 1 mL of Z-buffer. The assay was performed in 1.5 mL Eppendorf tubes. The cells were lysed by addition of 25 μL of 0.1% SDS followed by 50 μL of chloroform, mixed by a 10 s vortex. The reaction was started with the addition of 200 μL of 4mg/mL 2-nitrophenyl β-D-galactopyranoside (ONPG) in Z-buffer. The developing yellow color (proportional to the concentration of the product ONP) was monitored visually. Once sufficient yellow had developed in a tube (easily measurable by OD550 and OD420, without saturating the reading), the reaction was stopped by adding 200 μL of 2.5 M Na_2CO_3. (Typically 500 μL of a 1M solution is added in other protocols, but this change allows for the entire reaction to take place in a 1.5 mL Eppendorf tube.) Once all samples were stopped, the tubes were spun at $> 13,000$ g for 3 min in order to reduce the contribution of cell debris to the measurement. 200 μL of each sample were loaded into a 96 well plate and OD420 and OD550 measurements were taken on a Tecan Safire2 with the Z-buffer sample as a blank. In addition, the OD600 of 200 μL of each culture was taken with the same instrument. The absolute activity of LacZ is measured in Miller units,

$$MU = 1000\frac{OD420 - 1.75 \times OD550}{t \times v \times OD600}0.826, \tag{4.9}$$

where t is the reaction time in minutes, v is the volume of cells used in milliliters and OD refers to the optical density measurements obtained from the plate reader. The factor of 0.826 accounts for the use of 200 muL Na_2CO_3 as opposed to 500 μL which changes the concentration of ONPG in the final solution.

4.4.5 Single Cell mRNA FISH

Our assay is based on that used in reference [9]. Once a culture reaches OD600= $0.3 - 0.5$, it is immersed in ice for 15 minutes before being harvested in a large centrifuge chilled to $4°C$ for 5 minutes at 4500 g. The cells are then fixed by resuspending in 1 mL of 3.7% formaldehyde in 1x PBS which is then allowed to mix gently at room temperature for 30 minutes. Next, they are centrifuged (8 minutes at 400 g) and washed twice in 1 mL of 1x PBS twice. The cells are permeabilized by resuspension in 70% Ethanol which proceeds, with mixing, for 1 hour at room temperature. The cells are then pelleted (centrifuge at 600 g for 7 minutes) and resuspended in 1 mL of 20% wash solution (200 μL formamide, 100 μL 20x SSC, 700 μL water) and resuspended in 50 μL of DNA probes (consisting of an mix of 72 unique DNA probes, individual oligo sequences available as SI Text S5) labeled with ATTO532 dye (Atto-tec) in hybridization solution (0.1 g dextran sulfate, 0.2 mL formamide, 1 mg *E.coli* tRNA, 0.1 mL 20x SSC, 0.2 mg BSA, 10 μL of 200 mM Ribonucleoside vanadyl complex). This hybridization reaction is allowed to proceed overnight. The hybridized product is then washed four times in 20% wash solution before imaging in 2x SSC.

4.4.6 FISH data acquisition

Samples are imaged on a 1.5% agarose pad made from PBS buffer. Each field of view is imaged with phase contrast at the focal plane and with 532 nm epifluorescence (Verdi V2 laser, Coherent Inc.) both at the focal plane and in 8 z-slices spaced 200 nm above and below the focal plane, sufficient to cover the entire depth of the *E. coli.* The images are taken with an EMCCD camera (Andor Ixon2). The phase image is used for cell segmentation and the fluorescence images are used in mRNA detection. A total of 100 unique fields of view are imaged in each sample and a typical field of view has between 5 and 15 viable cells (cells which are touching and cells that have visibly begun to divide are ignored) resulting in roughly 1000 individual cells per sample.

4.4.7 FISH analysis

The FISH data is analyzed in a series of Matlab (The Mathworks) routines. The overview of the workflow is as follows: identifying individual cells, segmenting the fluorescence to identify possible mRNA, quantifying the mRNA which are found (because of the small size of *E. coli*, at high copy number mRNA can be difficult to distinguish and count by eye).

4.4.7.1 Cell identification and segmentation

In phase contrast imaging, *E. coli* are easily distinguishable from the background and automated programs can identify, segment and label cells with high fidelity. The results of the phase segmentation are manually checked for accuracy and bad segmentations are rejected. Cells which are touching or overlapping other cells, misidentification of cells or their boundaries or cells which have visibly begun to undergo division, etc are all discarded manually.

4.4.7.2 Fluorescence segmentation

First we perform several steps to process the raw intensity images. The images are flattened, a process to correct for any uneven elements in the illumination profile, using a fluorescence image of an agarose pad coated with a small drop of fluorescein (such that the drop spreads evenly across most of the pad), each pixel of every fluorescence image is scaled such that the corresponding pixel in the flattening image would be a uniform brightness (typically each pixel is scaled up to the level of the brightest pixel). This can be achieved by renormalizing each pixel in the data images and dividing by the ratio of the intensity of the corresponding pixel in the flattening image to the intensity of the brightest pixel. For instance, if one pixel in the flattening image was half as bright as the brightest pixel, the signal at that pixel's position in the raw intensity images would be doubled. We then subtract from every pixel the contribution to our signal associated with autofluorescence. The value for the autofluorescence is obtained by averaging over the fluorescence of every pixel in a control sample (one which underwent the entire FISH protocol but did not possess the *LacZ* gene). Finally, all local 3D maxima (where $x - y$ is the image plane) in fluorescence are identified. We require that the maxima be above a threshold in fluorescence (typically $300 - 400\%$ above the background autofluorescence signal). This threshold eliminates all fluorescence maxima in the control sample, which does not contain the *LacZ* gene.

4.4.7.3 mRNA quantification

Each identified maximum pixel is dilated in the image plane to a 5×5 box of surrounding pixels. If this causes maxima (herein called "spots" to avoid confusion) to overlap, the pixels which make up each overlapping spot are merged into one larger spot to avoid double counting the signal from any one pixel. Since, due to the small size of the *E. coli*, we can not guarantee that every spot corresponds to exactly one mRNA, we must divide the total summed intensity of each spot by the average intensity produced from a single mRNA. This value can be found by taking the average of the unmerged spots in very low expression samples (where the mean $\ll 1$ and mRNA are statistically very unlikely to overlap). We use several of our low expression strains to ensure that as we increase the mean expression it simply increases the frequency of spots with the single mRNA intensity but does not increase the mean intensity of each spot. The mean mRNA copy number can then be calculated by dividing each spot by the single mRNA intensity and averaging the total number of such mRNA in the entire collection of cells for each sample.

4.5 Supplementary information

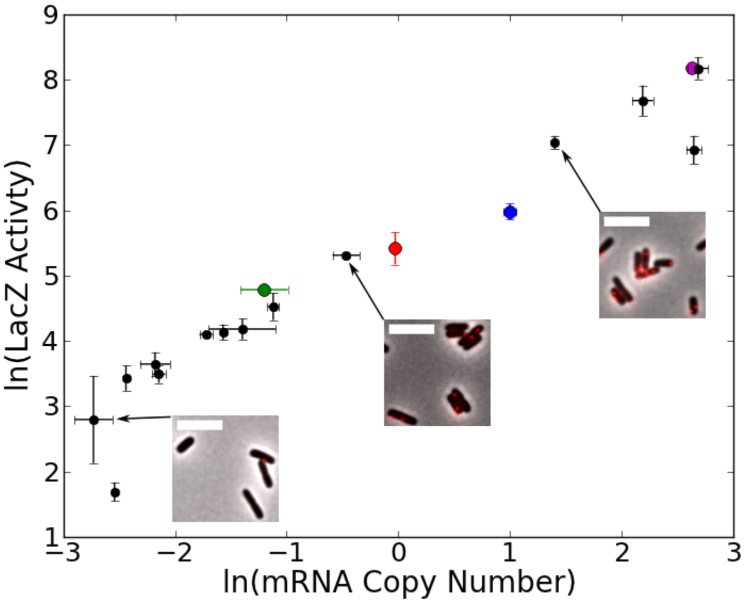


Figure 4.8: **mRNA vs. Protein Expression.** Scatter plot of mRNA vs. protein expression for each of our designed promoters. Each data point represents mRNA and protein expression measurements for a particular promoter. To obtain these values, expression of a LacZ reporter was measured at both the mRNA level (using mRNA FISH) and protein level (using the Miller assay of LacZ activity described in the methods). As would be expected from a simple model in which each mRNA produces a “burst” of translated protein molecules characterized by a fixed “burst size” b, these dual measurements display a linear relationship. The inset pictures are representative mRNA FISH images from the indicated strains. The scale bar is 5 μm.

4.5.1 Supplementary information text S1

Energy matrix for RNAP σ^{70} binding affinity Energy matrix for RNAP σ^{70} in arbitrary energy units. The energy matrix is determined from experiments in strain TK310 with no supplemental cAMP which means that these cells have no CRP. The matrix covers base pairs [-41:-1] where 0 denotes the transcription start site. Each row corresponds to a given position; each column corresponds to a value for that base pair. The columns are ordered [A,C,G,T].

```
# Energy matrix for RNAP in arbitrary units. Inferred from an
# experiment done in TK310 with no supplemental cAMP (and hence, no
# CRP present in the cells). The matrix covers base pairs [-41:-1]
# where 0 denotes the transcription start site. Each row corresponds
# to a given position; each column corresponds to a value for that
# base pair. The columns are ordered [A,C,G,T].
   3.3090086e-02   8.4338901e-01   1.9145915e-01   5.7156605e-01
   2.3776175e-01   1.5712752e+00   6.7058076e-03   1.3919617e+00
   1.0944116e+00   8.2535084e-01   7.1361981e-01   2.2328462e-06
```

```
5.9864426e-02   1.0066429e+00   6.9407124e-02   9.7620436e-01
2.6802048e+00   1.2734957e+00   0.0000000e+00   5.7258818e+00
4.3852720e+00   7.0449035e+00   4.7688539e+00   0.0000000e+00
3.0848289e+00   3.9709489e+00   3.4340292e+00   0.0000000e+00
1.2843899e+01   1.2775114e+01   0.0000000e+00   6.7068567e+00
0.0000000e+00   9.5273671e+00   1.2366599e+00   7.1684270e+00
9.7567254e+00   7.0366632e-01   1.0145991e+01   0.0000000e+00
0.0000000e+00   6.8593905e+00   4.3133704e+00   2.3905484e+00
0.0000000e+00   1.7594332e+00   1.3839752e+00   6.8668172e-01
7.5845192e-01   1.5786643e+00   0.0000000e+00   7.0599327e-01
2.8890547e-01   9.5169374e-01   2.8413340e-02   1.0598483e+00
5.3030278e-01   9.4433893e-01   6.7437472e-01   7.2803717e-05
0.0000000e+00   1.9163061e+00   9.9594277e-01   1.7259675e+00
1.4990845e+00   1.0768794e+00   7.7364760e-01   0.0000000e+00
0.0000000e+00   2.9917723e+00   2.1527347e+00   4.1632716e+00
4.1263772e-01   7.9893094e-03   1.9843027e-01   1.2690202e+00
4.9869143e-01   7.2434231e-01   5.6449291e-01   2.7238914e-04
2.5038165e-01   6.5802748e-01   2.1211249e-01   4.2288681e-02
0.0000000e+00   1.0634132e+00   1.0747566e+00   8.7305312e-01
2.8977506e-01   4.9904053e-01   8.8848304e-02   1.1179347e-01
3.2567358e-01   1.2689945e+00   1.1829313e+00   6.0211464e-03
2.7597944e+00   2.4891846e+00   2.6693995e+00   0.0000000e+00
0.0000000e+00   3.3573277e+00   1.2712026e+00   4.6265286e+00
1.8671571e+00   2.9598860e+00   0.0000000e+00   2.3774089e+00
4.2376464e+00   8.0605587e+00   0.0000000e+00   4.6122469e+00
1.9201763e+00   1.4430513e+00   0.0000000e+00   7.6884400e-01
4.9396224e+00   7.8252084e+00   9.9642909e+00   0.0000000e+00
0.0000000e+00   1.1449195e+01   1.0351181e+01   1.1048615e+01
1.3484172e+00   3.4139074e+00   4.2597235e+00   0.0000000e+00
0.0000000e+00   4.2758871e+00   5.5404763e+00   6.0569935e+00
0.0000000e+00   2.1330405e+00   5.5662408e+00   5.8880615e+00
7.0033761e+00   1.0815480e+01   9.2473926e+00   0.0000000e+00
0.0000000e+00   3.4444978e+00   1.7185707e+00   3.0026213e+00
2.0895130e-01   2.5615064e+00   9.1081798e-01   1.1727280e-02
1.3337890e-05   1.1660204e+00   1.1205350e+00   7.2778078e-01
1.9009344e-01   1.0398295e+00   2.5208391e-01   3.1778086e-02
```

```
   0.0000000e+00   3.1166170e+00   2.7723361e+00   2.4297976e+00
   4.3042402e-01   5.1900833e-01   8.7572299e-01   1.2296102e-03
```

4.5.2 Supplementary information text S2

Energy matrix for RNAP σ^{70} binding affinity Energy matrix for RNAP σ^{70} in units of $k_{\mathrm{B}}T$. The numerical values here are shown pictorially in Figure 4.2. The matrix covers base pairs [-41:-1] where 0 denotes the transcription start site. Each row corresponds to a given position; each column corresponds to a value for that base pair. The columns are ordered [A,C,G,T].

```
# Energy matrix for RNAP in kT. Inferred from an experiment done in
# TK310 with no supplemental cAMP (and hence, no CRP present in the
# cells). The matrix covers base pairs [-41:-1] where 0 denotes the
# transcription start site. Each row corresponds to a given position;
# each column corresponds to a value for that base pair. The columns
# are ordered [A,C,G,T].
-3.1342732e-01 -1.9897832e-01 -2.9105881e-01 -2.3737140e-01
-2.8079230e-01 -9.2442945e-02 -3.1342732e-01 -1.1776971e-01
-1.5884973e-01 -1.9685266e-01 -2.1263388e-01 -3.1342732e-01
-3.1342732e-01 -1.7970155e-01 -3.1207948e-01 -1.8400078e-01
6.5132678e-02 -1.3355505e-01 -3.1342732e-01 4.9531304e-01
3.0596138e-01 6.8161554e-01 3.6013961e-01 -3.1342732e-01
1.2228297e-01 2.4744117e-01 1.7160505e-01 -3.1342732e-01
1.5006827e+00 1.4909673e+00 -3.1342732e-01 6.3386882e-01
-3.1342732e-01 1.0322460e+00 -1.3875785e-01 6.9906237e-01
1.0646412e+00 -2.1403942e-01 1.1196223e+00 -3.1342732e-01
-3.1342732e-01 6.5541314e-01 2.9580578e-01 2.4220757e-02
-3.1342732e-01 -6.4919808e-02 -1.1795060e-01 -2.1643838e-01
-2.0630135e-01 -9.0452139e-02 -3.1342732e-01 -2.1371076e-01
-2.7663465e-01 -1.8302049e-01 -3.1342732e-01 -1.6774442e-01
-2.3853608e-01 -1.8005640e-01 -2.1818694e-01 -3.1342732e-01
-3.1342732e-01 -4.2762619e-02 -1.7275744e-01 -6.9646602e-02
-1.0169222e-01 -1.6132571e-01 -2.0415506e-01 -3.1342732e-01
-3.1342732e-01 1.0913939e-01 -9.3687490e-03 2.7460539e-01
-2.5627359e-01 -3.1342732e-01 -2.8652888e-01 -1.3531561e-01
-2.4302915e-01 -2.1115756e-01 -2.3373516e-01 -3.1342732e-01
-2.8403566e-01 -2.2645857e-01 -2.8944091e-01 -3.1342732e-01
-3.1342732e-01 -1.6322772e-01 -1.6162554e-01 -1.9011473e-01
```

```
-2.8504784e-01 -2.5549057e-01 -3.1342732e-01 -3.1018648e-01
-2.6827867e-01 -1.3504126e-01 -1.4719707e-01 -3.1342732e-01
7.6374146e-02 3.8152423e-02 6.3606505e-02 -3.1342732e-01
-3.1342732e-01 1.6077151e-01 -1.3387893e-01 3.4003717e-01
-4.9704568e-02 1.0463567e-01 -3.1342732e-01 2.2364895e-02
2.8511030e-01 8.2506967e-01 -3.1342732e-01 3.3801998e-01
-4.2215981e-02 -1.0960652e-01 -3.1342732e-01 -2.0483354e-01
3.8425946e-01 7.9182810e-01 1.0939584e+00 -3.1342732e-01
-3.1342732e-01 1.3036906e+00 1.1486039e+00 1.2471115e+00
-1.2297292e-01 1.6876299e-01 2.8822854e-01 -3.1342732e-01
-3.1342732e-01 2.9051153e-01 4.6912583e-01 5.4208023e-01
-3.1342732e-01 -1.2150416e-02 4.7276488e-01 5.1821978e-01
6.7575009e-01 1.2141828e+00 9.9270158e-01 -3.1342732e-01
-3.1342732e-01 1.7308367e-01 -7.0691348e-02 1.1067173e-01
-2.8557082e-01 4.6710971e-02 -1.8643711e-01 -3.1342732e-01
-3.1342732e-01 -1.4873706e-01 -1.5516155e-01 -2.1063531e-01
-2.9106640e-01 -1.7104718e-01 -2.8231068e-01 -3.1342732e-01
-3.1342732e-01 1.2677282e-01 7.8145573e-02 2.9764429e-02
-2.5280664e-01 -2.4029473e-01 -1.8991131e-01 -3.1342732e-01
```

4.5.3 Supplementary information text S3

Source code to adapt energy matrix from Kinney *et. al* [20] This code converts from the arbitrary units of SI text S1 to the values in units of $k_\mathrm{B}T$ as in SI Text S2. This code adds a constant offset to the matrix such that the average value of $E(S)$ across the *E. coli* genome is zero. The basis for this conversion is the reference of $-5.35\ k_\mathrm{B}T$ [45] for the binding energy of the wild-type promoter.

```
#!/usr/bin/python

import numpy as np

# import the original energy from Kinney et al
emat_orig = np.genfromtxt(’rnap-full-0_emat.txt’)
# set conversion factor from AU to k_B T. See "Methods" for details on
# how this number was obtained.
AU_per_kT = 7.08
# average energy value for non-specific binding as computed using the
# original, unmodified energy matrix
avg_nonspecific_AU = 91.3 # AU
# length of the binding site as defined by our matrix.
site_len = 41

# Set the best binding base of each row to zero. This is so that all
# values are in terms of an energy *difference*: the difference
# between that particular base and the best binding base. Remember,
# adding a constant to each element in a particular row has no
# physical meaning.
# For later, we’ll need to keep track of how much we’ve subtracted
# from the original matrix.
sum = 0 # how much we’ve subtracted
emat_zeroed = np.zeros(emat_orig.shape)
for i,row in enumerate(emat_orig):
    emat_zeroed[i,:] = row - min(row)
    sum = sum + min(row)

# Convert the units of the matrix from AU to k_B T. Now all nonzero
# matrix element are energy *differences* in terms of k_B T.
emat_rescaled = emat_zeroed/AU_per_kT

# Finally, add an offset to the matrix so that the average binding
# across the MG1655 genome is zero. This is for convenience only, and
# has no physical meaning. It’s not immediately obvious how to do this
# for the rescaled matrix, but it is easy to see how to do it for the
# original matrix. Since the original matrix has an average
# nonspecific binding of 91.3 AU, we could just subtract 91.3 from all
# elements in a particular row. Or, to make things a bit more
# equitable, we could subtract 91.3/41 = 2.23 AU from each row in the
# matrix. Now that we’ve figured out how many AU to subtract from each
# row in the original matrix, we realize that we can simply subtract
# 2.23/AU_per_KT = 2.23/7.08 = 0.315 from each row in the rescaled
# matrix to obtain the matrix we want. This is almost right, but
# remember that we’ve already subtracted the minimum of each each from
# each row. The total amount subtracted is contained in the variable
# <sum> defined above. If we used the emat_zeroed matrix to compute
# the average nonspecific energy, we would obtain
# <avg_nonspecific_AU - sum> as an answer. So "91.3" needs to be
# replaced with "91.3 - sum" everywhere in the preceding paragraph.

emat_rescaled_zeroed_to_MG1655 = (emat_rescaled -
                                  ((avg_nonspecific_AU-sum)/site_len)/AU_per_kT)

# save the matrix in a text file
np.savetxt('rnap-full-0_emat_relativetoMG1655background_kT.txt',
           emat_rescaled_zeroed_to_MG1655)
```

4.5.4 Supplementary information text S4

Promoter sequence for constitutive expression strains This spreadsheet contains the colloquial name and promoter sequence for each of the unique constitutive expression strains generated for this study. The following column contains the calculated energy for each promoter using the energy matrix in SI text S1 (from [20]). The final column is the result for the binding affinity of each promoter in units of $k_\mathrm{B}T$ and zeroed to the *E. coli* chromosome using the energy matrix given in Figure 4.2 and SI text S2, as described in the methods section.

```
Name,Sequence,Energy (AU),Energy (kT)
UV5,TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG,41.79623115,-6.992057748
WT,CAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG,53.44611685,-5.346593665
WTDL10,CAGGCATTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG,57.83138885,-4.727204964
WTDL20,CAGGCTTAAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG,69.02548383,-3.146118103
WTDL20v2,CAGGCCTTAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG,69.93334503,-3.01788912
WTDL30,CAGGCCTCAGACTTTATGCTTCCGGCTCGTATGTTGTGTGG,76.00160233,-2.160790631
WTDR30,CAGGCTTTACACTTTATGCTTCCGGCTCGGTTGTAGTGTGG,81.46239885,-1.389491687
5DL1,TCGCGTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG,42.74300962,-6.858331975
5DL5,TCGTGTTTACCCTTTATGCTTCCGGCTCGTATAATGTGTGG,49.57196158,-5.893790737
5DL10,TCGAGATTACACTTTATGCTTCCGGCTCGTATAATGTGTGG,46.18150315,-6.372669047
5DL20,TCGAGTTAAGACTTTATGCTTCCGGCTCGTATAATGTGTGG,57.37559813,-4.791582186
5DL30,TCGAGCTCAGACTTTATGCTTCCGGCTCGTATAATGTGTGG,64.35171663,-3.806254714
5DR1,TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGGGTGG,42.69532185,-6.865067536
5DR1v2,TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGGGAGG,42.8536372,-6.84270661
5DR5,TCGAGTTTACACTTTATGCTTCCGGCTCGAATAATGTGTGG,46.73585355,-6.294370968
```

```
5DR10,TCGAGTTTACACTTTATGCTTCCGGCTCGGATAATGTGTGG,51.76052205,-5.584672028
5DR20,TCGAGTTTACACTTTATGCTTCCGGCTCGGATAACGTGTGG,62.57600205,-4.057061858
5DR30,TCGAGTTTACACTTTATGCTTCCGGCTCGGTTAAAGTGTGG,69.81251315,-3.03495577
```

4.5.5 Supplementary information text S5

List of FISH probe sequences A list of all 72 probes and their sequences used in the mRNA FISH protocol. See Table 6.1.

Bibliography

[1] Howard M. Salis, Ethan A. Mirsky, and Christopher A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, 2009.

[2] Hernan G. Garcia and Rob Phillips. Quantitative dissection of the simple repression input–output function. *Proc. Nat. Acad. Sci*, 108(29):12173–12178, 2011.

[3] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[4] Mark Lies and Michael R Maurizi. Turnover of endogenous ssrA-tagged proteins mediated by ATP-dependent proteases in *Escherichia coli*. *The Journal of Biological Chemistry*, 283(34):22918–22929, 2008.

[5] Chris Grilly, Jesse Stricker, Wyming Lee Pang, Matthew R Bennett, and Jeff Hasty. A synthetic gene network for tuning protein degradation in *Saccharomyces cerevisiae*. *Mol. Syst. Biol*, 3(127):1–5, 2007.

[6] Trent A. Carrier and J. D. Keasling. Engineering mRNA stability in *E. coli* by the addition of synthetic hairpins using a 5' cassette system. *Biotechnology and Bioengineering*, 55(3):577–580, 1997.

[7] Trent A. Carrier and J. D. Keasling. Library of synthetic 5' secondary structures to manipulate mRNA stability in *Escherichia coli*. *Biotechnology Progress*, 15(1):58–64, 1999.

[8] Alvaro Sanchez, Hernan G. Garcia, Daniel Jones, Rob Phillips, and Jan Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol*, 7(3):e1001100, 03 2011.

[9] Lok-Hang So, Anandamohan Ghosh, Chenghang Zong, Leonardo A. Sepulveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in *Escherichia coli*. *Nature Genetics*, 43(6):554–560, 2011.

[10] M. T. Record, Jr., WS Reznikoff, ML Craig, KL McQuade, and PJ Schlax. *Escherichia coli* RNA polymerase (sigma70) promoters and the kinetics of the steps of transcription initiation. In Neidhardt FC *et al.*, editor, *In Escherichia coli and Salmonella Cellular and Molecular Biology*, pages 792–821. ASM Press, Washington DC, 1996.

[11] Carol A. Gross, Cathleen L. Chan, and Michael A. Lonetto. A structure/function analysis of *Escherichia coli* RNA polymerase. *Philosophical Transactions: Biological Sciences*, 351(1339):475–482, 1996.

[12] Araceli M. Huerta and Julio Collado-Vides. Sigma70 promoters in *Escherichia coli*: Specific transcription in dense regions of overlapping promoter-like signals. *Journal of Molecular Biology*, 333(2):261–278, 2003.

[13] Ewa Heyduk, Konstantin Kuznedelov, Konstantin Severinov, and Tomasz Heyduk. A consensus adenine at position -11 of the nontemplate strand of bacterial promoter is important for nucleation of promoter melting. *J. Biol. Chem.*, 281(18):12362–12369, 2006.

[14] Ryan K. Shultzaberger, Zehua Chen, Karen A. Lewis, and Thomas D. Schneider. Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Research*, 35(3):771–788, 2007.

[15] Y Takeda, A Sarai, and V M Rivera. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Nat. Acad. Sci*, 86(2):439–443, 1989.

[16] P H von Hippel and O G Berg. On the specificity of DNA-protein interactions. *Proc. Nat. Acad. Sci*, 83(6):1608–1612, March 1986.

[17] Panayiotis V. Benos, Martha L. Bulyk, and Gary D. Stormo. Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Research*, 30(20):4442–4451, 2002.

[18] G D Stormo. DNA binding sites: Representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, January 2000.

[19] Eran Segal, Tali Raveh-sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451(7178):535–540, January 2008.

[20] Justin B. Kinney, Anand Murugan, Curtis G. Callan, Jr., and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Nat. Acad. Sci*, 107(20):9158–9163, 2010.

[21] Madeline A. Shea and Gary K. Ackers. The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation. *Journal of Molecular Biology*, 181(2):211–230, 1985.

[22] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proc. Nat. Acad. Sci*, 100(9):5136–5141, 2003.

[23] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev*, 15(2):116–124, 2005.

[24] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Curr Opin Genet Dev*, 15(2):125–139, 2005.

[25] Rob Phillips, Jane Kondev, and Julie Theriot. *Physical Biology of the Cell.* Garland Science, New York, 2009.

[26] Ulrich Gerland, J David Moroz, and Terence Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Nat. Acad. Sci*, 99(19):12015–12020, September 2002.

[27] A.M. Sengupta, Marko Djordjevic, and B.I. Shraiman. Specificity and robustness in transcription control networks. *Proc. Nat. Acad. Sci*, 99:2072–2076, February 2002.

[28] A Ishihama and H Yoshikawa, editors. *Control of cell growth and division*, pages 121–140. Japan Scientific Society Press, 1991.

[29] FC Neidhardt, editor. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, chapter 97, page 1559. ASM Press, 2nd edition, 1996.

[30] IL Grigorova, NJ Phleger, VK Mutalik, and CA Gross. Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Nat. Acad. Sci*, 103(14):5332–5337, April 4 2006.

[31] Stefan Klumpp and Terence Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc. Nat. Acad. Sci*, 105(51):20245–20250, 2008.

[32] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X. Sunney Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–1603, 2006.

[33] Nir Friedman, Long Cai, and X. Sunney Xie. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.*, 97:168302–168306, Oct 2006.

[34] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc. Nat. Acad. Sci*, 98(15):8614–8619, 2001.

[35] D. Kennell and H. Riezman. Transcription and translation initiation frequencies of the *Escherichia coli lac* operon. *J Mol Biol*, 114(1):1–21, 1977.

[36] J. H. Miller. *Experiments in Molecular Genetics.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1972.

[37] J. Lederberg. The beta-d-galactosidase of *Escherichia coli*, strain K-12. *J Bacteriol*, 60(4):381–392, 1950.

[38] Kurt Wallenfels and Rudolf Weil. Beta-galactosidase. *The Enzyme*, 7:617–663, 1972.

[39] Hernan G. Garcia, Heun Jin Lee, James Q. Boedicker, and Rob Phillips. Comparison and calibration of different reporters for quantitative analysis of gene expression. *Biophysical Journal*, 101(3):535–544, AUG 3 2011.

[40] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides. RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*, 39(Database issue):D98–105, 2011.

[41] Peter Jensen and Karin Hammar. The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Applied and environmental microbiology*, 64(1):82–87, 1998.

[42] Hal Alper, Curt Fischer, Elke Nevoigt, and Gregory Stephanopoulos. Tuning genetic control through promoter engineering. *Proc. Nat. Acad. Sci*, 102(36):12678–12683, September 2005.

[43] Marjan De Mey, Jo Maertens, Gaspard J Lequeux, Wim K Soetaert, and Erick J Vandamme. Construction and model-based analysis of a promoter library for *E. coli*: An indispensable tool for metabolic engineering. *BMC biotechnology*, 7(34), January 2007.

[44] Mofang Liu, Michael Tolstorukov, Victor Zhurkin, Susan Garges, and Sankar Adhya. A mutant spacer sequence between -35 and -10 elements makes the P*lac* promoter hyperactive and cAMP receptor protein-independent. *Proc. Nat. Acad. Sci*, 101(18):6911–6916, May 2004.

[45] T. Kuhlman, Z. Zhang, Jr. Saier, M. H., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci U S A*, 104(14):6043–6048, 2007.

[46] R. Lutz, T. Lozinski, T. Ellinger, and H. Bujard. Dissecting the functional program of *Escherichia coli* promoters: The combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Res*, 29(18):3873–3881, 2001.

[47] S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, and D. L. Court. Recombineering: A homologous recombination-based method of genetic engineering. *Nat Protoc*, 4(2):206–223, 2009.

Chapter 5

Effect of promoter architecture on the cell-to-cell variability in gene expression

A version of this chapter originally appeared as: Sanchez A, Garcia HG, Jones D, Phillips R, Kondev J (2011) Effect of Promoter Architecture on the Cell-to-Cell Variability in Gene Expression. PLoS Comput Biol 7(3): e1001100. doi:10.1371/journal.pcbi.1001100

Author contribution note: for this chapter, I (DLJ) wrote code for and performed Gillespie simulations, and wrote code for and performed numerical computations of mRNA probability distribution functions.

5.1 Introduction

A fundamental property of all living organisms is their ability to gather information about their environment and adjust their internal physiological state in response to environmental conditions. This property, shared by all organisms, includes the ability of single-cells to respond to changes in their environment by regulating their patterns of gene expression. By regulating the genes they express, cells are able to survive, for example, changes in the extracellular pH or osmotic pressure, switch the mode of sugar utilization when the sugar content in their medium changes, or respond to shortages in key metabolites by adapting their metabolic pathways. Perhaps more interesting is the organization of patterns of gene expression in space and time resulting in the differentiation of cells into different types, which is one of the defining features of multicellular organisms. Much of this regulation occurs at the level of transcription initiation, and is mediated by simple physical interactions between transcription factor proteins and DNA, leading to genes being turned on or off. Understanding how genes are turned on or off (as well as the more nuanced expression patterns in which the level of expression takes intermediate levels) at a mechanistic level has been one of the great challenges of molecular biology and has attracted intense attention over the past 50 years.

The current view of transcription and transcriptional regulation has been strongly influenced by recent experiments with single-cell and and single-molecule resolution [111]. These experiments have confirmed the long-suspected idea that gene expression is stochastic [12,13], meaning that different steps on the path from gene to protein occur at random. This stochasticity also causes variability in the number of messenger RNAs (mRNA) and proteins produced from cell to cell in a colony of isogenic cells [11,1417]. The question of how transcriptional regulatory networks function reliably in spite of the noisy character of the inputs and outputs has attracted much experimental and theoretical interest [18,19]. A different, but also very relevant, question is whether cells actually exploit this stochasticity to fulfill any physiologically important task. This issue has been investigated in many different cell types and it has been found that stochasticity in gene expression plays a pivotal role in processes as diverse as cell fate determination in the retina of *Drosophila melanogaster* [20], entrance to the competent state of *B. subtilis* [7], resistance of yeast colonies to antibiotic challenge [17], maintenance of HIV latency [21], promoting host infection by pathogens [22] or the induction of the lactose operon in *E. coli* [23]. Other examples have been found and reviewed elsewhere [24,25]. The overall conclusion of all of these studies is that stochasticity in gene expression can have important physiological consequences in natural and synthetic systems and that the overall architecture of the gene regulatory network can greatly affect the level of stochasticity.

A number of theoretical and experimental studies have revealed multiple ways in which the architecture of the gene regulatory network affects cell-to-cell variability in gene expression. Examples of mechanisms for the control of stochasticity have been proposed and tested, including the regulation of translational efficiency [8], the presence of negative feedback loops [26,27,28], or the propagation of fluctuations from upstream regulatory components [29]. Another important source of stochasticity in gene expression is fluctuations in promoter activity, caused by stochastic association and dissociation of transcription factors, chromatin remodeling events, and formation of stable pre-initiation complexes [5,15,16,23,30]. In particular, it has been reported that perturbations to the architecture of yeast and bacterial promoters, such as varying the strength of transcription factor binding sites[17], the number and location of such binding sites [11,31], the presence of auxiliary operators that mediate DNA looping [23], or the competition of activators and repressors for binding to the same stretch of DNA associated with the promoter [32], may strongly affect the level of variability.

Our goal is to examine all of these different promoter architectures from a unifying perspective provided by stochastic models of transcription leading to mRNA production. The logic here is the same as in earlier work where we examined a host of different promoter architectures using thermodynamic models of transcriptional regulation [33,34]. We now generalize those systematic efforts to examine the same architectures, but now from the point of view of stochastic models. These models allow us to assess the unique signature provided by a particular regulatory architecture in terms of the cell-to-cell variability it produces.

First, we investigate in general theoretical terms how the architecture of a promoter affects the level of cell-to-cell variability. The architecture of a promoter is defined by the collection of transcription factor binding sites (also known as operators), their number, position within the promoter, their strength, as well as what kind of transcription factors bind them (repressors, activators or both), and how those transcription factors bind to the operators (independently, cooperatively, simultaneously). We apply the master-equation model of stochastic gene expression [35,36, 37,38] to increasingly complex promoter architectures [30], and compute the moments of the mRNA and protein distributions expected for these promoters. Our results provide an expectation for how different architectural elements affect cell-to-cell variability in gene expression.

The second point of this paper is to make use of stochastic kinetic models of gene regulation to put forth in vivo tests of the molecular mechanisms of gene regulation by transcription factors that have been proposed as a result of in vitro biochemical experiments. The idea of using spontaneous fluctuations in gene expression to infer properties of gene regulatory circuits is an area of growing interest, given its non-invasive nature and its potential to reveal regulatory mechanisms in vivo. Different theoretical methods have recently been proposed, which could be employed to distinguish between different modes (e.g. AND/OR) of combinatorial gene regulation, and to rule out candidate regulatory circuits [27,39,40] based solely on properties of noise in gene expression, such as the autocorrelation function of the fluctuations [27] or the three-point steady state correlations between multiple inputs and outputs [39,40].

Here, we make experimentally testable predictions about the level of cell-to-cell variability in gene expression expected for different bacterial promoters, based on the physical kinetic models of gene regulation that are believed to describe these promoters in vivo. In particular, we focus on how varying the different parameters (i.e., mutating operators to make them stronger or weaker, varying the intracellular concentration of transcription factors, etc.) should affect the level of variability. This way, cell-to-cell variability in gene expression is used as a tool for testing kinetic models of transcription factor-mediated regulation of gene expression in vivo.

The remainder of the paper is organized as follows: First we describe the theoretical formalism we use to determine analytic expressions for the moments of the probability distribution for both mRNA and protein abundances per cell. Next, we examine how the architecture of the promoter affects cell-to-cell variability in gene expression. We focus on simple and cooperative repression, simple and cooperative activation, and transcriptional regulation by distal operators mediated by DNA looping. We investigate how noise in gene expression caused by promoter activation differs from repression, how operator multiplicity affects noise in gene expression, the effect of cooperative binding of transcription factors, as well as DNA looping. For each one of these architectures we present a prediction of cell-to-cell variability in gene expression for a bacterial promoter that has been well characterized experimentally in terms of their mean expression values. These predictions suggest a new round of experiments to test the current mechanistic models of gene regulation at these promoters.

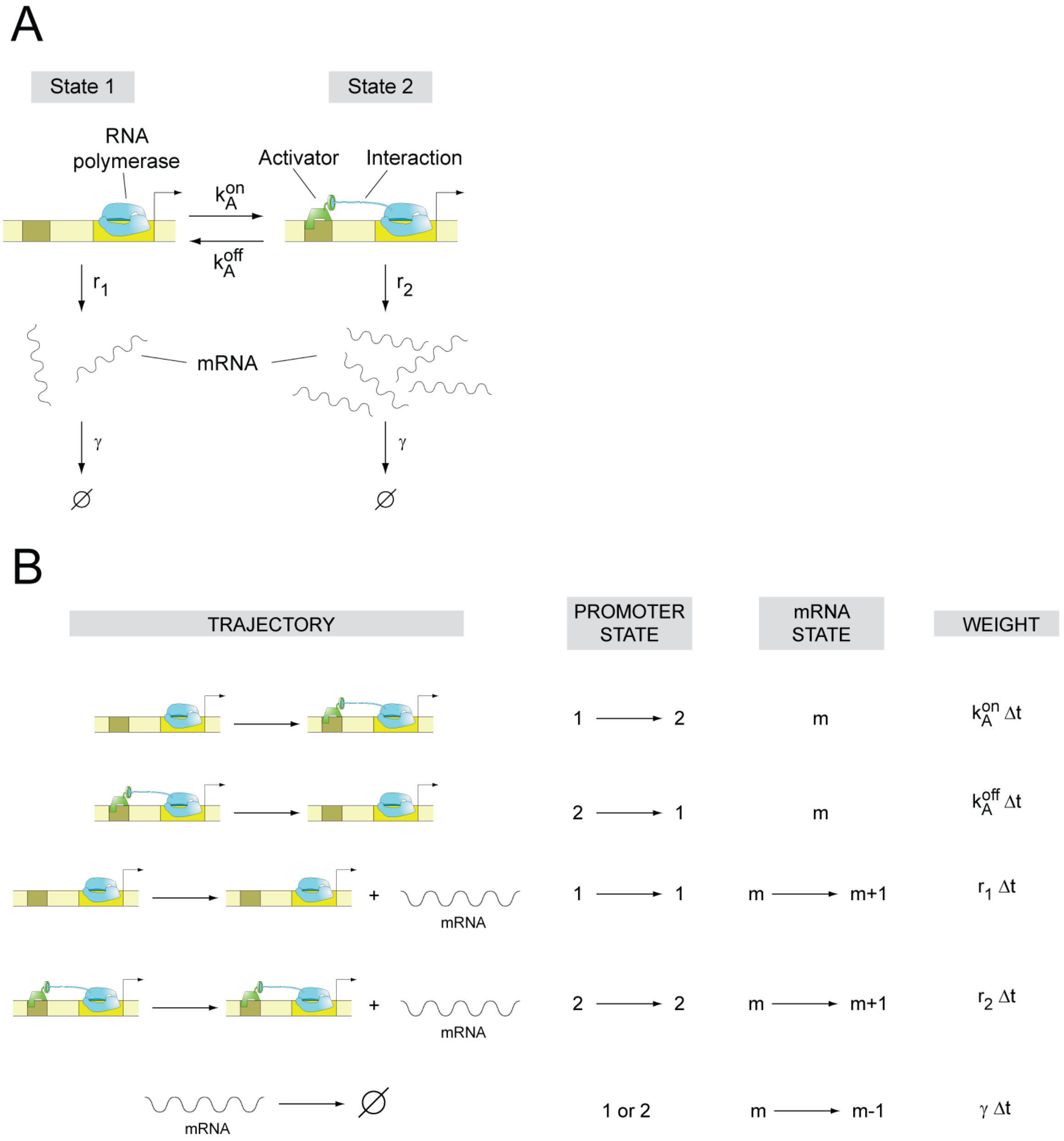
5.2 Methods

In order to investigate how promoter architecture affects cell-to- cell variability in gene expression, we use a model based on classical chemical kinetics (illustrated in Figure 5.1A), in which a promoter containing multiple operators may exist in as many biochemical states as allowed by the combinatorial binding of transcription factors to its operators. The promoter transitions stochastically between the different states as transcription factors bind and fall off. Synthesis of mRNA is assumed to occur stochastically at a constant rate that is different for each promoter state. Further, transcripts are assumed to be degraded at a constant rate per molecule.

This kind of model is the kinetic counterpart of the so-called thermodynamic model of transcriptional regulation [41], and it is the standard framework for interpreting the kinetics of gene regulation in biochemical experiments, both in vivo [2,23] and in vitro [42,43]. This class of kinetic models can easily accommodate stochastic effects, and it leads to a master equation from which the probability distribution of mRNA and protein copy number per cell can be computed. It is often referred to as the standard model of stochastic gene expression [38,44,45]. The degree of cell-to-cell variability in gene expression can be quantified by the stationary variance, defined as the ratio of the standard deviation and the mean of the probability distribution of mRNA or protein copy number per cell [35], or else by the Fano factor, the ratio between the variance and the mean. These two are the two most common metrics of noise in gene expression, and the relation between them will be discussed later.

In order to compute the noise strength from this class of models, we follow the same approach as in a previous article [30], which extends a master equation derived elsewhere [36,37,46] to promoters with arbitrary combinatorial complexity. The complexity refers to the existence of a number of discrete promoter states corresponding to different arrangements of transcription factors on the promoter DNA. Promoter dynamics are described by trajectories involving stochastic transitions between promoter states which are induced by the binding and unbinding of transcription factors. A detailed derivation of the equations which describe promoter dynamics can be found in the Text S1, but the essentials are described below.

There are only two stochastic variables in the model: the number of mRNA transcripts per cell, which is represented by the unitless state variable m, and the state of the promoter, which is defined by the pattern of transcription factors bound to their operator sites. The promoter state is described by a discrete and finite stochastic variable (s) (for an example, see Figure 5.1A). The example in Figure 5.1A illustrates the simplest model of transcriptional activation by a transcription factor. When the activator is not bound (state 1), mRNA is synthesized at rate r_1. When the activator is



TRAJECTORY	PROMOTER STATE	mRNA STATE	WEIGHT
→	1 → 2	m	$k_A^{on}\ \Delta t$
→	2 → 1	m	$k_A^{off}\ \Delta t$
→ + mRNA	1 → 1	m → m+1	$r_1\ \Delta t$
→ + mRNA	2 → 2	m → m+1	$r_2\ \Delta t$
mRNA → $\varnothing$	1 or 2	m → m-1	$\gamma\ \Delta t$

Figure 5.1: **Two-state promoter**. (A) Simple two-state bacterial promoter undergoing stochastic activation by a transcriptional activator binding to a single operator site. The rates of activator association and dissociation are given by $k_{\mathrm{A}}^{\mathrm{on}}$ and $k_{\mathrm{A}}^{\mathrm{off}}$, respectively and the rates of mRNA production for the basal and active states are r_1 and r_2 respectively. The mRNA degradation rate is assumed to be constant for each molecule, and is given by the parameter γ. (B) List of all possible stochastic transitions affecting either the copy number of mRNA (m) or the state of the promoter (s) and their respective statistical weights. State 1 has the operator free. State 2 is the activator bound state. The weights represent the probability that each change of state will occur during a time increment Δt. The master equation is constructed based on these rules.

bound to the promoter (state 2), mRNA is synthesized at the higher rate r_2. The promoter switches stochastically from state 1 to state 2 with rate $k_{\mathrm{A}}^{\mathrm{on}}$, and from state 2 to state 1 with rate $k_{\mathrm{A}}^{\mathrm{off}}$. Each mRNA molecule is degraded with rate γ.

The time evolution for the joint probability of having the promoter in states 1 or 2, with m mRNAs in the cell (which we write as $p(1, m)$ and $p(2, m)$, respectively), is given by a master equation, which we can build by listing all possible reactions that lead to a change in cellular state, either by changing m or by changing s (Figure 5.1b). The master equation takes the form:

$$\frac{d}{dt}p(1,m) = -k_{\mathrm{A}}^{\mathrm{on}}p(1,m) + k_{\mathrm{A}}^{\mathrm{off}}p(2,m) - r_1 p(1,m) - \gamma m p(1,m) + \\ r_1 p(1,m-1) + \gamma(m+1)p(1,m+1), \tag{5.1}$$

$$\frac{d}{dt}p(2,m) = k_{\mathrm{A}}^{\mathrm{on}}p(1,m) - k_{\mathrm{A}}^{\mathrm{off}}p(2,m) - r_2 p(2,m) - \gamma m p(2,m) + \\ r_2 p(2,m-1) + \gamma(m+1)p(2,m+1). \tag{5.2}$$

Inspecting this system of equations, we notice that by defining the vector:

$$\vec{p}(m) = \begin{pmatrix} p(1,m) \\ p(2,m) \end{pmatrix}, \tag{5.3}$$

and the matrices

$$\hat{K} = \begin{bmatrix} -k_{\mathrm{A}}^{\mathrm{on}} & k_{\mathrm{A}}^{\mathrm{off}} \\ k_{\mathrm{A}}^{\mathrm{on}} & -k_{\mathrm{A}}^{\mathrm{off}} \end{bmatrix}, \tag{5.4}$$

$$\hat{R} = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix}, \tag{5.5}$$

and

$$\hat{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.6}$$

we can rewrite the system of equations 5.1 and 5.2 in matrix form.

$$\frac{d}{dt}\vec{p}(m) = \left[\hat{K} - \hat{R} - m\gamma\hat{I}\right]\vec{p}(m) + \hat{R}\vec{p}(m-1) + (m+1)\gamma\hat{I}\vec{p}(m+1). \tag{5.7}$$

This has several advantages, but the most important one is that the matrix approach reduces the task of obtaining analytical expressions for the moments of the steady state mRNA distribution for an arbitrarily complex promoter to solving two simple linear matrix equations (more details are given in the Text S1).

The matrices appearing in equation 5.7 all have simple and intuitive interpretations. The matrix $\hat{K}$ describes the stochastic transitions between promoter states: the off-diagonal elements of the matrix $\hat{K}_{ij}$ are the rates of making transitions from promoter state j to promoter state i. The diagonal elements of the matrix $\hat{K}_{ij}$ are negative, and they represent the net probability flux out of state j: $\hat{K}_{ij} = \sum_{i \neq j} -\hat{K}_{ij}$. The matrix $\hat{R}$ is a diagonal matrix whose element $\hat{R}_{jj}$ gives the rate of transcription initiation when the promoter is in state j. Finally, the matrix $\hat{I}$ is the identity matrix.

An example of matrices $\hat{K}$ and $\hat{R}$ is presented pictorially in Figure 1 in Text S1. It is straightforward to see that even though equation 5.7 has been derived for a two-state promoter, it also applies to any other promoter architecture. What will change for different architectures are the dimensions of the matrices and vectors (these are given by the number of promoter states) as well as the values of the rate constants that make up the matrix elements of the various matrices.

An important limit of the master equation, which is often attained experimentally, is the steady state limit, where the probability distribution for mRNA number per cell does not change with time. Although the time dependence of the moments of the mRNA distribution can be easily computed from our model, for the sake of simplicity and because most experimental studies have been performed on cells in steady state, we focus on this limit. As shown in Text S1, analytic expressions for the first two moments of the steady state mRNA probability distribution are found by multiplying both sides of equation 5.7 by m and m^2 respectively, and then summing m from zero to infinity. After some algebra (elaborated in an earlier paper and in Text S1), we find that the first two moments can be written as:

$$\langle m \rangle = \frac{\vec{r} \cdot \vec{m}_{(0)}}{\gamma}, \tag{5.8}$$

$$\langle m^2 \rangle = \langle m \rangle + \frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma}. \tag{5.9}$$

The vector $\vec{r}$ contains the ordered list of rates of transcription initiation for each promoter state. For the two-state promoter shown in Figure 5.1, $\vec{r} = (r_1, r_2)$. The vector $\vec{m}_{(0)}$ contains the steady state probabilities for finding the promoter in each one of the possible promoter states, while $\vec{m}_{(1)}$ is the steady-state mean mRNA number in each promoter state. The vector $\vec{m}_{(0)}$ is the solution to the matrix equation

$$\hat{K}\vec{m}_{(0)} = 0, \tag{5.10}$$

while the vector $\vec{m}_{(1)}$ is obtained from

$$(\hat{K} - \gamma\hat{I})\vec{m}_{(1)} + \hat{R}\vec{m}_{(0)} = 0. \tag{5.11}$$

Figure 5.1 illustrates the following algorithm for computing the intrinsic variability of mRNA number for promoter of arbitrarily complex architecture:

1. Make a list of all possible promoter states and their kinetic transitions (Figure 5.1B)

2. Construct the matrices $\hat{K}$ and $\hat{R}$, and the vector $\vec{r}$ (Figure 1 in Text S1).

3. Solve equations 5.10-5.11 to obtain $\vec{m}_{(0)}$ and $\vec{m}_{(1)}$.

4. Plug solutions of 5.10-5.11 into equations 5.8-5.9 to obtain the moments.

The normalized variance of the mRNA distribution in steady state is then computed from the equation

$$\eta^2 = \frac{\mathrm{Var}(m)}{\langle m \rangle^2} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle^2} = \frac{1}{\langle m \rangle} + \frac{1}{\langle m \rangle^2} \left(\frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma} - \langle m \rangle^2 \right). \tag{5.12}$$

Equation 5.12 reveals that, regardless of the specific details characterizing promoter architecture, the intrinsic noise is always the sum of two components, and it can be written as

$$\eta^2 = \frac{1}{\langle m \rangle} + \eta^2_{\text{promoter}}. \tag{5.13}$$

The first component is due to spontaneous stochastic production and degradation of single mRNA molecules, is always equal to the Poissonian expectation of $1/\langle m \rangle$, and is independent of the architecture of the promoter. For an unregulated promoter that is always active and does not switch between multiple states (or does so very fast compared to the rates of transcription and mRNA degradation), the mRNA distribution is well described by a Poisson distribution [45,47], and the normalized variance is equal to $1/\langle m \rangle$. The second component ("promoter noise") results from promoter state fluctuations, and captures the effect of the promoters architecture on the cell-to-cell variability in mRNA:

$$\eta^2_{\text{promoter}} = \frac{1}{\langle m \rangle^2} \left(\frac{\vec{r} \cdot \vec{m}_{(1)}}{\gamma} - \langle m \rangle^2 \right). \tag{5.14}$$

In order to quantify the effect of the promoter architecture in the level of cell-to-cell variability in mRNA expression, we define the deviation in the normalized variance caused by gene regulation relative to the baseline Poisson noise for the same mean (see Figure 2):

$$\text{Fold-change mRNA noise} = \frac{\eta^2}{\eta^2_{\text{Poisson}}} \tag{5.15}$$

$$= \frac{\mathrm{Var}(m)/\langle m \rangle^2}{1/\langle m \rangle} = \frac{\mathrm{Var}(m)}{\langle m \rangle}. \tag{5.16}$$

Therefore, the deviation in the normalized variance caused by gene regulation is equal to the ratio between the variance and the mean. This parameter is also known as the Fano factor. Thus, for any given promoter architecture, the Fano factor quantitatively characterizes how large the mRNA noise is relative to that of a Poisson distribution of the same mean (i.e. how much the noise for the regulated promoter elevates with respect to the Poisson noise). This is the parameter that we will use throughout the paper as the metric of cell-to-cell variability in gene expression.

5.2.1 Promoter noise and variability of mRNA and protein numbers

For proteins, the picture is only slightly more complicated. As shown in the Text S1, in the limit where the lifetime of mRNA is much shorter than that of the protein it encodes for (a limit that is often fulfilled [30]), the noise strength of the probability distribution of proteins per cell takes the following form (where we define n as a state variable that represents the copy number of proteins per cell):

$$\frac{\mathrm{Var}(n)}{\langle n\rangle^2} = \frac{\langle n^2\rangle - \langle n\rangle^2}{\langle n\rangle^2} = \frac{1+b}{\langle n\rangle} + \frac{1}{\langle n\rangle^2}\left(b\frac{\vec{r}\cdot\vec{n}_{(1)}}{\delta} - \langle n\rangle^2\right), \qquad (5.17)$$

where δ stands for the protein degradation rate, and the constant b is equal to the protein burst size (the average number of proteins produced by one mRNA molecule). The mean protein per cell is given by

$$\langle n\rangle = b\frac{\vec{r}\cdot\vec{m}_{(0)}}{\delta}, \qquad (5.18)$$

and the vector $\vec{n}_{(1)}$ is the solution to the algebraic equation:

$$(\hat{K} - \delta\hat{I})\vec{n}_{(1)} + b\hat{R}\vec{m}_{(0)} = 0. \qquad (5.19)$$

The reader is referred to the Text S1 for a detailed derivation and interpretation of these equations. In the previous section we have shown that the noise for proteins and mRNA take very similar analytical forms. Indeed, if we define $\vec{r}_n = b\vec{r}$ and $\hat{R}_n = b\hat{R}$ as the vector and matrix containing the average rates of protein synthesis for each promoter state, it is straightforward to see that equations 5.11 and 5.19 are mathematically equivalent, with the only difference being that in equation 5.19, the matrix $\hat{R}_n$ represents the rates of protein synthesis, so all the rates of transcription are multiplied by the translation burst size b. Therefore, the vectors $\vec{m}_{(0)}$ and $\vec{m}_{(1)}$ are only going to differ in the prefactor b multiplying all the different transcription rates. We conclude that the promoter contribution to the noise takes the exact same analytical form both for proteins and for mRNA, with the only other quantitative difference being the different rates of degradation for proteins and mRNA. Therefore, promoter architecture has the same qualitative effect on cell-to-cell variability in mRNA and protein numbers. All the conclusions about the effect of promoter architecture on cell-to-cell variability in mRNA expression are also valid for proteins, even though quantitative differences do generally exist. For the sake of simplicity we focus on mRNA noise for the remainder of the paper.

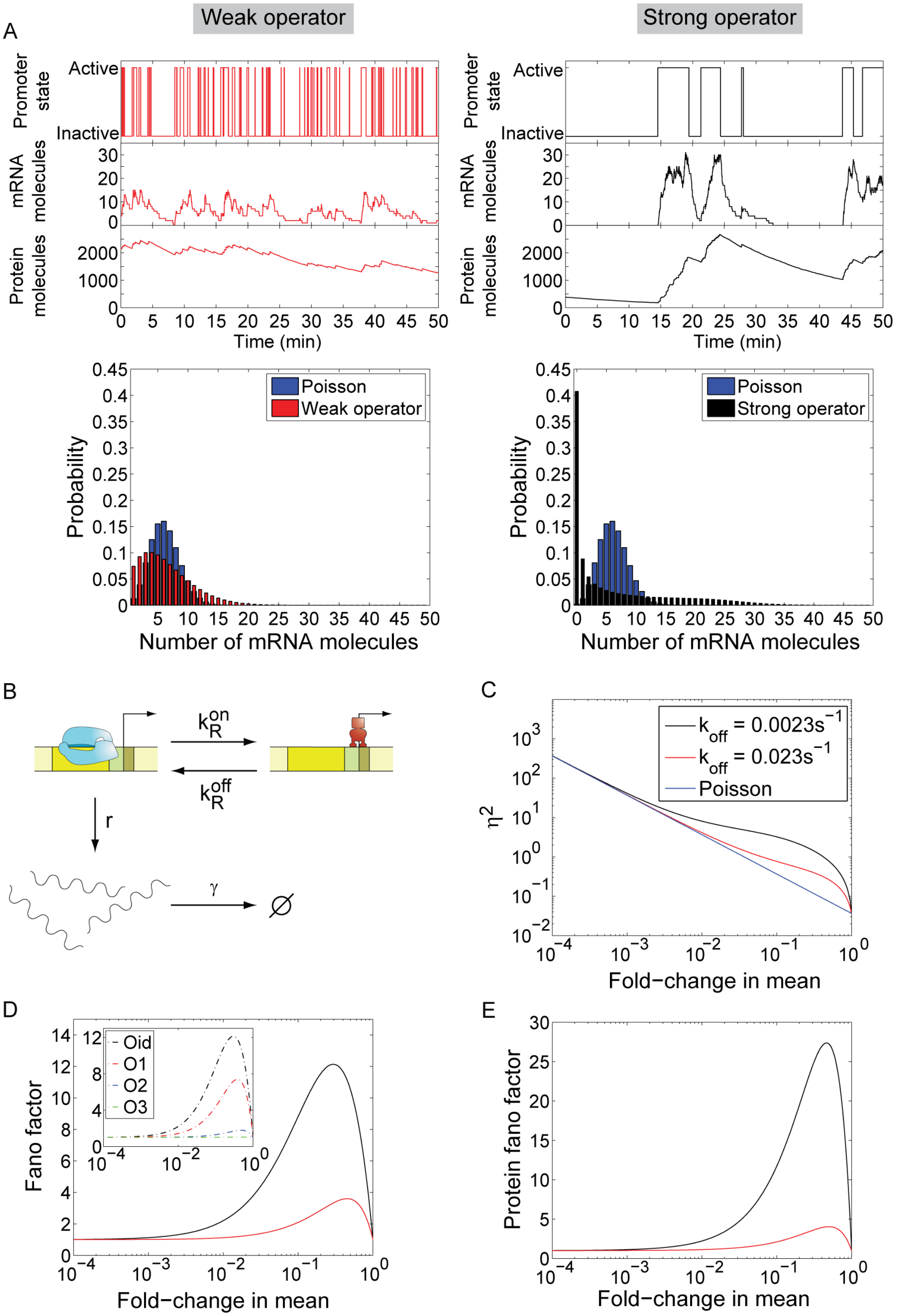


Figure 5.2: **Simple repression architecture.** (Caption continues on next page.)

Figure 5.2: (A) Time traces for promoter activity, mRNA and protein copy number are shown for both the weak operator and the strong operator. The mRNA histograms are also shown. The weaker operator with a faster repressor dissociation rate leads to small promoter noise and an mRNA probability distribution resembling a Poisson distribution (shown by the blue-bar histogram), in which most cells express mRNA near the population average. In contrast, the stronger operator with a slower repressor dissociation rate leads to larger promoter noise and strongly non-Poissonian mRNA statistics. (B) Kinetic mechanism of repression for an architecture involving a single repressor binding site. The repressor turns off the gene when it binds to the promoter (with rate $k_{\mathrm{R}}^{\mathrm{on}}$), and transcription occurs at a constant rate r when the repressor falls off (with rate $k_{\mathrm{R}}^{\mathrm{off}}$). (C) Normalized variance as a function of the fold-change in mean mRNA copy number. The parameters used are drawn from Table 1. The value of $k_{\mathrm{R}}^{\mathrm{off}} = 0.0023\mathrm{s}^{-1}$ from Table 1 corresponds to the in vitro dissociation constant of the Lac repressor from the Oid operator (black). The results for an off-rate 10 times higher are also plotted (red). As a reference for the size of the fluctuations, we show the normalized variance for a Poisson promoter. (D) Fano factor for two promoters bearing the same off-rates as in (B). Inset. Prediction for the Fano factor for the Δ_{O3} Δ_{O2} $\mathrm{P}_{\mathrm{lacUV5}}$ promoter, a variant of the $\mathrm{P}_{\mathrm{lacUV5}}$ promoter for which the two auxiliary operators have been deleted. The fold-change in mRNA noise is plotted as a function of the fold-change in mean mRNA copy number for mutants of the promoter that replace O1 for Oid, O2 or O3. The parameters are taken from Table 1 and [33]. Lifetimes of the operator-repressor complex are 7 min for Oid, 2.4 min for O1, 11s for O2 and 0.47 s for O3. (E) Fold-change in protein noise as a function of the fold-change in mean expression. As expected, the effect of operator strength is the same as observed for mRNA noise.

5.2.2 Parameters and assumptions

In order to evaluate the equations in our model, we use parameters that are consistent with experimental measurements of rates and equilibrium constants *in vivo* and *in vitro*, which we summarize in Table 1. Although these values correspond to specific examples of *E. coli* promoters, like the $\mathrm{P}_{\mathrm{lac}}$ or the P_{RM} promoter, we extend their reach by using them as “typical” parameters characteristic of bacterial promoters, the idea being that we are trying to demonstrate the classes of effects that can be expected, rather than dissecting in detail any particular promoter. The rate of association for transcription factors to operators *in vivo* is assumed to be the same as the recently measured value for the Lac repressor, which is close to the diffusion limited rate [48]. In order to test whether the particular selection of parameters in Table 1 is biasing our results, we have also done several controls (see Figures 24 in Text S1) in which the kinetic parameters were randomly sampled. We found that the conclusions reached for the set of parameters in Table 1 are valid for other parameter sets as well.

Operator strength reflects how tightly operators bind their transcription factors, and it is quantitatively characterized by the equilibrium dissociation constant K_{O-TF}. The dissociation constant has units of concentration and is equal to the concentration of free transcription factor at which the probability for the operator to be occupied is one half. K_{O-TF} is related to the association and dissociation rates by $K_{O-TF} = k_{\mathrm{off}}/k_{\mathrm{on}}^{0}$, where k_{off} is the rate (*i.e.*, the probability per unit time) at which a transcription factor dissociates from the promoter, and k_{on}^{0} is a second order rate constant, which represents the association rate per unit of concentration of transcription factors, *i.e.*, $k_{\mathrm{on}} = k_{\mathrm{on}}^{0}[N_{TF}]$. Note that in the last formula, k_{on}, which has units of s^{-1}, is written as a

Table 5.1: Kinetic parameters used to the make the quantitative estimates in the text and plots in the figures.

Kinetic rate	**Symbol**	**Value**	**Reference**
Unregulated promoter transcription rate	r	0.33 s^{-1}	[99]
Repressor and activator association rates	k_R^0, k_A^0	0.0027 $(\text{s nM})^{-1}$	[2]
Repressor and activator dissociation rates	$k_{\text{R}}^{\text{off}}$, $k_{\text{A}}^{\text{off}}$	$0.0023\,\text{s}^{-1}$	[42]
mRNA decay rate	γ	$0.011\,\text{s}^{-1}$	[10]
Ratio between transcription rates due to activation	$f = r_1/r_2$	11	[50]
Cooperativity in repression	$\Omega_{\text{repression}}$	0.013	[50]
Cooperativity in activation	$\Omega_{\text{activation}}$	0.1	[33]
Looping J-factor	$[J]$	660 nM	[33]
Protein translation burst size	b	31.2 proteins/mRNA	[5]
Protein decay rate	δ	$0.00083\,\text{s}^{-1}$	[99]

product of two quantities: $[N_{TF}]$, which is the concentration (in units of (mol/liter)) of transcription factors inside the cell, and k_{on}^0, a second order rate constant that has units of $(\text{mol/liter})^{-1}\text{s}^{-1}$. For simplicity, we assume that the binding reaction is diffusion limited; namely, k_{on}^0 is already close to its maximum possible value, so the only parameter that can differ from operator to operator is the dissociation rate: strong operators have slow dissociation rates, and weak operators have large dissociation rates.

Throughout this paper, we also make the assumption that the mean expression level is controlled by varying the intracellular concentration of transcription factors, a scenario that is very common experimentally [49,50,51]. We also assume that changing the intracellular concentration of transcription factors only affects the association rate of transcription factors to the operators, but the dissociation rate and the rates of transcription at each promoter state are not affected. In other words, k_{off} is a constant parameter for each operator, and it is not changed when we change the mean by titrating the intracellular repressor level. All of these general assumptions need to be revisited when studying a specific gene-regulatory system. Here our focus is on illustrating the general principles associated with different promoter architectures typical of those found in prokaryotes.

5.2.3 Simulations

To generate mRNA time traces, we applied the Gillespie algorithm [52] to the master equation described in the text. A single time step of the simulation is performed as follows: one of the set of possible trajectories is chosen according to its relative weight, and the state of the system is updated appropriately. At the same time, the time elapsed since the last step is chosen from an exponential distribution, whose rate parameter equals the sum of rate parameters of all possible trajectories. This process is repeated iteratively to generate trajectories that exactly reflect dynamics of the underlying master equation. For the figures, simulation lengths were set long enough for the system to reach steady state and for several promoter state transitions to occur.

To generate the probability distributions, it is convenient to reformulate the entire system of mRNA master equations in terms of a single matrix equation. To do this, we first define a vector

$$\vec{P} = \begin{pmatrix} p(1,0) \\ p(2,0) \\ \vdots \\ p(N,0) \\ p(1,1) \\ \vdots \\ p(N,1) \\ p(1,2) \\ \vdots \\ p(N,2) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vec{p}(0) \\ \vec{p}(1) \\ \vec{p}(2) \\ \vdots \end{pmatrix}, \tag{5.20}$$

where $p(i,m)$ is the probability of having m mRNAs and being in the ith promoter state. Then the master equation for time evolution of this probability vector is

$$\frac{d\vec{P}}{dt} = \begin{pmatrix} \hat{K}-\hat{R} & \gamma\hat{I} & 0 & \cdots \\ \hat{R} & \hat{K}-(\hat{R}+\gamma\hat{I}) & 2\gamma\hat{I} & \cdots \\ 0 & \hat{R} & \hat{K}-(\hat{R}+2\gamma\hat{I}) & \cdots \\ 0 & 0 & \hat{R} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \vec{p}(0) \\ \vec{p}(1) \\ \vec{p}(2) \\ \vec{p}(3) \\ \vdots \end{pmatrix}, \tag{5.21}$$

where each element of the matrix is itself an N by N matrix as described in the text. Then finding the steady-state distribution $\vec{P}_{ss}$ is equivalent to finding the eigenvector of the above matrix associated with eigenvalue 0. To perform this calculation numerically, one must first choose an upper bound on mRNA copy number in order to work with finite matrices. In this work, we chose an upper bound six standard deviations above mean mRNA copy number as an initial guess, and then modified this bound if necessary. Computations were performed using the SciPy (Scientific Python) software package.

5.3 Results

5.3.1 Promoters with a single repressor binding site

We first investigate a promoter architecture consisting of a single repressor binding site, and examine how operator strength affects intrinsic variability in gene expression. Although this particular mode

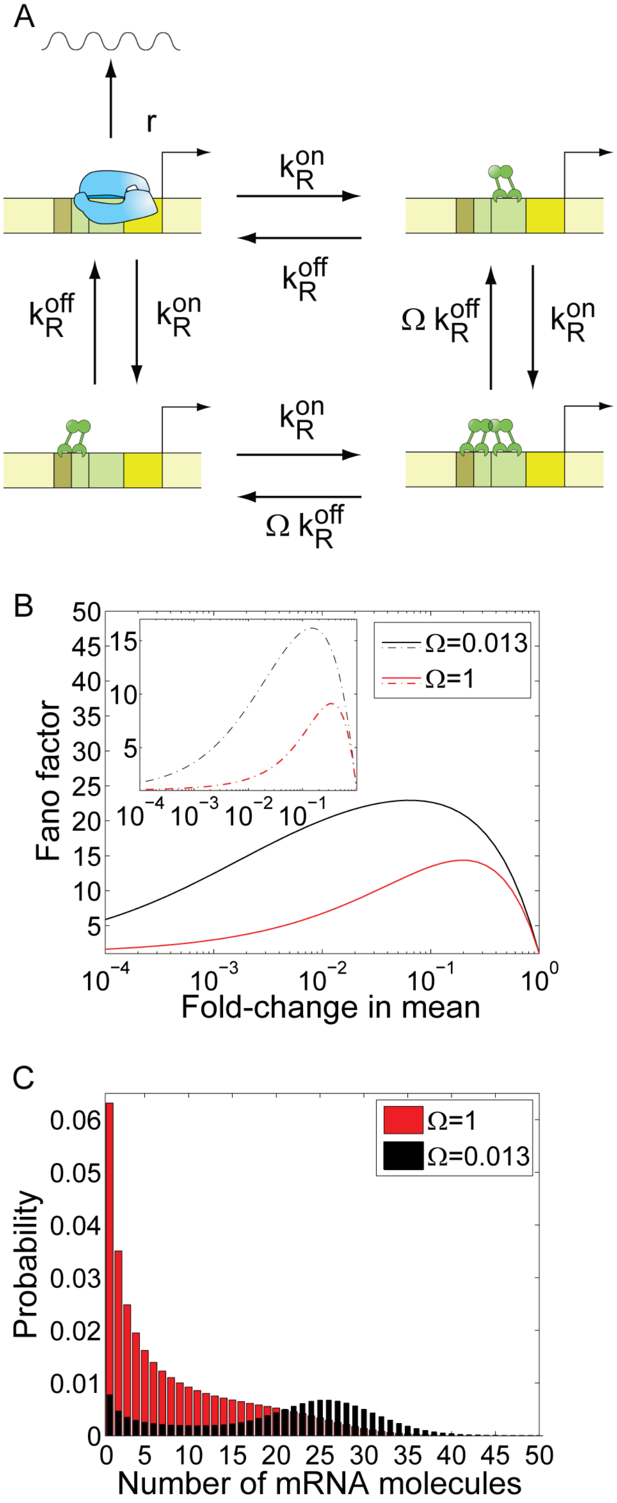


Figure 5.3: **Dual repression architecture.** (A) Kinetic mechanism of repression for a dual-repression architecture. The parameters $k_{\mathrm{R}}^{\mathrm{off}}$ and $k_{\mathrm{R}}^{\mathrm{on}}$ are the rates of repressor dissociation and association to the operators, and Ω is a parameter reflecting the effect of cooperative binding on the dissociation rate. For independent binding, $\Omega = 1$ and for cooperative binding $\Omega = 0.013$ (see Table 1). (B) Fold-change in the mRNA noise caused by gene regulation for independent (red) and cooperative (black) repression as a function of the mean mRNA copy number. Inset: Prediction for a variant of the λ P$_{\mathrm{R}}$ promoter where the upstream operators O$_{\mathrm{L1}}$, O$_{\mathrm{L2}}$, and O$_{\mathrm{L3}}$ are deleted. The promoter mRNA noise is plotted as a function of the mean mRNA number for both wild-type cI repressor (blue line) and a repressor mutant (Y210H) that abolishes cooperativity (red line). Parameters taken from [43,97]. The lifetime of the O$_{\mathrm{R1}}$-cI complex is 4 min. Lifetime of O$_{\mathrm{R2}}$-cI complex is 9.5s. (C) mRNA distribution for the same parameters used in (B).

of gene regulation has been well studied theoretically before [1,16,36,37,45], it is a useful starting point for illustrating the utility of this class of models. Within this class of models, when the repressor is bound to the operator, it interferes with transcription initiation and transcription does not occur. When the repressor dissociates and the operator is free, RNAP can bind and initiate transcription at a constant rate r. The probability per unit time that a bound repressor dissociates is $k_{\mathrm{R}}^{\mathrm{off}}$, and the probability per unit time that a free repressor binds the empty operator is $k_{\mathrm{R}}^{\mathrm{on}} = k_{\mathrm{on}}^{0}[N_R]$, where k_{on}^{0} is the second-order association constant and $[N_R]$ is the intracellular repressor concentration. The rate of mRNA degradation per molecule is γ. This mechanism is illustrated in Figure 5.2B.

We compute the mean and the Fano factor for this architecture following the algorithm described in Text S1. The kinetic rate and transcription rate matrices $\hat{K}$ and $\hat{R}$ are shown in Table S1 in Text S1. For this simple architecture, the mean of the mRNA probability distribution and the normalized variance take simple analytical forms:

$$\langle m \rangle = \frac{r}{\gamma} \frac{k_{\mathrm{R}}^{\mathrm{off}}}{k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}} = \frac{r}{\gamma} \frac{1}{1 + k_{\mathrm{R}}^{\mathrm{on}}/k_{\mathrm{R}}^{\mathrm{off}}}, \tag{5.22}$$

$$\eta^2 = \frac{1}{\langle m \rangle} + \frac{k_{\mathrm{R}}^{\mathrm{off}}}{k_{\mathrm{R}}^{\mathrm{on}}} \frac{\gamma}{1 + k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}}. \tag{5.23}$$

Using the relationship between $k_{\mathrm{R}}^{\mathrm{on}}$ and the intracellular concentration of repressor, we can write the mean as:

$$\langle m \rangle = \frac{r}{\gamma} \frac{1}{1 + k_{\mathrm{on}}^{0}[N_R]/k_{\mathrm{R}}^{\mathrm{off}}} = \langle m \rangle_{\max} \frac{1}{1 + [N_R]/K_{OR}}. \tag{5.24}$$

Here we have defined the equilibrium dissociation constant between the repressor and the operator as $K_{OR} = k_{\mathrm{R}}^{\mathrm{off}}/k_{\mathrm{on}}^{0}$. It is interesting to note that equation 5.24 could have been derived using the thermodynamic model approach [33,34,41,53]. In particular we see that this expression is equal to the product of the maximal activity in the absence of repressor $\langle m \rangle_{\max} = r/\gamma$, and the so-called fold-change in gene expression [34]:

$$\text{fold-change} = (1 + k_{\mathrm{R}}^{\mathrm{on}}/k_{\mathrm{R}}^{\mathrm{off}})^{-1} = (1 + [N_R]/K_{OR})^{-1}. \tag{5.25}$$

The fold-change is defined as the ratio of the level of expression in the presence of the transcription factor of interest, and the level of expression in the absence of the transcription factor.

The Fano factor for the mRNA distribution can be computed from equation 5.16, and we obtain:

$$\text{Fano} = 1 + \left(\frac{k_{\mathrm{R}}^{\mathrm{on}}}{k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}} \right) \frac{r}{\gamma + k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}}, \tag{5.26}$$

which is also shown as the first entry of Table S2 in Text S1. In many experiments [4,15,31,50], the concentration of repressor $[N_R]$ (and therefore the association rate $k_{\mathrm{R}}^{\mathrm{on}} = k_{\mathrm{on}}^{0}[N_R]$) can be varied by either expressing the repressor from an inducible promoter, or by adding an inducer that binds

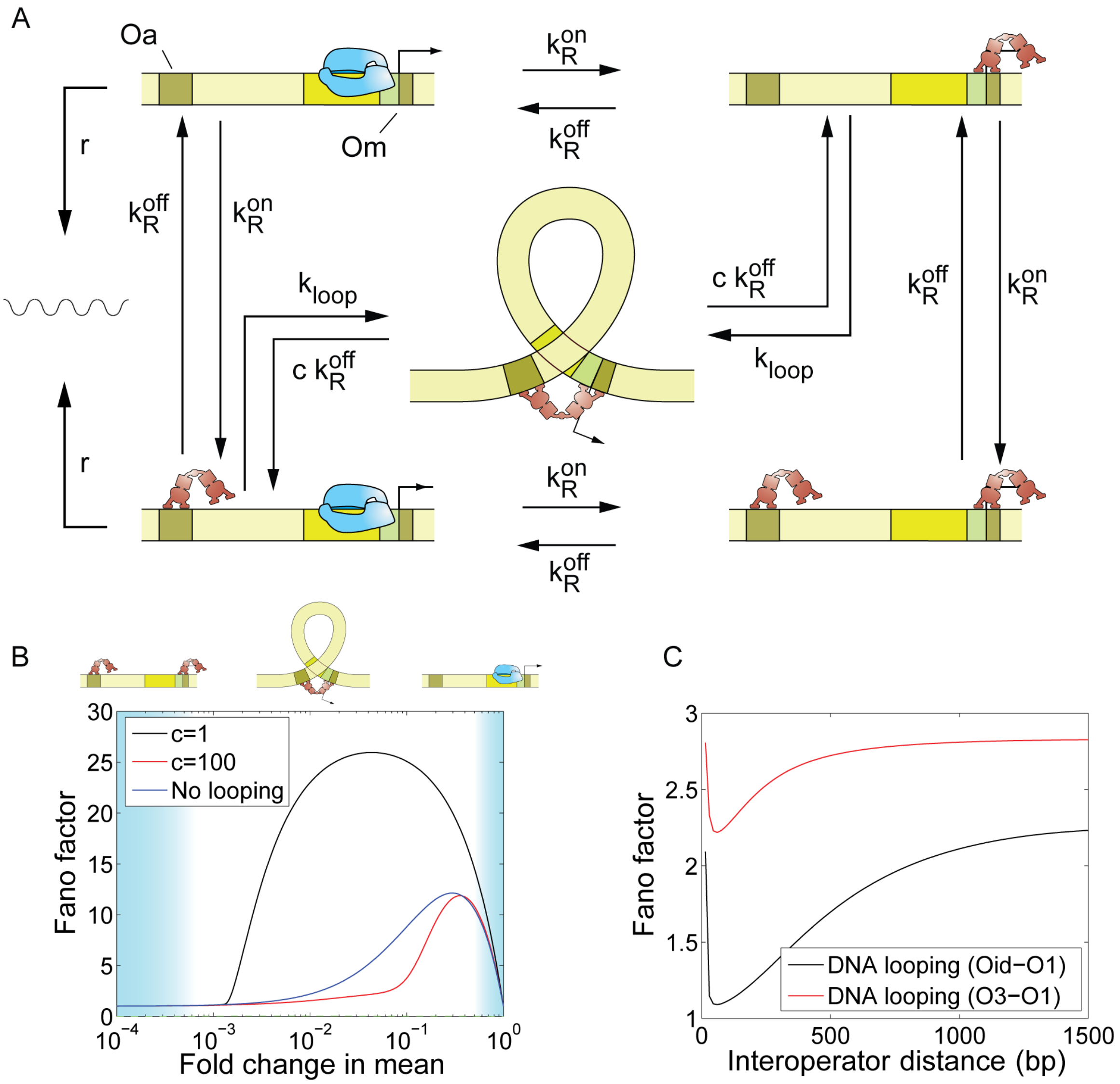


Figure 5.4: **Repression by DNA looping.** (A) Kinetic mechanism of repression. $k_{\mathrm{R}}^{\mathrm{off}}$ and $k_{\mathrm{R}}^{\mathrm{on}}$ are the rates of repressor dissociation and association. The rate of loop formation is $k_{\mathrm{loop}} = [J]k_R^0$, where $[J]$ can be thought of as the local concentration of repressor in the vicinity of one operator when it is bound to the other operator. The rate of dissociation of the operator-repressor complex in the looped conformation is given by $k_{\mathrm{unloop}} = ck_{\mathrm{R}}^{\mathrm{off}}$. The parameter c captures the rate of repressor dissociation in the looped state relative to the rate of dissociation in a non-looped state. (B) Effect of DNA looping on cell-to-cell variability. The Fano factor is plotted as a function of the fold-change in the mean expression level, in the absence (blue) and presence (black) of the auxiliary operator, and assuming that dissociation of the operator from Om is the same in the looped and the unlooped state ($c = 1$). The presence of the auxiliary operator, which enables repression by DNA looping, increases the cell-to-cell variability. The regions over which the state with two repressors bound, the state with one repressor bound, or the looped DNA state are dominant are indicated by the shading in the background. The noise is larger at intermediate repression levels, where only one repressor is found bound to the promoter region, simultaneously occupying the auxiliary and main operators through DNA looping. (Caption continues on next page.)

Figure 5.4: (Caption continued from previous page.) The rate of DNA loop formation is $k_{\text{loop}}(660\,\text{nM})k_R^0$ [33]. We also show the effect of DNA looping in the case where the kinetics of dissociation from the looped state are 100 times faster than the kinetics of dissociation from the unlooped state: $c = k_{\text{unloop}}/k_{\text{R}}^{\text{off}} = 100$ (red). In this limit, the presence of the auxiliary operator leads to less gene expression noise. (C) Prediction for a library of P_{lacUV5} promoter variants, harboring an O2 deletion, and with the position of O3 moved upstream by multiples of 11 bp while keeping its identity (red), or replaced by the operator by Oid (black). Parameters are taken from the analysis in [33] of the data in [98]. We assume a concentration of 50 Lac repressor tetramers per cell. The association rate of the tetrameric repressor to the operators is taken from Table 1. The lifetimes of the operator-repressor complex are given in the caption to Figure 5.2. The dependence of the rate of DNA looping on the inter-operator distance is taken from [33], and equal to: $k_{\text{loop}} = k_{\text{R}}^{\text{on}} \times \exp\left[-\frac{u}{D} - v\log(D) + wD + z\right]$, where $u = 140.6$, $v = 2.52$, $w = 0.0014$, and $z = 19.9$. Note that the Fano factor is not plotted as a function of the mean, but as a function of the inter-operator distance D. In this case, as we change D, we vary both the mean and the Fano factor.

directly to the repressor rendering it incapable of binding specifically to the operators in the promoter region. When such an operation is performed, the only parameter that is varied is typically k_{R}^{on}, and all other kinetic rates are constant. The Fano factor can thus be rewritten as a function of the mean mRNA, and we obtain:

$$\text{Fano} = 1 + \langle m \rangle \left(\frac{1 - \langle m \rangle / \langle m \rangle_{\max}}{k_{\text{R}}^{\text{off}}/\gamma + \langle m \rangle / \langle m \rangle_{\max}} \right). \tag{5.27}$$

Therefore, for any given value of the mean, the Fano factor depends only on two parameters: the maximal mRNA or protein expression per cell, and a parameter that reflects the strength of binding between the repressor and the operator: $k_{\text{R}}^{\text{off}}/\gamma$. Equations 5.24 and 5.27 reveal that changes in the mean due to repressor titration affect the noise as well as the mean. Since neither the repressor dissociation rate $k_{\text{R}}^{\text{off}}$ nor the mRNA degradation rates are affected by the concentration of repressor, $k_{\text{R}}^{\text{off}}/\gamma$ is a constant parameter that will determine how large the cell-to-cell variability is: the Fano factor is maximal for promoter with very strong operators ($k_{\text{R}}^{\text{off}} << 1$), and it goes to one (*i.e.*, the distribution tends to a Poisson distribution) when the operator is very weak and the rate of dissociation extremely fast ($k_{\text{R}}^{\text{off}} >> 1$). In the latter limit of fast promoter kinetics, the fast fluctuations in promoter occupancy are filtered by the long lifetime of mRNA. Effectively, mRNA degradation acts as a low-pass frequency filter [54,55], and fast fluctuations in promoter occupancy are not propagated into mRNA fluctuations. Therefore, promoters with strong operators are expected to be noisier than promoters with weak operators [56]. From this discussion it should also be clear that the mRNA degradation rate critically affects cell-to-cell variability. Any processes that tend to accelerate degradation will tend to increase noise, and mRNA stabilization (i.e., protection of the transcript by RNA binding proteins) leads to reduction of variability. However, the focus of this article is on promoter architecture and transcriptional regulation. Therefore, we do not consider regulation of transcription by mRNA degradation, and assume that all the promoters transcribe the same mRNA as is often the case in experimental studies.

The effect of operator strength on the output of transcription and translation is illustrated in Figure 5.2A, where we show results from a stochastic simulation of the model depicted in Figure 5.2B, for the case of a weak and a strong operator. The simulation yields trajectories in time for the promoter state, the mRNA, and protein number, as well as the steady state distribution of mRNA number. Concentrations of repressor in the simulations were chosen so that the mean expression level was equal for the two different promoter architectures. As expected from the general arguments presented above, we clearly see that the level of variability is smaller for the weak operator than for the strong operator, due to faster promoter switching leading to smaller mRNA fluctuations and a more Poisson-like mRNA distribution (Figure 5.2A, weak promoter). Slow dissociation from a strong operator, on the other hand, causes slow promoter state fluctuations and a highly non-Poissonian mRNA distribution, with few cells near the mean expression level (see Figure 5.2A, strong promoter).

In order to show that the effect of operator strength on the cell-to-cell variability is general and does not depend on the particular set of parameters chosen in the simulation, in figures 2C and 2D, we show the normalized variance and the Fano factor as a function of the fold-change in the mean mRNA concentration for a strong operator whose dissociation rate is $k_{\mathrm{R}}^{\mathrm{off}} = 0.0027\mathrm{s}^{-1}$ (a value that is representative of well-characterized repressor-operator interactions such as Lac repressor with the *lac* Oid binding site, or the lambda phage cI dimer with O_{R1}), and for a single weak operator whose dissociation rate $k_{\mathrm{R}}^{\mathrm{off}}$ is 10 times larger.

The Fano factor has a characteristic shape whereby it takes values approaching one at low and high transcription levels with a peak at intermediate values. The reason for this shape is that for very low transcription levels the promoter is nearly always inactive, firing only very rarely. In this limit successive transcription events become uncorrelated and the time in between them is exponentially distributed, leading to a distribution of mRNA per cell that approaches a Poisson distribution characterized by a Fano factor equal to one. In contrast, for very high transcription levels the promoter is nearly always active, switching off very rarely and staying in the off state for short times. In this limit, transcription events are again uncorrelated and exponentially distributed, leading once again to a Poisson distribution of mRNA number. It is only for intermediate values of the mean that the promoter is switching between a transcriptionally active and an inactive state. This causes transcription to occur in bursts, and the mRNA distribution to deviate from Poisson, leading to a Fano factor that is larger than one.

In Figure 5.2E we plot the fold-change in protein noise due to gene regulation for the simple repression architecture. As expected, we find that the effect of operator strength in protein noise is qualitatively identical to that which we found for mRNA. Since the same can be said of all the rest of the architectures studied, we will limit the discussion to mRNA noise for the rest of the paper, with the understanding that for the class of models considered here, all the conclusions about the effect of promoter architecture in cell-to-cell variability that are valid for mRNA, are true for intrinsic protein noise as well.

In Figure 5.2, and throughout this paper, we plot the Fano factor as a function of transcription level, which is characterized by the fold-change in gene expression. The fold-change in gene expression is defined as the mean mRNA number in the presence of the transcription factor, normalized by the mean mRNA in the absence of the transcription factor. For architectures based on repression, the fold-change in gene expression is always less than one, since the repressor reduces the level of transcription. For example, a fold-change in gene expression of 0.1 means that in the presence of repressor, the transcription level is 10% of the value it would have if the repressor concentration dropped to zero. For the case of activators, the fold-change is always greater than one, since activators raise the level of transcription.

An example of the single repressor-binding site architecture is a simplified version of the $\mathrm{P}_{\mathrm{lacUV5}}$ promoter. Based on a simple kinetic model of repression, in which the Lac repressor competes with RNAP for binding at the promoter, we can write down the $\hat{K}$ and $\hat{R}$ matrices and compute the cell-to-cell variability in mRNA copy number. The matrices are presented in Table S1 in Text S1. Based on our previous analysis, we know that stronger operators are expected to cause larger noise and higher values of the Fano factor than weaker operators. Therefore, we expect that if we replace the wild-type O1 operator by the 10 times weaker O2 operator, or by the ≈ 500 times weaker operator O3, the fold-change in noise should go down. Using our best estimates and available measurements for the kinetic parameters involved, we find that noise is indeed much larger for O1 than for O2, and it is negligible for O3. This prediction is presented as an inset in Figure 5.2C.

5.3.2 Promoter with two repressor-binding operators

Dual repression occurs when promoters contain two or more repressor binding sites. Here, we consider three different scenarios for architectures with two operators: 1) repressors bind independently to the two operators, 2) repressors bind cooperatively to the two operators and 3) one single repressor may be bound to the two operators simultaneously thereby looping the intervening DNA. At the molecular level, cooperative repression is achieved by two weak operators that form long-lived repressor-bound complexes when both operators are simultaneously occupied. Transcription factors may stabilize each other either through direct protein- protein interactions [53], or through indirect mechanisms mediated by alteration of DNA conformation [57].

5.3.2.1 Cooperative and independent repression.

The kinetic mechanisms of gene repression for both the cooperative and independent repressor architectures are reproduced in Figure 5.3A. For simplicity, we assume that both sites are of equal strength, so the rates of association and dissociation to both sites are equal. Cooperative binding is reflected in the fact that the rate of dissociation from the state where the two operators are occupied is slower (by a factor $\Omega << 1$) than the dissociation from a single operator. This parameter is related to the cooperativity factor ω often found in thermodynamic models [54] by $\Omega = 1/\omega$. Typical values of Ω are $\sim 10^{-3} - 10^{-2}$ [50,53]. By way of contrast, independent binding is characterized by a value of $\Omega = 1$, which reflects the fact that the rate of dissociation from each operator is not affected by the presence of the other operator.

The $\hat{K}$ and $\hat{R}$ matrices for these two architectures are defined in Table S1 in Text S1. Using these matrices, we can compute the mean gene expression and the Fano factor for these two architectures as a function of the concentrations of repressor. The resulting expression for the fold-change in noise is shown as entry number 3 of Table S2 in Text S1. As shown in Figure 5.3B, the noise for cooperative repression is substantially larger than for the independent repression architecture. The high levels of intrinsic noise associated with cooperative repression can be understood intuitively in terms of the kinetics of repressor-operator interactions. At low repressor concentration, the lifetime of the states where only one repressor is bound to either one of the two operators can be shorter than the time it takes for a second repressor to bind. This makes simultaneous binding of two repressors to the two operators a rare event. However, when it occurs, the two repressors stabilize each other, forming a very long-lived complex with the operator DNA. This mode of repression, with rare but long-lived repression events, is intrinsically very noisy, since the promoter switches slowly between active (unrepressed) and inactive (repressed) states, generating wide bimodal distributions of mRNA (see Figure 3C). On the other hand, independent binding to two operators causes more frequent transitions between repressed and unrepressed states, leading to lower levels of intrinsic noise and long-tailed mRNA distributions (see Figure 5.3C). In order to illustrate these conclusions, we have evaluated the model with a specific parameter set that is representative of this kind of bacterial promoter, and plotted the Fano factor as a function of the mean, under the assumption that we vary the mean by titrating the amount of repressor inside the cell. Furthermore, so as to demonstrate that our conclusions are not dependent on choice of parameters, we have randomly generated 10,000 different sets of kinetic parameters and compared the Fano factor for cooperative and independent binding. The result of this analysis is shown in Figure 2 in Text S1, where we demonstrate that cooperative binding always results in larger cell-to-cell variability than non-cooperative binding.

As an example of the two repressor-binding sites architecture, we consider a simplified version of the lytic phage λ $\mathrm{P_R}$ promoter, which is controlled by the lysogenic repressor cI. The wild-type $\mathrm{P_R}$ promoter consists of three proximal repressor binding sites, $\mathrm{O_{R1}}$, $\mathrm{O_{R2}}$, and $\mathrm{O_{R3}}$, with different affinities for the repressor ($\mathrm{O_{R2}}$ is about 25 times weaker than $\mathrm{O_{R1}}$) [58], and three distal operators $\mathrm{O_{L1}}$, $\mathrm{O_{L2}}$, and $\mathrm{O_{L3}}$. For simplicity, we consider a simpler version of $\mathrm{P_R}$, harboring a deletion of the three distal operators. In the absence these operators, the $\mathrm{O_{R3}}$ operator plays only a very minor role in the repression of this promoter, and can thus be ignored [50,59]. We are then left with only $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$. The cI repressor binds cooperatively to $\mathrm{O_{R1}}$ and $\mathrm{O_{R2}}$, and that cooperativity is mediated by direct protein-protein interactions between cI bound at each operator [59]. Mutant forms of cI that are cooperativity deficient (*i.e.*, not able to bind cooperatively to the promoter) have been designed [60]. In the inset in Figure 5.3B, we compare the normalized variance of the mRNA distribution, both for wild-type cI repressor, and for a cooperativity deficient mutant such as Y210H [60]. The cooperative repressor is predicted to have significantly larger promoter noise than the cooperativity deficient mutant.

5.3.2.2 Simultaneous binding of one repressor to two operators: DNA looping.

Repression may also be enhanced by the presence of distant operators, which stabilize the repressed state by allowing certain repressors to simultaneously bind to both distant and proximal operators, forming a DNA loop [61,62]. The P lac promoter is a prominent example of this architecture. The kinetic mechanism of repression characterizing this promoter architecture is presented in Figure 5.4A. The repressor only prevents transcription when it is bound to the main operator Om, but not when it is only bound to the auxiliary operator Oa. DNA loop formation is characterized by a kinetic rate $k_{\mathrm{loop}} = k^0_{\mathrm{on}}[J]$, where $[J]$, the looping J factor, can be thought of as the local concentration of repressor in the vicinity of one operator when the repressor is bound to the other operator [33,34]. The rate of dissociation of the operator-repressor complex in the looped conformation is given by $k_{\mathrm{unloop}} = ck^{\mathrm{off}}_{\mathrm{R}}$. The parameters $[J]$ and c have both been measured in vitro for the particular case of the Lac repressor [42,63], and also estimated from in vivo data [33,64]. The $\hat{K}$ and $\hat{R}$ matrices for this architecture are defined in Table S1 in Text S1. We use these matrices to compute the mean and the noise strength according to equations 5.8-5.16, resulting in the fifth entry of Table S2 in Text S1.

We first examine how the presence of the auxiliary operator affects the level of cell-to-cell variability in mRNA expression. In Figure 5.4B we compare the Fano factor in the absence of the auxiliary operator with the Fano factor in the presence of the auxiliary operator, which is assumed to be of the same strength as the main operator. We use parameters in Table 1, and we first assume that the dissociation rate of the operator-repressor complex in the looped state is the same as the dissociation rate in the unlooped state, so $c = 1$ and $k_{\mathrm{unloop}} = k^{\mathrm{off}}_{\mathrm{R}}$. This assumption is supported by single-molecule experiments in which the two operators are on the same side of the DNA double-helix, separated by multiples of the helical period of DNA [42,63]. Under these conditions we find that the presence of an auxiliary operator results in a larger Fano factor, in spite of the fact that the auxiliary operator Oa does not stabilize the binding of the repressor to the main operator Om. Interestingly, we find that the Fano factor is maximal at intermediate concentrations of repressor for which only one repressor is bound to the promoter, making the simultaneous occupancy of the auxiliary and main operators mediated by DNA looping possible. In contrast, the Fano factor is identical to that of the simple repression case if the concentration of repressor is so large that it saturates both operators and looping never occurs. It had been previously hypothesized that DNA looping might be a means to reduce noise in gene expression, due to rapid re-association kinetics between Om and a repressor that is still bound to Oa, which may cause short and frequent bursts of transcription [64,65]. Here, by applying a simple stochastic model of gene regulation, we show that the presence of the auxiliary operator does not, by itself, decrease cell-to-cell variability. On the contrary, it is expected to increase it. The reason for this increase is that the rate of dissociation from the main operator is not made faster by DNA looping; instead the presence of the auxiliary operator causes the repressor to rapidly rebind the main operator, extending the effective period of time when the promoter is repressed.

Indeed, we find that it is only if the dissociation rate for a repressor in the looped state is faster than in the unlooped state that the presence of the auxiliary operator might reduce the cell-to-cell variability. To illustrate this limit, we have assumed a value of $c = 100$, so that $k_{\mathrm{unloop}} = 100k_{\mathrm{R}}^{\mathrm{off}}$, and find that the Fano factor goes down, below the expectation for the simple repression architecture. A modest increase in the dissociation rate in the looped conformation has been reported in recent single-molecule experiments for promoter architectures in which the two operators are out of phase (located on different faces of the DNA) [42]. In order to verify the general validity of these conclusions, we have randomly chosen 10,000 different sets of kinetic parameters and compared the Fano factor for an architecture with an auxiliary operator and an architecture without the auxiliary operator (simple repressor). In this analysis the operator strength, rate of transcription, rate of DNA loop formation and mean mRNA are randomly sampled over up to four orders of magnitude. The results are shown in Figure 4 in Text S1. In the limit where dissociation of the repressor from the operator is not affected by DNA looping ($c = 1$), we find that the presence of the auxiliary operator leads to an increase in noise (Figure 4A in Text S1). In contrast, we find that when this parameter c is allowed to be larger than one, the presence of the auxiliary operator reduces cell-to-cell variability in many instances (Figure 4B in Text S1).

An example of this type of architecture is a simplified variant of the $\mathrm{P}_{\mathrm{lacUV5}}$ promoter, which consists of one main operator and one auxiliary operator upstream from the promoter. The kinetic mechanism of repression is believed to be identical to the one depicted in Figure 5.4A [23,42,63,64]. We can use the stochastic model of gene regulation described in the theory section to make precise predictions that will test this kinetic model of gene regulation by DNA looping. We find that the kinetic model predicts that, if we move the center of the auxiliary operator further upstream from its wild-type location, in increments of distance given by the helical period of the DNA, such that both operators stay in phase, the fold-change in noise should behave as represented in Figure 5.4C. In order to model the effect of DNA looping, we assume that the dependence of the rate of DNA looping on the inter-operator distance D (in units of base-pairs) is given by ([33]):

$$k_{\text{loop}} = k_{\text{R}}^{\text{on}} \times \exp\left[-\frac{u}{D} - v\log(D) + wD + z\right], \tag{5.28}$$

where $u = 140.6$, $v = 2.52$, $w = 0.0014$, $z = 19.9$, and we assume the same concentration of repressors (and therefore the same value for k_{R}^{on}) for all of the different loop lengths. Note that in Figure 5.4C, the Fano factor is not plotted as a function of the mean, but as a function of the inter-operator distance D. That is, we keep the number of repressors constant, and instead we alter the distance between the two operators. In particular, as the operator distance is changed, both the mean and the variance will change, and therefore a direct comparison between Figures 4C and 4B cannot be made. If we had plotted the Fano factor as a function of the mean (as we do in Figure 4B) we would have seen that, for the same mean, the Fano factor for looping is always larger than for a simple repression motif, consistent with Figure 5.4B.

5.3.3 Simple activation

Transcriptional activators bind to specific sites at the promoter from which they increase the rate of transcription initiation by either direct contact with one or more RNAP subunits or indirectly by modifying the conformation of DNA around the promoter [57]. The simplest example of an activating promoter architecture consists of a single binding site for an activator in the vicinity of the RNAP binding site. When the activator is not bound, transcription occurs at a low basal rate. When the activator is bound, transcription occurs at a higher, activated rate. Stochastic association and dissociation of the activator causes fluctuations in transcription rate which in turn cause fluctuations in mRNA copy number.

This simple activation architecture is illustrated in Figure 5.1A. The $\hat{K}$ and $\hat{R}$ matrices for this architecture are given in Table S1 in Text S1. Solving equations 5.8-5.11 for this particular case, we find that the mean mRNA copy number per cell takes the form:

$$\langle m \rangle = \frac{r_2}{\gamma}\frac{k_{\text{A}}^{\text{on}}}{k_{\text{A}}^{\text{on}} + k_{\text{A}}^{\text{off}}} + \frac{r_1}{\gamma}\frac{k_{\text{A}}^{\text{off}}}{k_{\text{A}}^{\text{on}} + k_{\text{A}}^{\text{off}}}. \tag{5.29}$$

The mean mRNA can be changed by adjusting the intracellular concentration of the activator. The rate at which one of the activators binds to the promoter is proportional to the activator concentration: $k_{\text{A}}^{\text{on}} = k_{\text{on}}^{0}[N_A]$. Following the same argument as we used in the simple repression case, the equilibrium dissociation constant for the activator-promoter interaction is given by $K_{OA} = k_{\text{A}}^{\text{off}}/k_{\text{on}}^{0}$. Finally, it is convenient to define the enhancement factor: the ratio between the rate of transcription in the active and the basal states $f = r_2/r_1$. The mean mRNA can be written in terms

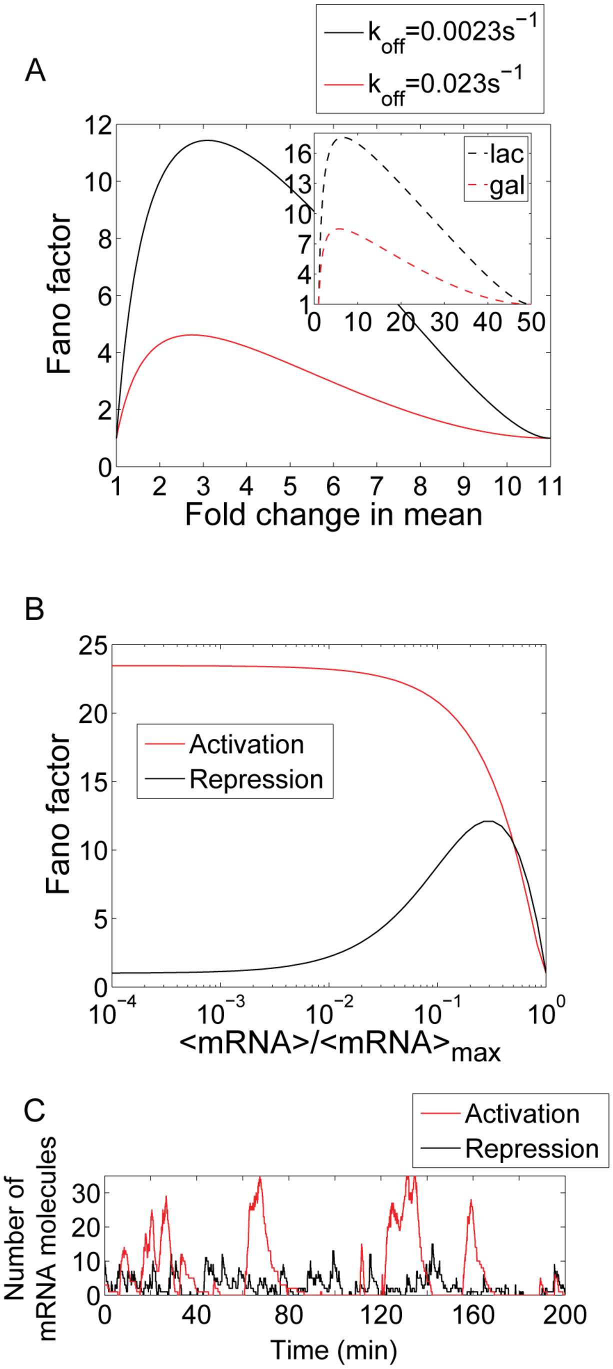


Figure 5.5: **Simple activation architecture.**(A) The Fano factor is plotted as a function of the fold-change gene expression (blue line). In red, we show the effect of reducing operator strength (i.e., reducing the lifetime of the operator-activator complex) by a factor of 10. Just as we observed with single repression, weak activator binding operators generate less promoter noise than strong activating operators. The parameters used are shown in Table 1 with the exception of $r_1 = 0.33\text{s}^{-1}/f$, where f is the enhancement factor. Inset: Prediction for the activation of the P lac promoter. The fold-change in noise is plotted as a function of the fold-change in mean mRNA expression for both the wild-type P_{lac} (CRP dissociation time = 8 min), represented by a blue line, and a P_{lac} promoter variant where the lac CRP binding site has been replaced by the weaker *gal* CRP binding site (dissociation time = 1 min). The enhancement factor was set to $f = 50$ [33]. These parameters are taken from [67] and [33]. The remaining parameters are taken from Table 1. (B) Fano factor as a function of $\langle\text{mRNA}\rangle/\langle\text{mRNA}\rangle_{\text{max}}$ for a repressor (black) and an activator (red) with the same transcription factor affinity. The transcription rate in the absence of activator is assumed to be zero. The transcription rate in the fully activated case is equal to the transcription rate of the repression construct in the absence of repressor and is $r = 0.33\text{s}^{-1}$ as specified by Table 1. For low expression levels, $\langle\text{mRNA}\rangle/\langle\text{mRNA}\rangle_{\text{max}} < 0.5$, simple activation is considerably noisier than simple repression. (C) The results of a stochastic simulation for the simple activation and simple repression architectures. We assume identical dissociation rates for the activator and repressor, and identical rates of transcription in their respective active states. As shown in (B), low concentrations of an activator result in few, but very productive transcription events, whereas high concentrations of a repressor lead to frequent but short lived excursions into the active state.

of these parameters as:

$$\langle m\rangle = \frac{r_1}{\gamma}\left(\frac{K_{OA}}{[N_A]+K_{OA}} + f\frac{[N_A]}{[N_A]+K_{OA}}\right). \tag{5.30}$$

The Fano factor can be computed using equations 5.8-5.16 and is shown as entry 2 of Table S2 in Text S1. We can rewrite the equation appearing in Table S2 in Text S1 by writing $k_{\mathrm{A}}^{\mathrm{on}}$ as a function of the mean:

$$\text{Fano} = 1 + \langle m\rangle \left(\frac{f - \langle m\rangle/\langle m\rangle_{\text{basal}}}{\langle m\rangle/\langle m\rangle_{\text{basal}}}\right)^2 \frac{\langle m\rangle\langle m\rangle_{\text{basal}} - 1}{(f - \langle m\rangle/\langle m\rangle_{\text{basal}}) + \frac{k_{\mathrm{A}}^{\mathrm{off}}}{\gamma}(f-1)}. \tag{5.31}$$

With these equations in hand, we explore how operator strength affects noise in gene expression in the case of activation. Stronger operators bind to the activator more tightly than weak operators, leading to longer residence times of the promoter in the active state.

In Figure 5.5A we plot the Fano factor as a function of the fold- change in mean expression for a strong operator as well as a 10 times weaker operator. We have used the parameters in Table 1. Just as we saw for the simple repression architecture, it is also true for the simple activation architecture that stronger operators cause larger levels of noise for activators than weaker operators.

To get a sense of the differences between these two standard regulatory mechanisms, we compare simple repression with simple activation. In Figure 5.5B, we plot the Fano factor as a function of the mean for a repressor and an activator with identical dissociation rates. We assume that the promoter switches between a transcription rate $r = 0$ in its inactive state (which happens when the repressor is bound in the simple repression case, or the activator is not bound in the simple activation case), and a rate equal to $r = 0.33\mathrm{s}^{-1}$ (see Table 1) in the active state (repressor not bound in the simple repression case, activator bound in the simple activation case). As shown in Figure 5.5B, at low expression levels the simple activation is considerably (> 20 times) noisier than the simple repression promoter. At high expression levels both architectures yield very similar noise levels, with the simple repression architecture being slightly noisier. A low level of gene expression may be achieved either by low concentrations of an activator, or by high concentrations of a repressor. Low concentrations of an activator will lead to rare activation events. High concentrations of a repressor will lead to frequent but short lasting windows of time for which the promoter is available for transcription. As a result, and as we illustrate in Figure 5.5C, the activation mechanism leads to bursty mRNA expression whereas the repressor leads to Poissonian mRNA production. This result suggests that in order to maintain a homogeneously low expression level, a repressive strategy in which a high concentration of repressor ensures low expression levels may be more adequate than a low activation strategy. We confirmed that this statement is true for other parameter sets in addition to the particular choice used above. We randomly sampled the rates of activator and repressor dissociation, as well as the rates of basal and maximum transcription. As shown in Figure 3 in Text S1, the statement that the simple activation architecture is noisier than the simple repression architecture at low expression (less than 10 mRNA/cell) levels is valid for a wide range of parameter values, with over 99% of the conditions sampled leading to this conclusion.

An example of simple activation is the wild-type P$_{lac}$ promoter, which is activated by CRP when complexed with cyclic AMP (cAMP). CRP is a ubiquitous transcription factor, and is involved in the regulation of dozens of promoters, which contain CRP binding sites of different strengths [66]. In the inset of Figure 5.5A we include CRP as an example of simple activation, and make predictions for how changing the wild-type CRP binding site in the P$_{lac}$ promoter by the CRP binding site of the P$_{gal}$ (which is ≈ 8 times weaker [67]) should affect the Fano factor. As expected from our analysis of this class of promoters, the noise goes down.

5.3.4 Dual activation: Independent and cooperative activation

Dual activation architectures have two operator binding sites. Simultaneous binding of two activators to the two operators may lead to a larger promoter activity in different ways. For instance, in some promoters each of the activators may independently contact the polymerase, recruiting it to the promoter. As a result, the probability of finding RNAP bound at the promoter increases and so does the rate of transcription [33,68]. In other instances, there is no increase in enhancement factor when the two activators are bound. However, the first activator recruits the second one through protein-protein or protein-DNA interactions, stabilizing the active state and increasing the fraction of time that the promoter is active [59]. These two modes are not mutually exclusive, and some promoters exhibit a combination of both mechanisms [69].

We first investigate the effect of dual activation in the limit where binding of the two transcription factors is not cooperative. Assuming that activators bound at the two operators independently recruit the polymerase, we compare this architecture with the simple activation architecture. The mechanism of activation is depicted in Figure 5.6A, and matrices $\hat{K}$ and $\hat{R}$ are presented in Table S1 in Text S1. For simplicity, we assume that both operators have the same strength, and both have the same enhancement factor $f = r_2/r_1 = r_3/r_1$. When the two activators are bound, the total enhancement factor is given by the product of the individual enhancement factors, which in this case is $f \times f = r_4/r_1$ [33]. All of the other relevant kinetic parameters are given in Table 1. The Fano factor is plotted in Figure 5.6B. We find that compared to the single operator architecture, the second operator increases the level of variability, even when binding to the operators is non-cooperative.

We then ask whether this is also true when the binding of activators is cooperative. We assume a small cooperativity factor $\Omega = 0.1$. Just as we found for repressors, cooperative binding of activators generates larger cell-to-cell variability than independent binding, which in turn generates larger cell-to-cell variability than simple activation. This is illustrated in the stochastic simulation in Figure 5.6C. As expected the dual activation architectures are noisier than the simple activation, characterized by rare but long-lived activation events that lead to large fluctuations in mRNA levels.

In contrast, the simple activation architecture leads to more frequent but less intense activation events.

Together with the results from the dual repressor mechanism, these results indicate that multiplicity in operator number may introduce significant intrinsic noise in gene expression. Multiple repeats of operators commonly appear in eukaryotic promoters [1,70,71], but are often found in prokaryotic promoters as well [59,68,72]. It is interesting to note that this prediction of the model is in qualitative agreement with the findings of Raj *et al* [2] who report an increase in cell-to-cell variability in mRNA when the number of activator binding sites was changed from one to seven.

An example of cooperative activation is the lysogenic phage $\lambda\mathrm{P}_{\mathrm{RM}}$ promoter [59]. This promoter contains three operators (O_{R1}, O_{R2}, and O_{R3}) for the cI protein, which acts as an activator. When O_{R2} is occupied, cI activates transcription. O_{R1} has no direct effect on the transcription rate, but it helps recruit cI to O_{R2}, since cI binds cooperatively to the two operators. Finally, O_{R3} binds cI very weakly, but when it is occupied, P_{RM} becomes repressed. There are variants of this promoter [50] that harbor mutations in O_{R3} that make it unable to bind cI. In Figure 6D, we include one of these variants, r1-P_{RM} [51] as an example of dual activation, and we present a theoretical prediction for the promoter noise as a function of the mean mRNA. We examine the role of cooperativity by comparing the wild-type cI, with a cooperativity deficient mutant. We find that the cooperative activator causes substantially larger cell-to-cell variability than the mutant, emphasizing our expectation that cooperativity may cause substantial noise in gene expression in bacterial promoters such as P_{RM}.

5.4 Discussion

The DNA sequence of a promoter encodes the binding sites for transcriptional regulators. In turn, the collection of these regulatory sites, known as the architecture of the promoter, determines the mechanism of gene regulation. The mechanism of gene regulation determines the transcriptional response of a promoter to a specific input in the form of the concentration of one or more transcription factors or inducer molecules. In recent years we have witnessed an increasing call for quantitative models of gene regulation that can serve as a conceptual framework for reflecting on the explosion of recent quantitative data, testing hypotheses, and proposing new rounds of experiments [34,73,74]. Much of this data has come from bulk transcription experiments with large numbers of cells, in which the average transcriptional response from a population of cells (typically in the form of the level of expression of a reporter protein) was measured as a function of the concentration of a transcription factor or inducer molecule [50,75]. Thermodynamic models [34,41,53] of gene regulation are a general framework for modeling gene regulation and dealing with this kind of bulk transcriptional regulation experiments This class of models has proven to be very successful at predicting gene expression patterns from the promoter architecture encoded in the DNA sequence [49,7377]. However, a new

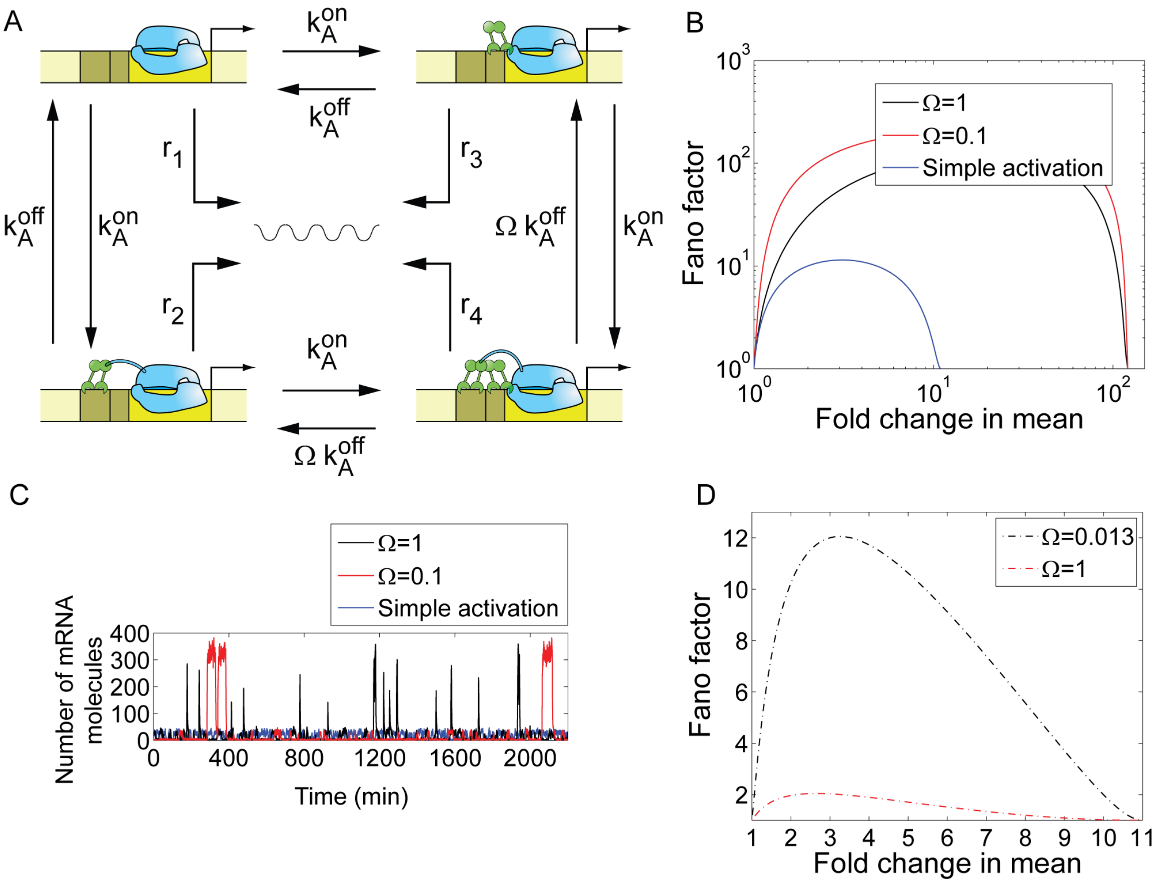


Figure 5.6: **Dual activation architecture.** (A) Kinetic mechanism of dual activation. The parameters k_A^{off} and k_A^{on} are the rates of activator dissociation and association, and Ω is a parameter reflecting the effect cooperative binding the dissociation rate. (B) Fano factor as a function of the mean mRNA for independent ($\Omega = 1$, black), cooperative ($\Omega = 0.1$, red), and for simple activation (blue). The parameters are taken from Table 1 and $r_1 = 0.33\,\mathrm{s}^{-1}/f$, $r_2 = f \times r_1$, $r_3 = f \times r_1$, and $r_4 = f^2 \times r_1$; f is the enhancement factor. (C) A stochastic simulation shows the effect of independent and cooperative binding in creating a sustained state of high promoter activity, resulting in high levels of mRNA in the active state and large cell-to-cell variability. (D) Prediction for the r1-P_{RM} promoter (a P_{RM} promoter variant that does not exhibit O_{R3} mediated repression [51]). This promoter is activated by cI, which binds cooperatively to O_{R1} and O_{R2}. The prediction is shown for wild-type cI ($\Omega = 0.013$) and for a cooperativity deficient mutant (Y210H, $\Omega = 1$). Parameters are taken from [33,43,58,97]. The lifetime of O_{R1}-cI complex is 4 min. Lifetime of O_{R2}-cI complex is 9.5 s.

generation of experiments now provides information about gene expression at the level of single-cells, with single-molecule resolution [2,4,5,6,9,10,23,31,47,51]. These experiments provide much richer information than just how the mean expression changes as a function of an input signal: they tell us how that response is spread among the population of cells, distinguishing homogeneous responses, in which all cells express the same amount of proteins or mRNA for the same input, from heterogeneous responses in which some cells achieve very high expression levels while others maintain low expression. Thermodynamic models are unable to explain the single-cell statistics of gene expression, and therefore are an incomplete framework for modeling gene regulation at the single-cell level.

A class of stochastic kinetic models have been formulated that make it possible to calculate either the probability distribution of mRNA, or proteins per cell, or its moments, for simple models of gene regulation involving one active and one inactive promoter state [36,37,45,78]. Recently, we have extended that formalism to account for any number of promoter states [30], allowing us to model any promoter architecture within the same mathematical framework. Armed with this model, we can now ask how promoter architecture affects not only the response function, but also how that response is distributed among different cells.

In this paper we have explored the feasibility of this stochastic analog of thermodynamic models as a general framework through which to understand gene regulation at the single-cell level. Using this approach we have examined a series of common promoter architectures of increasing complexity, and established how they affect the level of cell-to-cell variability of the number of mRNA molecules, and proteins, in steady state. We have found that, given the known kinetic rates of transcription factor association and dissociation from operators, the level of variability in gene expression for many well studied bacterial promoters is expected to be larger than the simple Poissonian expectation, particularly for mRNA and short-lived proteins. We have investigated how the level of variability generated by a simple promoter consisting of one single operator differs from more complex promoters containing more than one operator, and found that the presence of multiple operators increases the level of cell-to-cell variability even in the absence of cooperative binding. Cooperative binding makes the effect of operator multiplicity even larger. We also found that operator strength is one of the major determinants of cell-to-cell variability. Strong operators cause larger levels of cell-to-cell variability than weak operators. We have also examined the case where one single repressor may bind simultaneously to two operators by looping the DNA in between. We have found that the stability of the DNA loop is the key parameter in determining whether DNA looping increases or decreases the level of variability, suggesting a potential role of DNA mechanics in regulating cell-to-cell variability.

We have examined the difference between activators and repressors, and found that repressors tend to generate less cell-to-cell variability than activators at low expression levels, whereas at high expression levels repressors and activators generate similar levels of cell-to-cell variability. We conclude that induction of gene expression by increasing the concentration of an activator leads to a more heterogeneous response at low and moderate expression levels than induction of gene expression by degradation, sequestration or dilution of a repressor. In addition, we have used this model to make quantitative predictions for a few well characterized bacterial promoters, connecting the kinetic mechanism of gene regulation that we believe applies for these promoters *in vivo* with single-cell gene expression data. Direct comparison between the model and experimental data offers an opportunity to validate these kinetic mechanisms of gene regulation.

5.4.1 Intrinsic and extrinsic noise

There are two different classes of sources of cell-to-cell variability in gene expression. The first class has its origins in the intrinsically stochastic nature of the chemical reactions leading to the production and degradation of mRNAs and proteins, including the binding and unbinding of transcription factors, transcription initiation, mRNA degradation, translation, and protein degradation. The noise coming from these sources is known as intrinsic noise [79]. A different source of variability originates in cell-to-cell differences in cell size, metabolic state, copy number of transcription factors, RNA polymerases, ribosomes, nucleotides, etc. This second kind of noise is termed extrinsic noise [79]. The contributions from intrinsic and extrinsic sources can be separated experimentally, and the total noise can be written as the sum of intrinsic and extrinsic components [3]. In this paper we focus exclusively on intrinsic noise, and the emphasis is on bacterial promoters. This double focus requires us to discuss to what extent intrinsic noise is relevant in bacteria.

The experimental evidence gathered so far indicates that intrinsic noise is the dominant source of cell-to-cell variability in bacteria of the mRNA copy number. In a recent single-molecule study, transcription was monitored in real time for two different *E. coli* promoters, P_{RM} and $P_{lac/ara}$ [4]. The authors measured the rates of mRNA synthesis and dilution, as well as the rates of promoter activation and inactivation in single cells. The intrinsic noise contribution was calculated from all of these rates. It was found to be responsible for the majority of the total cell-to-cell variability, accounting for over 75% of the total variance. Another recent experiment in *B. subtilis* [7] found that mRNA expressed from the ComK promoter is also dominated by intrinsic noise. Furthermore, this study indicated that intrinsic mRNA noise is responsible for activation of a phenotypic switch that drives a fraction of the cells to competence for the uptake of DNA [7]. A third recent report investigated the activation of the genetic switch in *E. coli*, which drives the entrance of a fraction of cells into a lactose metabolizing phenotype [23]. The authors of the study found evidence that stochastic binding and unbinding of the Lac repressor to the main operator was responsible for the observed cell-to-cell variability in gene expression and, consequently the choice of phenotype. Furthermore, the authors discovered that the deletion of an auxiliary operator that permits transcriptional repression by DNA looping leads to a strong increase in the level of cell to cell variability in the expression of the lactose genes, indicating that promoter architecture plays a big role in determining the level of noise and variability in this system. Taken all together, these experiments suggest that intrinsic mRNA noise is dominant and may have important consequences for cell fate determination. In addition, at least in one case, promoter architecture has been shown to be of considerable importance.

At the protein level, the contribution of extrinsic and intrinsic noise to the total cell-to-cell variability has also been determined experimentally for a variety of promoters and different kinds of bacteria. The first reports examined intrinsic and extrinsic protein noise in *E. coli* and found that extrinsic noise was the dominant source of cell-to-cell variability in protein expressed from a variant of the $\mathrm{P_L}$ promoter in a variety of different strains [3]. However, the intrinsic component was non-negligible and for some strains, dominant [3]. A second team of researchers examined a different set of *E. coli* promoters involved in the biosynthetic pathway of lysine [80]. The authors found that the intrinsic noise contribution was significant for some promoters (i.e. lysA), but not for others. In a third study the total protein noise was measured for a Lac repressor-controlled promoter in *B. subtilis*, and it was reported that the data could be well explained by a model consisting only of intrinsic noise [8]. The authors found that the rates of transcription and translation could be determined by directly comparing the total cell-to-cell variability to the predictions of a simple stochastic model that considered only intrinsic sources of noise. They also found that the model had predictive power, and that mutations that enhanced the rate of translation or transcription produced expected effects in the total noise.

In summary, all studies that have measured mRNA noise in bacteria so far report that intrinsic noise contributes substantially to the total cell-to-cell variability. This is further supported by observations that most of the mRNA variability comes from intrinsic sources in yeast [31] and mammalian cells [1]. The issue is less clear for protein noise. Some reports indicate that it is mostly extrinsic [3], but others suggest that intrinsic noise may also be important [8,23,80]. It seems likely that the relative importance of intrinsic and extrinsic noise depends on the context, and that for some promoters and genes extrinsic noise will be larger, whereas for others the intrinsic component may dominate. In any case, it is clear that both contributions are important, and both need to be understood.

5.4.2 Comparison with experimental results

The aim of this paper is to formulate a set of predictions that reflect the class of kinetic models of gene regulation in bacteria that one routinely finds in the literature [42,64,8184]. Our analysis indicates that if these models are correct, and if the kinetic and thermodynamic parameters that have been measured over the years are also reasonably close to their real values in live cells [85], the effect of promoter architecture in cell-to-cell variability in bacteria should be rather large and easily observable. In this sense, our intention is more to motivate new experiments than to explain or fit any currently available data. We only know of one published report in which the effect of perturbing the architecture of a bacterial promoter on the cell-to-cell variability in gene expression has been determined [23]. Given that there are several examples of promoters in bacteria for which a molecular kinetic mechanism of gene regulation has been formulated [42,64,81-84,86], we hope that the computational analysis in this paper may serve as an encouragement for researchers to do for bacteria the same kind of experiments that have been already performed in eukaryotes [1,11,15,17,31]. Indeed, several different studies have examined the effect of promoter architectural elements in cell-to-cell variability in protein and mRNA in eukaryotic cells. Although our efforts in this paper have focused on bacterial promoters rather than eukaryotic promoters, it is worthwhile to discuss the findings of these studies and compare them (if only qualitatively) with the predictions made in this paper.

Two recent studies measured intrinsic mRNA noise in yeast [31] and mammalian cells [1]. Both papers concluded that stochastic promoter activation and inactivation was the leading source of intrinsic noise. While stochastic chromatin remodeling is suspected to be the origin of those activation events, neither one of these studies was conclusive about the precise molecular mechanism responsible for promoter activation. However, both studies found that promoter architecture had an important role and strongly affected the level of total mRNA noise. In both studies, the authors found that when the number of binding sites for a transcriptional activator was raised from one to seven, the normalized variance increased several-fold. This qualitative behavior is in agreement with our prediction that dual activation causes larger intrinsic mRNA noise than simple activation. It is possible that this agreement is coincidental, since the actual mechanism of gene regulation at these promoters could be much more complicated than the simple description of gene activation at a bacterial promoter adopted here.

Other studies [11,15,17] have measured the total protein noise from variants of the GAL1 promoter in yeast, and found that their data could be well explained by a model that considered only intrinsic noise sources. These studies also concluded that the main sources of intrinsic noise were stochastic activation and inactivation of the promoter due to chromatin remodeling. However, it was also found that the stable formation of pre-initiation complex at the TATA box and the stochastic binding and unbinding of transcriptional repressors contributed to the total noise [11,15,17]. The authors of these studies found that for point mutations in the TATA box of the GAL1 promoter in yeast, which made the box weaker, the level of cell-to-cell variability went down significantly. This is also in good agreement with our prediction that the stronger the binding site of a transcriptional activator, the larger the intrinsic noise should be. However, since this study measured the total noise strength, and did not isolate the intrinsic noise, the observed decrease in noise strength as a result of making the TATA box weaker may have other origins. These experiments were conducted under induction conditions that minimize repression by nucleosomes and activation by chromatin remodeling. A more recent report by the same lab [11] found that the copy number and location of a transcriptional repressor binding site greatly affects the total protein noise. The authors found that when they increased the number of repressor binding sites, the noise went up. This is also in qualitative agreement with our prediction that operator number positively correlates with intrinsic noise in the case of dual repression. However, the same caveat applies here as in the previous case studies, which is that only the total noise was measured. Although the authors of this study attributed all of the noise to intrinsic sources, it is still possible that extrinsic noise was responsible for the observed dependence of noise strength on operator number.

Finally, it is worth going back to bacteria, and discussing the only study that has yet examined the effect of a promoter architecture motif on cell-to-cell variability in gene expression. In this paper, the authors investigated the effect of DNA looping on the total cell-to-cell variability for the $\mathrm{P}_{\mathrm{lacUV5}}$ promoter in *E. coli* [23]. Using a novel single-protein counting technique, Choi and co-workers measured protein distributions for promoters whose auxiliary operator had been deleted (leaving them with a simple repression architecture), and compared them to promoters with the auxiliary operator O3 present, which allows for DNA looping. They report a reduction in protein noise due to the presence of O3, which according to our analysis, may indicate that the dissociation of the repressor from the looped state is faster than the normal dissociation rate. The authors attributed this looping-dependent decrease in noise to intrinsic origins, related to the different kinetics of repressor binding and rebinding to the main operator in the presence of the auxiliary operator, and in its absence. However, their measurements also reflect the total noise,

More recently, several impressive experimental studies have measured the noise in mRNA in bacteria for a host of different promoters ([87], and Ido Golding, private communication). In both of these cases, simplified low-dimensional models which do not consider the details of the promoter architecture have been exploited to provide a theoretical framework for thinking about the data. Our own studies indicate that the differences between a generic two-state model and specific models that attempt to capture the details of a given architecture are sometimes subtle and that the acid test of ideas like those presented in this paper can only come from experiments which systematically tune parameters, such as the repressor concentration, for a given transcriptional architecture.

5.4.3 Future Directions

Some recent theoretical work has analyzed the effect of cooperative binding of activators in the context of particular examples of eukaryotic promoters [88,89]. The main focus of this study is bacterial promoters. The simplicity of the microscopic mechanisms of transcriptional regulation for bacterial promoters makes them a better starting point for a systematic study like the one we propose. However, many examples of eukaryotic promoters have been found whose architecture affects the cell-to-cell variability [1,11,17,31,32]. Although the molecular mechanisms of gene regulation in these promoters are much more complex, with many intervening global and specific regulators [90], the stochastic model employed in this paper can be applied to any number of promoter states, and thus can be applied to these more complex promoters. Recent experimental work is starting to reveal the dynamics of nucleosomes and transcription factors with single-molecule sensitivity [91,92], allowing the formulation of quantitative kinetic and thermodynamic mechanistic models of transcriptional regulation at the molecular level [73,77]. The framework for analyzing gene expression at the single-cell level developed in this paper will be helpful to investigate the kinetic mechanisms of gene regulation in eukaryotic promoters, as the experimental studies switch from ensemble, to single-cell.

5.4.4 Shortcomings of the approach

Although the model of transcriptional regulation used in this paper is standard in the field, it is important to remark that it is a very simplified model of what really happens during transcription initiation. There are many ways in which this kind of model can fail to describe real situations. For instance, mRNA degradation requires the action of RNases. These may become saturated if the global transcriptional activity is very large and the degradation becomes non-linear [55]. Transcription initiation and elongation are assumed to be jointly captured in a single constant rate of mRNA synthesis for each promoter state. This is an oversimplification also. When considered explicitly, and in certain parameter ranges, the kinetics of RNAP-promoter interaction may cause noticeable effects in the overall variability [46]. Similarly, as pointed out elsewhere [93,94,95], translational pausing, back-tracking or road-blocking may also cause significant deviations in mRNA variability from the predictions of the model used in this paper. How serious these deviations are depends on the specifics of each promoter-gene system. The model explored in this paper also assumes that the cell is a well-mixed environment. Deviations from that approximation can significantly affect cell-to-cell variability [56,96]. Another simplification refers to cell growth and division, which are not treated explicitly by the model used in this paper: cell division and DNA replication cause doubling of gene and promoter copy number every cell cycle, as well as binomial partitioning of mRNAs between mother and daughter cells [3]. In eukaryotes, mRNA often needs to be further processed by the splicing apparatus before it becomes transcriptionally active. It also needs to be exported out of the nucleus, where it can be translated by ribosomes.

To study the effect of transcription factor dynamics on mRNA noise we assume that the unregulated promoter produces mRNA in a Poisson manner, at a constant rate. This assumption can turn out to be wrong if there is another process, independent of transcription factors, that independently turns the promoter on and off. In eukaryotes examples of such processes are nucleosome positioning and chromatin remodeling, while in prokaryotes analogous processes are not as established, but could include the action of non-specifically bound nucleoid proteins such as HU and HNS, or DNA supercoiling. Experiments that measure cell-to-cell distributions of mRNA copy number in the absence of transcription factors (say without Lac repressor for the lac operon case) can settle this question. In case the Fano factor for this distribution is not one (as expected for a Poisson distribution) this can signal a possible transcription factor-independent source of variability. The stochastic models studied here can be extended to account for this situation. For example, the promoter can be made to switch between an on and an off state, where the transcription factors are allowed to interact with promoter DNA only while it is in the on state. In this case the mRNA fluctuations produced by an unregulated promoter will not be Poissonian. One can still investigate the affect of transcription factors by measuring how they change the nature of mRNA fluctuations from this new baseline. Comparison of this extended model with single-cell transcription experiments would then have the exciting potential for uncovering novel modes of transcriptional regulation in prokaryotes.

For the purpose of isolating the effect of individual promoter architectural elements on cell-to-cell variability in gene expression, we have artificially changed the value of one of those parameters, while keeping the other parameters constant. For instance, we have investigated the effect of altering the strength of an operator on the total cell-to-cell variability. In order to do this, we ask how changes in the dissociation rate of the transcription factor alter the cell-to-cell variability, given that all other rates (say the rate of transcription, or mRNA degradation) remain constant. This assumption is not necessarily always correct, since very often the operator sequence overlaps the promoter, and therefore changes in the sequence that alter operator strength also affect the sequence from which RNAP initiates transcription, which can potentially affect the overall rates of transcription. As is usually the case, biology presents us with a great diversity of forms, shapes and functions, and promoters are no exception. One needs to examine each promoter independently on the basis of the assumptions made in this paper, as many of these assumptions may apply for some promoters, but not for others.

For the same reason of isolating the effect of promoter architecture and cis-transcriptional regulation on cell-to-cell variability in gene expression, when we compare different architectures we make the simplifying assumption that they are transcribing the same gene, and therefore that the mRNA transcript has the same degradation rate. Care must be taken to take this into account when promoters transcribing different genes are investigated, since the mRNA degradation rate has a large effect on the level of cell-to-cell variability.

We have also assumed that when transcription factors dissociate from the operator, they dissociate into an averaged out, well-mixed, mean-field concentration of transcription factors inside the cell. The possibility of transcription factors being recaptured by the same or another operator in the promoter right after they fall off the operator is not captured by the class of models considered here. Recent in vivo experiments suggest that this scenario may be important in yeast promoters containing arrays of operators [31].

In spite of all of the simplifications inherent in the class of models analyzed in this paper, we believe they are an adequate jumping off point for developing an intuition about how promoter architecture contributes to variability in gene expression. Our approach is to take a highly simplified model of stochastic gene expression, based on a kinetic model for the processes of the central dogma of molecular biology, and add promoter dynamics explicitly to see how different architectural features affect variability. This allows us to isolate the effect of promoter dynamics, and develop an intuitive understanding of how they affect the statistics of gene expression.

It must be emphasized, however, that the predictions made by the model may be wrong if any of the complications mentioned above are significant. This is not necessarily a bad outcome. If the comparison between experimental data and the predictions made by the theory for any particular system reveals inconsistencies, then the model will need to be refined and new experiments are required to identify which of the sources of variability that are not accounted for by the model are in play. In other words, experiments that test the quantitative predictions outlined stand a chance of gaining new insights about the physical mechanisms that underlie prokaryotic transcriptional regulation.

5.4.5 Supporting information

Text S1. Mathematical derivations and supplementary information. A derivation of all equations in the text is presented, together with its corresponding tables and figures.

The moments of the mRNA probability distribution

We start by considering the same mechanism as in the text (see figure 1), in which the promoter switches between one active and one inactive state. There are only two stochastic variables in the model: the number of mRNA transcripts per cell (*m*), and the state of the promoter which reflects which transcription factors are bound where. The promoter state is always a discrete and finite stochastic variable (*s*) (for an example, see figure 1a). The example in figure 1a illustrates the simplest model of transcriptional activation by a transcription factor.

When the activator is bound to the promoter (state 1) mRNA is synthesized at rate r_1. When the activator is not bound (state 2) mRNA is synthesized at a lower rate r_2. The promoter switches stochastically from state 1 to state 2 with rate k_A^{off}, and from state 2 to state 1 with rate k_A^{on}. Each mRNA molecule is degraded with rate γ.

The time evolution for the joint probability of having the promoter in states 1 or 2, with m mRNAs in the cell (which we write as $p(1,m)$ and $p(2,m)$, respectively), is given by a master equation, which we can build by listing all possible reactions that lead to a change in cellular state, either by changing m or by changing s (figure 1b). The master equation takes the form:

$$\begin{aligned}
\frac{d}{dt}p(1,m) &= -k_A^{off}\,p(1,m)+k_A^{on}\,p(2,m)-r_1p(1,m)-\gamma mp(1,m)+r_1p(1,m-1)+\gamma(m+1)p(1,m+1),\\
\frac{d}{dt}p(2,m) &= k_A^{off}\,p(1,m)-k_A^{on}\,p(2,m)-r_2p(2,m)-\gamma mp(2,m)+r_2p(2,m-1)+\gamma(m+1)p(2,m+1)\quad .
\end{aligned}\tag{1}$$

Inspecting this system of equations, we notice that by defining the vector:

$$\vec{p}(m)=\begin{pmatrix} p(1,m) \\ p(2,m) \end{pmatrix}, \tag{2}$$

and the matrices

$$\hat{K}=\begin{bmatrix} -k_A^{off} & k_A^{on} \\ k_A^{off} & -k_A^{on} \end{bmatrix} \;;\; \hat{R}=\begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix} \;;\; \hat{I}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

we can rewrite the system of equations (1) in matrix form.

$$\frac{d}{dt}\vec{p}(m)=\left[\hat{K}-\hat{R}-m\gamma\hat{I}\right]\vec{p}(m)+\hat{R}\,\vec{p}(m-1)+(m+1)\gamma\hat{I}\;\vec{p}(m+1) \;\; . \tag{3}$$

This approach can be generalized to any mechanism of transcriptional regulation at the promoter level. The only difference between the mechanisms rests on the particular dimensionality and form of the three matrices defined above. Examples of those matrices for all of the architectures and mechanisms investigated on this paper are given in Table S1 in Text S1. In steady state, the left hand side of equation (4) is equal to 0:

$$0=\left[\hat{K}-\hat{R}-m\gamma\hat{I}\right]\vec{p}(m)+\hat{R}\,\vec{p}(m-1)+(m+1)\gamma\hat{I}\;\vec{p}(m+1) \;\; . \tag{4}$$

In order to find the first two moments of the steady state mRNA probability distribution, we follow the same strategy as in references [1,2]: we multiply both sides of equation (5) by m and m^2 respectively, and then sum over all values of m, from 0 to ∞. We start from the first moment of the mRNA distribution, which requires us to multiply equation (5) by m and then sum:

$$\begin{aligned}\sum_{m=0}^{\infty} m\left(\left[\hat{K}-\hat{R}-m\gamma\hat{I}\right]\vec{p}(m)+\hat{R}\,\vec{p}(m-1)+(m+1)\gamma\hat{I}\;\vec{p}(m+1)\right) &= \sum_{m=0}^{\infty} m\;\hat{K}\;\vec{p}(m)-\sum_{m=0}^{\infty} m^2\gamma\hat{I}\;\vec{p}(m) \\ -\sum_{m=0}^{\infty} m\;\hat{R}\,\vec{p}(m)+\sum_{m=0}^{\infty} m\;\hat{R}\,\vec{p}(m-1)+\sum_{m=0}^{\infty} m\;(m+1)\;\gamma\hat{I}\;\vec{p}(m+1) \;\; . \end{aligned} \tag{5}$$

Since none of the three matrices $\hat{K}$, $\hat{R}$ and $\hat{I}$ are functions of *m*, they can be taken out of the sums, and we find:

$$0=\hat{K}\sum_{m=0}^{\infty} m\;\vec{p}(m)-\gamma\hat{I}\sum_{m=0}^{\infty} m^2\vec{p}(m)-\hat{R}\sum_{m=0}^{\infty} m\;\vec{p}(m)+\hat{R}\sum_{m=0}^{\infty} m\;\vec{p}(m-1)+\gamma\hat{I}\sum_{m=0}^{\infty} m\;(m+1)\;\vec{p}(m+1) \;\; . \tag{6}$$

It will be convenient in what follows to define the following vectors of partial moments of the mRNA probability distribution:

$$\vec{m}_{(0)} = \sum_{m=0}^{\infty} m^0 \vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m^0 p(1,m) \\ \sum_{m=0}^{\infty} m^0 p(2,m) \end{pmatrix} = \begin{pmatrix} p(1) \\ p(2) \end{pmatrix},$$

$$\vec{m}_{(1)} = \sum_{m=0}^{\infty} m\ \vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m\ p(1,m) \\ \sum_{m=0}^{\infty} m\ p(2,m) \end{pmatrix}, \tag{7}$$

$$\vec{m}_{(2)} = \sum_{m=0}^{\infty} m^2\ \vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m^2\ p(1,m) \\ \sum_{m=0}^{\infty} m^2\ p(2,m) \end{pmatrix} .$$

The usefulness of these vectors of partial moments of the mRNA distribution lies in the fact that they are related to the moments of the probability distribution. For instance, the mean mRNA is given by

$$\langle m \rangle = \sum_{s=1}^{2} \sum_{m=0}^{\infty} m\ p(s,m) = \sum_{m=0}^{\infty} m\ p(1,m) + \sum_{m=0}^{\infty} m\ p(2,m) \quad . \tag{8}$$

If we define, again for convenience, the vector $\vec{u} = (1,1)$, we find that the mean of the mRNA distribution is related to the vectors of partial moments by $\langle m \rangle = \vec{u}\ \vec{m}_{(1)}$. Following this example, it is also straightforward to prove that the second moment of the mRNA distribution is given by: $\langle m^2 \rangle = \vec{u}\ \vec{m}_{(2)}$.

Given these definitions, we return to equation (7) which we can now write as:

$$\hat{K}\ \vec{m}_{(1)} - \gamma \hat{I}\ \vec{m}_{(2)} - \hat{R}\ \vec{m}_{(1)} + \hat{R} \sum_{m=0}^{\infty} m\ \vec{p}(m-1) + \gamma \hat{I} \sum_{m=0}^{\infty} m\ (m+1)\ \vec{p}(m+1) = 0 \quad . \tag{9}$$

We can re-arrange terms in the last two sums so that we write them as operations on the vectors of partial moments of the probability distributions. For instance, by making the change of variables: $m \to m+1$, and taking into account the fact that the number of mRNA molecules inside the cell can never fall below 0 (so that $\vec{p}(-1) = 0$), we find:

$$\sum_{m=0}^{\infty} m\ \vec{p}(m-1) = \sum_{m=0}^{\infty} (m+1)\ \vec{p}(m) = \vec{m}_{(1)} + \vec{m}_{(0)} \quad . \tag{10}$$

Similarly, by making the change of variables $m+1 \rightarrow m$, the last sum takes the simpler form:

$$\sum_{m=0}^{\infty} m(m+1)\ \vec{p}(m+1) = \sum_{m=0}^{\infty} m(m-1)\ \vec{p}(m) = \vec{m}_{(2)} - \vec{m}_{(1)} \quad . \tag{11}$$

Entering these results into equation (10), we finally find:

$$\hat{K}\ \vec{m}_{(1)} - \gamma\hat{I}\ \vec{m}_{(2)} - \hat{R}\ \vec{m}_{(1)} + \hat{R}\left(\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(\vec{m}_{(2)} - \vec{m}_{(1)}\right) = \hat{K}\ \vec{m}_{(1)} - \gamma\hat{I}\ \vec{m}_{(1)} + \hat{R}\ \vec{m}_{(0)} \quad . \tag{12}$$

The vector of partial moments $\vec{m}_{(1)}$ is therefore the solution to the matrix equation:

$$\left(\hat{K} - \gamma\hat{I}\right)\ \vec{m}_{(1)} + \hat{R}\ \vec{m}_{(0)} = 0 \quad . \tag{13}$$

The final step is to multiply both sides of equation (14) by the vector $\vec{u} = (1,1)$. Because of how it was constructed (i.e. *p(1,m)*s loss is *p(2,m)*s gain during transitions between promoter states), the matrix $\hat{K}$ has the property that the sum of the elements of any one of its columns is always 0. Therefore, we find that $\vec{u}\ \hat{K} = 0$. The matrix $\hat{R}$ is diagonal, so if we multiply matrix $\hat{R}$ on the left by vector $\vec{u}$, we get a vector that is equal to the list of diagonal elements of matrix $\hat{R}$. Thus, we define the vector $\vec{r} = \left(\hat{R}_{11}, \hat{R}_{22}\right) = \left(r_1, r_2\right)$, as the vector for which it is true that $\vec{u}\ \hat{R} = \vec{r}$. Finally, the identity matrix is $\hat{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Therefore, multiplying $\hat{I}$ on the left by the vector $\vec{u}$ leads us to: $\vec{u}\ \hat{I} = \vec{u}$. Therefore, when we multiply equation (14) by the vector $\vec{u}$ we find:

$$0 = \vec{u}\ \hat{K}\ \vec{m}_{(1)} - \vec{u}\ \gamma\hat{I}\ \vec{m}_{(1)} + \vec{u}\ \hat{R}\ \vec{m}_{(0)} = -\gamma\vec{u}\ \vec{m}_{(1)} + \vec{r}\ \vec{m}_{(0)} \quad . \tag{14}$$

Knowing that the mean of the mRNA distribution is related to the vector of partial moments by: $\langle m \rangle = \vec{u}\ \vec{m}_{(1)}$, we find that:

$$\langle m \rangle = \frac{\vec{r}\ \vec{m}_{(0)}}{\gamma} \quad . \tag{15}$$

Note that, by definition,

$$\vec{m}_{(0)} = \sum_{m=0}^{\infty} m^0\,\vec{p}(m) = \begin{pmatrix} \sum_{m=0}^{\infty} m^0\,p(1,m) \\ \sum_{m=0}^{\infty} m^0\,p(2,m) \end{pmatrix} = \begin{pmatrix} p(1) \\ p(2) \end{pmatrix} \quad . \tag{16}$$

In other words, the first element of vector $\vec{m}_{(0)}$ is the steady state probability to find the promoter in state 1, and the second element is the steady state probability to find the promoter in state 2. This vector is straightforward to obtain by summing equation (5) over all *m*, and it is the solution of $\hat{K}\ \vec{m}_{(0)} = 0$, normalized so that $p(1) + p(2) = 1$.

In order to find the second moment, we just multiply equation (5) by m^2 and sum over all *m* from 0 to ∞. As a result of this manipulation, we find:

$$\begin{aligned}
&\sum_{m=0}^{\infty} m^2 \left(\left[\hat{K} - \hat{R} - m\gamma\hat{I} \right] \vec{p}(m) + \hat{R}\ \vec{p}(m-1) + (m+1)\gamma\hat{I}\ \vec{p}(m+1) \right) = \sum_{m=0}^{\infty} m^2\ \hat{K}\ \vec{p}(m) - \sum_{m=0}^{\infty} m^3\ \gamma\hat{I}\ \vec{p}(m) \\
&-\sum_{m=0}^{\infty} m^2\ \hat{R}\ \vec{p}(m) + \sum_{m=0}^{\infty} m^2\ \hat{R}\ \vec{p}(m-1) + \sum_{m=0}^{\infty} m^2\ (m+1)\ \gamma\hat{I}\ \vec{p}(m+1) = \\
&\hat{K}\ \vec{m}_{(2)} - \gamma\hat{I}\ \vec{m}_{(3)} - \hat{R}\ \vec{m}_{(2)} + \sum_{m=0}^{\infty} m^2\ \hat{R}\ \vec{p}(m-1) + \sum_{m=0}^{\infty} m^2\ (m+1)\ \gamma\hat{I}\ \vec{p}(m+1)\ \ .
\end{aligned} \tag{17}$$

The last two terms of the right hand side of equation (18) can be simplified by writing the two sums in terms of the vectors of partial moments. In order to do that, we must make the same changes of variables that we invoked above when dealing with the mean. First, the change of variables $m \to m+1$ allows us to rewrite the first sum as:

$$\sum_{m=0}^{\infty} m^2\ \vec{p}(m-1) = \sum_{m=0}^{\infty} (m+1)^2\ \vec{p}(m) = \vec{m}_{(2)} + 2\vec{m}_{(1)} + \vec{m}_{(0)}\ \ . \tag{18}$$

Finally, the change of variables $m+1 \to m$, allows us to re-write the last sum as:

$$\sum_{m=0}^{\infty} m^2 (m+1)\ \vec{p}(m+1) = \sum_{m=0}^{\infty} m(m-1)^2\ \vec{p}(m) = \vec{m}_{(3)} - 2\vec{m}_{(2)} + \vec{m}_{(1)}\ \ . \tag{19}$$

Entering these last two sums in equation (18), we find:

$$\begin{aligned}
&\hat{K}\ \vec{m}_{(2)} - \gamma\hat{I}\ \vec{m}_{(3)} - \hat{R}\ \vec{m}_{(2)} + \hat{R}\left(\vec{m}_{(2)} + 2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(\vec{m}_{(3)} - 2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = \\
&\hat{K}\ \vec{m}_{(2)} + \hat{R}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\hat{I}\left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = 0\ \ .
\end{aligned} \tag{20}$$

As we did before, we can transform this equation into an equation for the moments of the mRNA distribution by multiplying both sides of this equation on the left by the vector $\vec{u}$. Performing these operations, we find:

$$\begin{aligned}
&\vec{u}\ \hat{K}\ \vec{m}_{(2)} + \vec{u}\ \hat{R}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \vec{u}\ \gamma\hat{I}\left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = \vec{r}\left(2\vec{m}_{(1)} + \vec{m}_{(0)}\right) + \gamma\vec{u}\ \left(-2\vec{m}_{(2)} + \vec{m}_{(1)}\right) = \\
&2\ \vec{r}\ \vec{m}_{(1)} + \vec{r}\ \vec{m}_{(0)} - 2\ \gamma\left\langle m^2 \right\rangle + \gamma\left\langle m \right\rangle = 0\ \ .
\end{aligned} \tag{21}$$

Therefore, the second moment of the mRNA distribution in steady state is given by:

$$\left\langle m^2 \right\rangle = \frac{\vec{r}\ \vec{m}_{(1)}}{\gamma} + \frac{\vec{r}\ \vec{m}_{(0)} + \gamma \left\langle m \right\rangle}{2\gamma} \quad . \tag{22}$$

Using the fact that the first moment is given by:

$$\left\langle m \right\rangle = \frac{\vec{r}\ \vec{m}_{(0)}}{\gamma} \quad . \tag{23}$$

We can further simplify the second moment as:

$$\left\langle m^2 \right\rangle = \left\langle m \right\rangle + \frac{\vec{r}\ \vec{m}_{(1)}}{\gamma} \quad . \tag{24}$$

Therefore, the normalized variance can be written as:

$$\eta^2 = \frac{\left\langle m^2 \right\rangle - \left\langle m \right\rangle^2}{\left\langle m \right\rangle^2} = \frac{1}{\left\langle m \right\rangle} + \frac{1}{\left\langle m \right\rangle^2}\left(\frac{\vec{r}\ \vec{m}_{(1)}}{\gamma} - \left\langle m \right\rangle^2 \right) \quad . \tag{25}$$

The moments of the protein probability distribution

We can use the same method to compute the normalized variance of the protein distribution. We will start from a promoter that is constitutively active, and then extend our analysis to a promoter that switches between two or more active and inactive states. We assume that each transcription event leads to the production of multiple proteins (a "burst"). The number of proteins produced per mRNA (which we denote as β) obeys a geometric distribution [3,4,5] with an average burst size b. Therefore, the probability for β is given by: $h(\beta) = \frac{b^\beta}{(1+b)^{\beta+1}}$. We assume that proteins are also degraded with a constant rate per molecule of δ. In order to write down the master equation for this process, we have to consider all the possible ways in which the cell can enter or leave a state with *n* proteins during a small increment of time dt. If we assume that mRNA lifetime is much shorter than protein lifetime (an approximation that is realistic in many experimental systems –see refs [4,5,6]), then all of the proteins may be assumed to be made simultaneously. Therefore, we need to consider the possibility that the cell will jump from a state with *n* proteins to a state with $n+\beta$, for all possible values of β. The probability that the cell will leave a state with *n* proteins, by making a transition to a state with $n+\beta$ proteins is equal to the product of the probability that the cell is in a state with *n* proteins $(p(n))$, the probability that the cell will make a transcript during dt (rdt), and the probability that the transcript makes

β proteins before it is degraded $\left(h\left(\beta\right)\right)$. Thus, the total probability per unit time to abandon a state with *n* proteins is given by $r\ h\left(\beta\right)p(n)$. Since β can in principle take any integer value, the total probability to abandon the state with *n* proteins by the occurrence of a protein burst is given by the sum of $r\ h\left(\beta\right)p(n)$ over all possible values of β. This term will be given by: $\sum_{\beta=1}^{\infty} r\ h\left(\beta\right)p(n) = r\ p(n)\sum_{\beta=1}^{\infty} h\left(\beta\right)$. Also, we need to consider that the cell may enter a state with *n* proteins from any state with less than *n* proteins. The probability per unit time that the cell enters a state with n proteins, from a state with $n-\beta$ proteins is given by: $rh\left(\beta\right)p(n-\beta)$. Therefore, following the same logic as we did before, the net probability per unit time that the cell enters a state with *n* proteins is $\sum_{\beta=1}^{n} r\ h\left(\beta\right)p(n-\beta)$. With these considerations, the master equation for a constitutive promoter is given by:

$$\frac{d}{dt}p(n) = -\sum_{\beta=1}^{\infty} rh\left(\beta\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1) \quad . \tag{26}$$

As discussed above, the first sum can be further simplified to:

$$\sum_{\beta=1}^{\infty} rh\left(\beta\right)p(n) = r\ p(n)\sum_{\beta=1}^{\infty} h\left(\beta\right) = r\ p(n)\sum_{\beta=1}^{\infty}\frac{b^{\beta}}{\left(1+b\right)^{\beta+1}} = r\left(\frac{b}{1+b}\right)p(n)\,. \tag{27}$$

As a result, the master equation takes the form:

$$\frac{d}{dt}p(n) = -r\left(\frac{b}{1+b}\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1)\,. \tag{28}$$

In steady state, the right hand side of equation (29) is equal to 0, and we have:

$$0 = -r\left(\frac{b}{1+b}\right)p(n) + \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta) - \delta np(n) + \delta(n+1)p(n+1)\,. \tag{29}$$

The first two moments of the steady state protein distribution $p(n)$ can be obtained, in exactly the same way we used to find out the moments of the mRNA distribution in the previous section: by multiplying both sides of equation (30) by n and n^2 respectively, and then summing over all n. Before we do that, it is useful to evaluate the sums $\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta)$ and $\sum_{n=0}^{\infty} n \sum_{\beta=1}^{n} rh\left(\beta\right)p(n-\beta)$. We can find the general term of the first sum by expanding the series:

$$\begin{aligned}&\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} h(\beta)\, p(n-\beta) = 1^2\left(h(1)p(0)\right) + 2^2\left(h(1)p(1)+h(2)p(0)\right) + 3^2\left(h(1)p(2)+h(2)p(1)+h(3)p(0)\right)+\ldots = \\ &\left(1^2 h(1) + 2^2 h(2) + 3^2 h(3)\ldots\right)p(0) + \left(2^2 h(1) + 3^2 h(2) + 4^2 h(3)\ldots\right)p(1) + \left(3^2 h(1) + 4^2 h(2) + 5^2 h(3)\ldots\right)p(2)+\ldots = \\ &\sum_{n=0}^{\infty} p(n)\left(\sum_{\beta=1}^{\infty} h(\beta)(n+\beta)^2\right) = \sum_{n=0}^{\infty}\left(b + 2b^2 + 2bn + \frac{b}{1+b}n^2\right)p(n)\ \ .\end{aligned} \tag{30}$$

We can do the same for the second sum, and we find:

$$\begin{aligned}&\sum_{n=0}^{\infty} n \sum_{\beta=1}^{n} h(\beta)\, p(n-\beta) = 1\left(h(1)p(0)\right) + 2\left(h(1)p(1)+h(2)p(0)\right) + 3\left(h(1)p(2)+h(2)p(1)+h(3)p(0)\right)+\ldots = \\ &\left(h(1) + 2h(2) + 3h(3)\ldots\right)p(0) + \left(2h(1) + 3h(2) + 4h(3)\ldots\right)p(1) + \left(3h(1) + 4h(2) + 5h(3)\ldots\right)p(2)+\ldots = \\ &\sum_{n=0}^{\infty} p(n)\left(\sum_{\beta=1}^{\infty} h(\beta)(n+\beta)\right) = \sum_{n=0}^{\infty}\left(b + \frac{b}{1+b}n\right)p(n)\ \ .\end{aligned} \tag{31}$$

Likewise, it will be necessary to recall from the first section of this supplement, that the sum $\sum_{n=0}^{\infty} n(n+1)\,p(n+1)$ can be computed by using the change of variables: $n+1 \to n$, and we find:

$$\sum_{n=0}^{\infty} n(n+1)\,p(n+1) = \sum_{n=0}^{\infty} n(n-1)\,p(n)\ \ . \tag{32}$$

With these results in hand, we can finally solve the first two moments of the protein distribution $p(n)$. As explained above, we can find the first moment by multiplying both sides of equation (30) by n and then summing over all *n*. In order to find the second moment, we multiply both sides of equation (30) by n^2 and then sum over all *n*. For the first moment, we find:

$$\begin{aligned}0 &= -r\left(\frac{b}{1+b}\right)\sum_{n=0}^{\infty} np(n) + r\sum_{n=0}^{\infty} n\sum_{\beta=1}^{n} h(\beta)\, p(n-\beta) - \delta\sum_{n=0}^{\infty} n^2 p(n) + \delta\sum_{n=0}^{\infty} n(n+1)p(n+1) = \\ &= -r\left(\frac{b}{1+b}\right)\langle n\rangle + r\left(b + \frac{b}{1+b}\langle n\rangle\right) - \delta\langle n^2\rangle + \delta\langle n^2\rangle - \delta\langle n\rangle = rb - \delta\langle n\rangle.\end{aligned} \tag{33}$$

Solving this equation, we find that the mean protein per cell is equal to:

$$\langle n\rangle = \frac{rb}{\delta}. \tag{34}$$

For the second moment, we find:

$$\begin{aligned}
0 &= -r\left(\frac{b}{1+b}\right)\sum_{n=0}^{\infty} n^2 p(n) + r\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} h\left(\beta\right) p(n-\beta) - \delta\sum_{n=0}^{\infty} n^3 p(n) + \delta\sum_{n=0}^{\infty} n^2 (n+1) p(n+1) = \\
&= -r\left(\frac{b}{1+b}\right)\left\langle n\right\rangle + r\left(b + 2b^2 + 2b\left\langle n\right\rangle + \frac{b}{1+b}\left\langle n^2\right\rangle\right) - \delta\left\langle n^3\right\rangle + \delta\left\langle n^3\right\rangle - 2\delta\left\langle n^2\right\rangle + \delta\left\langle n\right\rangle = \qquad (35) \\
&= r\left(b + 2b^2 + 2b\left\langle n\right\rangle\right) - 2\delta\left\langle n^2\right\rangle + \delta\left\langle n\right\rangle.
\end{aligned}$$

Solving this last equation, we find that the second moment of the protein distribution is equal to:

$$\left\langle n^2\right\rangle = \frac{r}{2\delta}b + \frac{r}{2\delta}2b^2 + \frac{r}{2\delta}2b\left\langle n\right\rangle + \frac{\left\langle n\right\rangle}{2} = \left(1+b\right)\left\langle n\right\rangle + \left\langle n\right\rangle^2. \qquad (36)$$

Therefore, the normalized variance of the protein distribution for a constitutive promoter takes the form:

$$\frac{Var(n)}{\left\langle n\right\rangle^2} = \frac{\left\langle n^2\right\rangle - \left\langle n\right\rangle^2}{\left\langle n\right\rangle^2} = \frac{\left(1+b\right)\left\langle n\right\rangle + \left\langle n\right\rangle^2 - \left\langle n\right\rangle^2}{\left\langle n\right\rangle^2} = \frac{\left(1+b\right)}{\left\langle n\right\rangle} \quad . \qquad (37)$$

If now we consider that the promoter can exist in two states, characterized by having different rates of transcription, then the cell’s state is characterized not only by the number of proteins present, but also by the state of the promoter. Therefore, the master equation must consider two variables: one characterizing the state of the promoter (*s*), and one representing the number of proteins per cell (*n*). By analogy with the mRNA master equation, and the master equation for the protein distribution of a constitutive promoter, the two-state master equation for the protein distribution can be written as:

$$\begin{aligned}
&\frac{d}{dt}p(1,n) = -k_A^{on} p(1,n) + k_A^{off} p(2,n) - \sum_{\beta=1}^{\infty} r_1 h\left(\beta\right) p(1,n) + \sum_{\beta=1}^{n} r_1 h\left(\beta\right) p(1,n-\beta) - \delta n p(1,n) + \delta(n+1) p(1,n+1), \\
&\frac{d}{dt}p(2,n) = k_A^{on} p(1,n) - k_A^{off} p(2,n) - \sum_{\beta=1}^{\infty} r_2 h\left(\beta\right) p(2,n) + \sum_{\beta=1}^{n} r_2 h\left(\beta\right) p(2,n-\beta) - \delta n p(2,n) + \delta(n+1) p(2,n+1) \quad .
\end{aligned} \qquad (38)$$

Just as we did in order to compute the moments of the mRNA distribution, we can define the vector $\vec{p}(n) = \left(p(1,n), p(2,n)\right)$. By doing so, we will be able to re-write the master equation (39) as a matrix equation, that will be applicable to any promoter with any number of states. This matrix equation can be written in terms of exactly the same matrices we used for the mRNA probability distribution. We find:

$$\frac{d}{dt}\vec{p}(n) = \left[\hat{K} - \frac{b}{1+b}\hat{R} - n\delta\,\hat{I}\right]\vec{p}(n) + \hat{R}\sum_{\beta=1}^{n} h(\beta)\vec{p}(n-\beta) + (n+1)\delta\,\hat{I}\,\vec{p}(n+1) \quad . \qquad (39)$$

In steady state, the left side of equation (40) is equal to 0, and the master equation has the form:

$$0=\left[\hat{K}-\frac{b}{1+b}\hat{R}-n\delta\ \hat{I}\right]\vec{p}(n)+\hat{R}\sum_{\beta=1}^{n}h(\beta)\vec{p}(n-\beta)+(n+1)\delta\ \hat{I}\ \vec{p}(n+1)\ \ . \tag{40}$$

Just as we did in order to calculate the moments of the mRNA distribution, it will be convenient to define the vectors of partial moments:

$$\begin{aligned}
\vec{n}_{(0)}&=\sum_{n=0}^{\infty}n^{0}\vec{p}(n)=\begin{pmatrix}\sum_{n=0}^{\infty}n^{0}p(1,n)\\ \sum_{n=0}^{\infty}n^{0}p(2,n)\end{pmatrix}=\begin{pmatrix}p(1)\\ p(2)\end{pmatrix},\\
\vec{n}_{(1)}&=\sum_{n=0}^{\infty}n\ \vec{p}(n)=\begin{pmatrix}\sum_{n=0}^{\infty}n\ p(1,n)\\ \sum_{n=0}^{\infty}n\ p(2,n)\end{pmatrix},\\
\vec{n}_{(2)}&=\sum_{n=0}^{\infty}n^{2}\ \vec{p}(n)=\begin{pmatrix}\sum_{n=0}^{\infty}n^{2}\ p(1,n)\\ \sum_{n=0}^{\infty}n^{2}\ p(2,n)\end{pmatrix}\quad .
\end{aligned}\tag{41}$$

It is straightforward to see that the vector $\vec{n}_{(0)}$ is exactly identical to the vector $\vec{m}_{(0)}$. The next two vectors $\vec{n}_{(1)}$ and $\vec{n}_{(2)}$ can be obtained by multiplying equation (41) by n and n^{2} respectively, and then summing over all n. We end up with the following two equations:

$$\begin{aligned}
0&=\sum_{n=0}^{\infty}n\left[\hat{K}-\frac{b}{1+b}\hat{R}-n\delta\ \hat{I}\right]\vec{p}(n)+\hat{R}\sum_{n=0}^{\infty}n\sum_{\beta=1}^{n}h(\beta)\vec{p}(n-\beta)+\sum_{n=0}^{\infty}n(n+1)\delta\ \hat{I}\cdot\vec{p}(n+1)=\\
&=\hat{K}\ \vec{n}_{(1)}-\frac{b}{1+b}\hat{R}\ \vec{n}_{(1)}-\delta\ \hat{I}\ \vec{n}_{(2)}+\delta\ \hat{I}\left(\vec{n}_{(2)}-\vec{n}_{(1)}\right)+\hat{R}\left(\frac{b}{1+b}\vec{n}_{(1)}+b\vec{n}_{(0)}\right)\\
&=\left(\hat{K}-\delta\ \hat{I}\right)\ \vec{n}_{(1)}+b\ \hat{R}\ \vec{n}_{(0)},
\end{aligned}\tag{42}$$

and

$$\begin{aligned} 0 &= \sum_{n=0}^{\infty} n^2 \left[\hat{K} - \frac{b}{1+b}\hat{R} - n\delta\, \hat{I} \right] \vec{p}(n) + \hat{R}\sum_{n=0}^{\infty} n^2 \sum_{\beta=1}^{n} h(\beta)\,\vec{p}(n-\beta) + \sum_{n=0}^{\infty} n^2 (n+1)\delta\, \hat{I}\cdot\vec{p}(n+1) = \\ &= \hat{K}\ \vec{n}_{(2)} - \frac{b}{1+b}\hat{R}\ \vec{n}_{(2)} - \delta\ \hat{I}\ \vec{n}_{(3)} + \delta\ \hat{I}\left(\vec{n}_{(3)} - 2\vec{n}_{(2)} + \vec{n}_{(1)}\right) + \hat{R}\left(\frac{b}{1+b}\vec{n}_{(2)} + 2b\vec{n}_{(1)} + b(1+2b)\vec{n}_{(0)}\right) = \\ &= \hat{K}\ \vec{n}_{(2)} + \delta\hat{I}\left(-2\vec{n}_{(2)} + \vec{n}_{(1)}\right) + \hat{R}\left(2b\vec{n}_{(1)} + b(1+2b)\vec{n}_{(0)}\right) = \\ &= \left(\hat{K}\ - 2\delta\hat{I}\right)\vec{n}_{(2)} + \left(\delta\hat{I} + 2b\hat{R}\right)\vec{n}_{(1)} + b(1+2b)\hat{R}\ \vec{n}_{(0)} \quad . \end{aligned} \tag{43}$$

Now by multiplying the vector $\vec{u} = (1,1)$ on the left of equations (43) and (44), we find

$$0 = -\delta\langle n\rangle + b\ \vec{r}\ \vec{n}_{(0)}\,, \tag{44}$$

and:

$$0 = -2\delta\left\langle n^2\right\rangle + \delta\langle n\rangle + 2b\ \vec{r}\ \vec{n}_{(1)} + b(1+2b)\vec{r}\ \vec{n}_{(0)} \quad . \tag{45}$$

Thus, we find analytical equations for the first two moments of the protein distribution:

$$\langle n\rangle = \frac{b\ \vec{r}\ \vec{n}_{(0)}}{\delta}, \tag{46}$$

$$\left\langle n^2\right\rangle = (1+b)\langle n\rangle + \frac{b\ \vec{r}\ \vec{n}_{(1)}}{\delta}. \tag{47}$$

Where $\vec{n}_{(1)}$ is the solution of equation (43):

$$0 = \left(\hat{K} - \delta\ \hat{I}\right)\ \vec{n}_{(1)} + b\ \hat{R}\ \vec{n}_{(0)}, \tag{48}$$

Armed with these equations, we can finally compute the stationary variance of the protein distribution:

$$\frac{Var(n)}{\langle n\rangle^2} = \frac{(1+b)\langle n\rangle + \frac{b\ \vec{r}\cdot\vec{n}_{(1)}}{\delta} - \langle n\rangle^2}{\langle n\rangle^2} = \frac{(1+b)}{\langle n\rangle} + \frac{1}{\langle n\rangle^2}\left(\frac{b\ \vec{r}\ \vec{n}_{(1)}}{\delta} - \langle n\rangle^2\right) \quad . \tag{49}$$

Exploration of the space of parameter values

In order to test how some of the key qualitative and quantitative conclusions discussed in the main text depend on choice of rate constants that characterize the different architectures, we computed the Fano factor for a large set of parameter values drawn randomly from the space of possible values. The results of these calculations are shown in figures S2, S3, and S4.

REFERENCES

1. Sanchez A, Kondev J (2008) Transcriptional control of noise in gene expression. Proc Natl Acad Sci: 105,5081-5086.
2. Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys J: 81,3116-3136.
3. Berg O (1978) A model for statistical fluctuations of protein numbers in a microbial-population. J Theor Biol: 71:587-603.
4. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature: 440, 358-362.
5. Yu J, Xiao, J., Ren, X., Lao, K., S. Xie (2006) Probing gene expression in live cells one protein at a time. Science: (311)1600-1603.
6. Kennell D, Riezman, H. (1977) Transcription and translation initiation frequencies of the Escherichia coli lac operon. J Mol Biol 114: 1-21.

SUPPLEMENTARY FIGURES

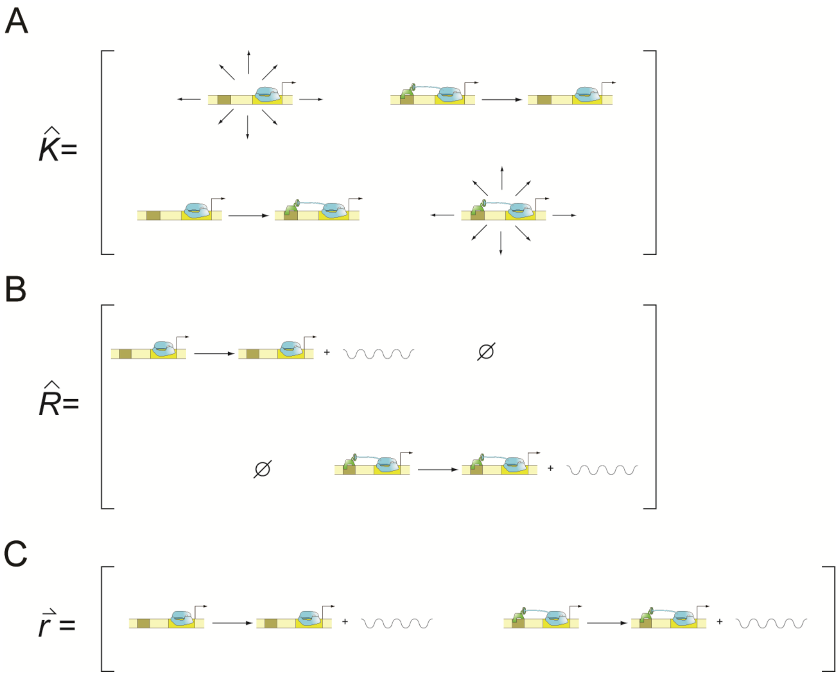


Figure 1. Cartoon depiction of the construction of kinetic rate matrices and vectors. (A) Cartoon representation of the kinetic rate matrix $\hat{K}$. The diagonal elements represent the net rate at which the promoter abandons each state. For instance, element $\{\hat{K}\}_{11}$ is the rate at which the promoter abandons state 1 due to stochastic association of the activator with the promoter: $\{\hat{K}\}_{11} = -k_A^{on}$, and element $\{\hat{K}\}_{22} = -k_A^{off}$ is the rate of dissociation of the activator from the promoter, abandoning state 2. The non-diagonal element $\{\hat{K}\}_{21} = k_A^{on}$ is the rate at which the promoter makes a transition from state 1 to state 2 (by dissociation association of one activator to the promoter), and the non-diagonal element $\{\hat{K}\}_{12} = k_A^{off}$ is the rate at which the promoter makes a transition from state 2 to state 1 (by dissociation of the activator). (B) The transcription rate matrix $\hat{R}$ contains, in its diagonal elements, the net rate of transcription at each promoter state. Element $\{\hat{R}\}_{11} = r_1$ is the rate of transcription in promoter state 1 and $\{\hat{R}\}_{22} = r_2$ is the rate of transcription in promoter state 2. (C) The vector $\vec{r} = (r_1, r_2)$ contains the rates of transcription at states 1 and 2, and is identical to the diagonal of matrix $\hat{R}$.

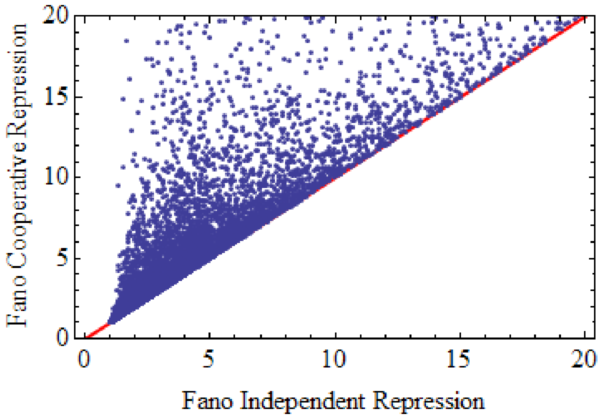


Figure 2. Effect of parameter choice on Fano factor for independent and cooperative repression architectures. We sample the parameter space by randomly selecting 10,000 different values for the mean mRNA $\langle m \rangle$ (within 0.005 and 100), k_R^{off} (from k_R^{off}/γ =0.01 to k_R^{off}/γ =100), $\langle m \rangle_{\max}$ (from 5 to 100), and Ω (from 0.001 to 1). The Fano factor is calculated for both independent and cooperative repression architectures, when the mean is the same for both. In the X axis we plot the Fano Factor for independent repression. In the Y axis we plot the Fano factor for cooperative repression. As is the case throughout the paper, we assume that we vary the mean by titrating the amount of repressor inside the cell. Each point in the figure corresponds to two architectures with the same mean. We find that cooperative binding always results in larger cell-to-cell variability than non-cooperative binding. The red solid line marks the region where the Fano factor is the same for both architectures.

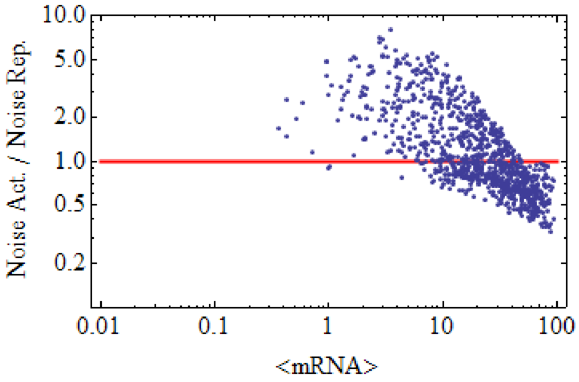


Figure 3. Simple activation tends to be noisier than simple repression at low expression levels. We follow the same procedure as in figure S1, and sample the parameter space by randomly selecting 1,000 different values for the mean mRNA $\langle m \rangle$ (within 0.01 and 100), k_R^{off} and k_A^{off} (from $0.01\,\gamma$ to $100\,\gamma$), $\langle m \rangle_{\max}$, the enhancement factor f (from 10 to 100). For each one of these 10,000 sets of parameters, we compute the Fano factor for the simple activation and the simple repression architectures. We plot the ratio between the Fano factor for simple activation and repression as a function of the mean. We find that at low mRNA levels $\left(\langle m \rangle < 10\right)$, the simple activation architecture is noisier than the simple repression architecture in over 99% of the sets of rates tested here. In contrast, at high mRNA levels, it is the other way around. In order for the comparison between both architectures to be meaningful, we have assumed that the repressor and the activator have the same affinity for their operators (even if we vary this affinity over 4 orders of magnitude). The red solid line marks the region where the Fano factor is the same for both architectures (and thus the ratio between the two is 1)

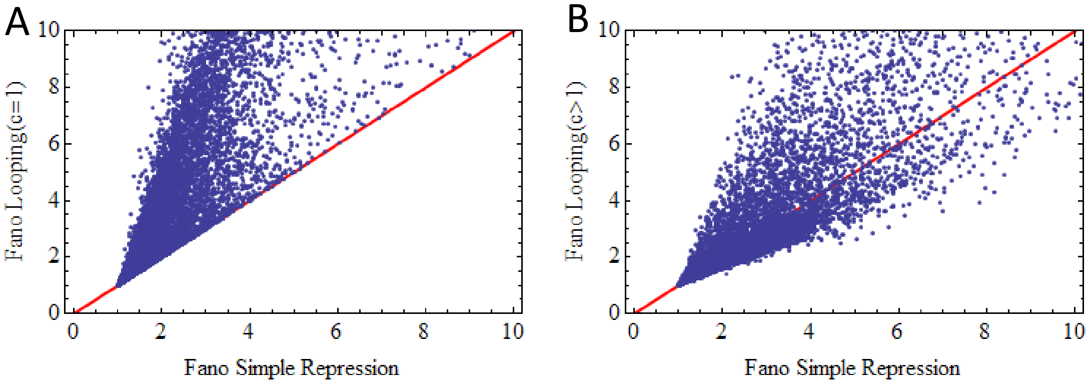
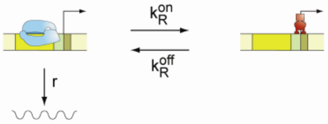
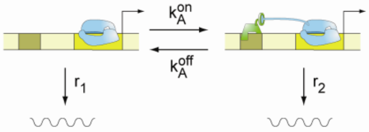
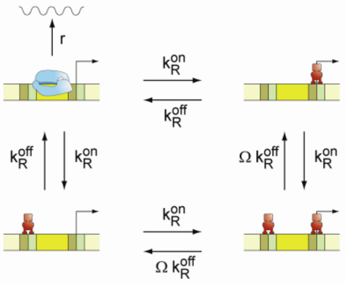


Figure 4. Effect of parameter choice on Fano factor for the repression by DNA looping architecture We sample the parameter space by randomly selecting 10,000 different values for the mean mRNA $\langle m \rangle$ (within 0.005 and 100), k_R^{off} (from k_R^{off}/γ =0.01 to k_R^{off}/γ =100), $\langle m \rangle_{\max}$ (from 5 to 100), k_{loop} (from 0.01 γ to 100 γ), and the parameter characterizing the rate of dissociation in the presence of the auxiliary operator, relative to that in its absence (*c*). We first assume that *c* = 1 for all parameter sets (A), and then we randomly sample it within 1 and 10 (B). In the X axis we plot the Fano Factor for simple repression. In the Y axis we plot the Fano factor for repression by DNA looping. As is the case throughout the paper, we assume that we vary the mean by titrating the amount of repressor inside the cell. Each point in the figure corresponds to two architectures with the same mean. We find that whether DNA looping enhances or diminishes noise depends on the value of *c*. If *c* = 1, meaning that DNA looping does not affect the rate of dissociation of the repressor from the operator, the Fano factor for the DNA looping architecture is larger than the Fano factor for the simple repression architecture. On the other hand, if $c > 1$, DNA looping may decrease noise (as observed for ~40% of the parameters chosen). The red solid line marks the region where the Fano factor is the same for both architectures.

Mechanism	Matrices and vectors	
	$\hat{K}$	$\vec{u}\cdot\hat{R}$
	$\begin{pmatrix} -k_R^{on} & k_R^{off} \\ k_R^{on} & -k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ 0 \end{pmatrix}$
	$\begin{pmatrix} -k_A^{on} & k_A^{off} \\ k_A^{on} & -k_A^{off} \end{pmatrix}$	$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$
 r k_R^{on} k_R^{off} k_R^{off} k_R^{on} $\Omega\, k_R^{off}$ k_R^{on} k_R^{on} $\Omega\, k_R^{off}$	$\begin{pmatrix} -2k_R^{on} & k_R^{off} & k_R^{off} & 0 \\ k_R^{on} & -(k_R^{off}+k_R^{on}) & 0 & \Omega\, k_R^{off} \\ k_R^{on} & 0 & -(k_R^{off}+\ k_R^{on}) & \Omega\, k_R^{off} \\ 0 & k_R^{on} & k_R^{on} & -2\Omega\, k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ 0 \\ 0 \\ 0 \end{pmatrix}$

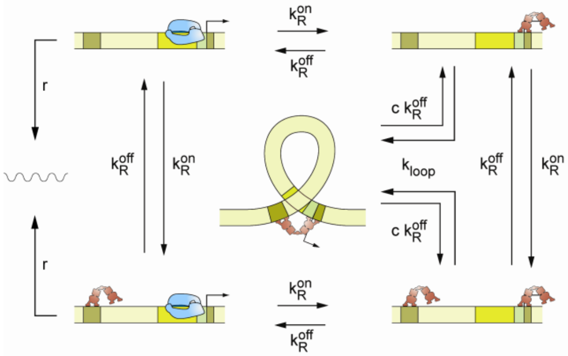
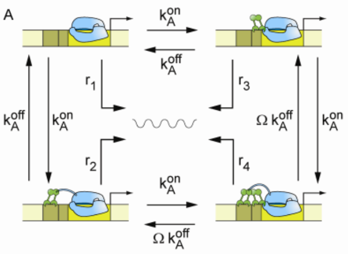
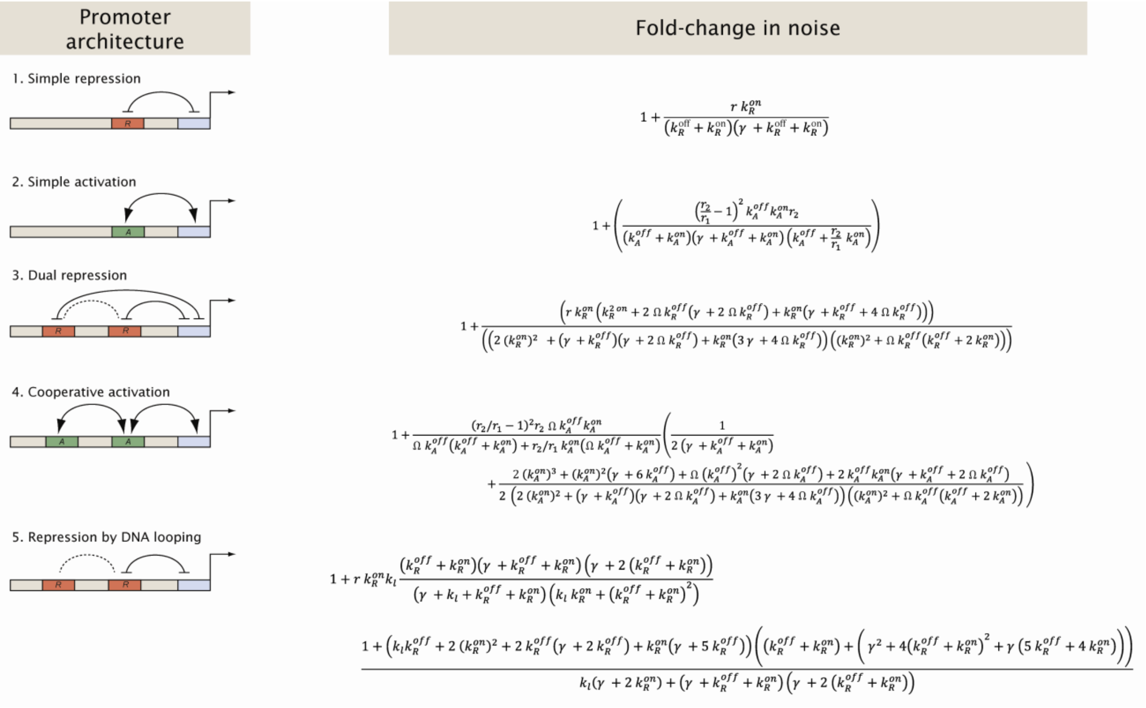
Mechanism	Matrices and vectors	
	$\hat{K}$	$\vec{u}\cdot\hat{R}$
 k_R^{on} k_R^{off} r k_R^{off} k_R^{on} $c\ k_R^{off}$ k_{loop} $c\ k_R^{off}$ k_R^{off} k_R^{on} r k_R^{on} k_R^{off}	$\begin{pmatrix} -2k_R^{on} & k_R^{off} & 0 & 0 & 0 \\ k_R^{on} & -(k_R^{off}+k_R^{on}+k_l) & 0 & k_R^{off} & c\,k_R^{off} \\ k_R^{on} & 0 & -(k_R^{off}+k_R^{on}+k_l) & k_R^{off} & c\,k_R^{off} \\ 0 & k_R^{on} & k_R^{on} & -2k_R^{off} & 0 \\ 0 & k_l & k_l & 0 & -2c\,k_R^{off} \end{pmatrix}$	$\begin{pmatrix} r \\ r \\ 0 \\ 0 \\ 0 \end{pmatrix}$
 A k_A^{on} k_A^{off} k_A^{off} k_A^{on} r_1 r_2 r_3 r_4 $\Omega\, k_A^{off}$ k_A^{on} k_A^{on} $\Omega\, k_A^{off}$	$\begin{pmatrix} -2k_A^{on} & k_A^{off} & k_A^{off} & 0 \\ k_A^{on} & -(k_A^{off}+k_A^{on}) & 0 & \Omega\, k_A^{off} \\ k_A^{on} & 0 & -(k_A^{off}+k_A^{on}) & \Omega\, k_A^{off} \\ 0 & k_A^{on} & k_A^{on} & -2\Omega\, k_A^{off} \end{pmatrix}$	$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$

Table S1: Kinetic rate matrices for all mechanisms in the text. In the first column, we represent the kinetic mechanisms of gene regulation for all of the architectures considered in the text. In the second and third columns, we show the corresponding promoter kinetic transition rate matrices $\hat{K}$ and the vector $\vec{r} = \vec{u}\ \hat{R}$ for all of the mechanisms.



Promoter architecture	Fold-change in noise
1. Simple repression	$1 + \frac{r\, k_R^{on}}{\left(k_R^{\mathrm{off}} + k_R^{\mathrm{on}}\right)\left(\gamma + k_R^{\mathrm{off}} + k_R^{\mathrm{on}}\right)}$
2. Simple activation	$1 + \left(\frac{\left(\frac{r_2}{r_1} - 1\right)^2 k_A^{off} k_A^{on} r_2}{\left(k_A^{off} + k_A^{on}\right)\left(\gamma + k_A^{off} + k_A^{on}\right)\left(k_A^{off} + \frac{r_2}{r_1}\, k_A^{on}\right)}\right)$
3. Dual repression	$1 + \frac{\left(r\, k_R^{on}\left(k_R^{2\,on} + 2\,\Omega\, k_R^{off}\left(\gamma + 2\,\Omega\, k_R^{off}\right) + k_R^{on}\left(\gamma + k_R^{off} + 4\,\Omega\, k_R^{off}\right)\right)\right)}{\left(\left(2\,(k_R^{on})^2 + \left(\gamma + k_R^{off}\right)\left(\gamma + 2\,\Omega\, k_R^{off}\right) + k_R^{on}\left(3\,\gamma + 4\,\Omega\, k_R^{off}\right)\right)\left((k_R^{on})^2 + \Omega\, k_R^{off}\left(k_R^{off} + 2\, k_R^{on}\right)\right)\right)}$
4. Cooperative activation	$1 + \frac{(r_2/r_1 - 1)^2 r_2\ \Omega\, k_A^{off} k_A^{on}}{\Omega\, k_A^{off}\left(k_A^{off} + k_A^{on}\right) + r_2/r_1\, k_A^{on}\left(\Omega\, k_A^{off} + k_A^{on}\right)}\left(\frac{1}{2\left(\gamma + k_A^{off} + k_A^{on}\right)} + \frac{2\,(k_A^{on})^3 + (k_A^{on})^2\left(\gamma + 6\, k_A^{off}\right) + \Omega\left(k_A^{off}\right)^2\left(\gamma + 2\,\Omega\, k_A^{off}\right) + 2\, k_A^{off} k_A^{on}\left(\gamma + k_A^{off} + 2\,\Omega\, k_A^{off}\right)}{2\left(2\,(k_A^{on})^2 + \left(\gamma + k_A^{off}\right)\left(\gamma + 2\,\Omega\, k_A^{off}\right) + k_A^{on}\left(3\,\gamma + 4\,\Omega\, k_A^{off}\right)\right)\left((k_A^{on})^2 + \Omega\, k_A^{off}\left(k_A^{off} + 2\, k_A^{on}\right)\right)}\right)$
5. Repression by DNA looping	$1 + r\, k_R^{on} k_l \frac{\left(k_R^{off} + k_R^{on}\right)\left(\gamma + k_R^{off} + k_R^{on}\right)\left(\gamma + 2\left(k_R^{off} + k_R^{on}\right)\right)}{\left(\gamma + k_l + k_R^{off} + k_R^{on}\right)\left(k_l\, k_R^{on} + \left(k_R^{off} + k_R^{on}\right)^2\right)}$ $\frac{1 + \left(k_l k_R^{off} + 2\,(k_R^{on})^2 + 2\, k_R^{off}\left(\gamma + 2\, k_R^{off}\right) + k_R^{on}\left(\gamma + 5\, k_R^{off}\right)\right)\left(\left(k_R^{off} + k_R^{on}\right) + \left(\gamma^2 + 4\left(k_R^{off} + k_R^{on}\right)^2 + \gamma\left(5\, k_R^{off} + 4\, k_R^{on}\right)\right)\right)}{k_l\left(\gamma + 2\, k_R^{on}\right) + \left(\gamma + k_R^{off} + k_R^{on}\right)\left(\gamma + 2\left(k_R^{off} + k_R^{on}\right)\right)}$

Table S2: Fold-change in noise for different promoter architectures. The fold-change in promoter noise is shown as a function of the different kinetic parameters corresponding to each promoter architecture considered throughout the text. Refer to Table I for the definition and value of each rate.

5.6 Bibliography

1. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 10: e309.

2. Elf J, Li GW, Xie XS (2007) Probing transcription factor dyamics at the single molecule level in a single cell. *Science* 316: 11911194.

3. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297: 11831186.

4. Golding I, Paulsson J, Zawilski SM, Cox E (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123: 10251036.

5. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* 440: 358362.

6. Chubb JR, Trcek T, Shenoy SM, Singer RH (2006) Transcriptional pulsing of a developmental gene. *Curr Biol* 16: 10181025.

7. Maamar H, Raj A, Dubnau D (2007) Noise in Gene Expression Determines. *Science* 317: 526529.

8. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31: 6973.

9. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2009) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5: 877879.

10. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells one protein at a time. *Science* 311: 16001603.

11. Murphy KF, Balazsi G, Collins JJ (2007) Combinatorial promoter design for engineering noisy gene expression. *Proc Natl Acad Sci* 104: 1272612731.

12. Rigney DR, Schieve WC (1977) Stochastic model of linear, continuous protein synthesis in bacterial populations. *J Theor Biol* 69: 761766.

13. Berg O (1978) A model for statistical fluctuations of protein numbers in a microbial-population. *J Theor Biol* 71: 587603.

14. Bar-Even A, Paulsson J, Maheshri N, Carmi M, OShea EK, et al. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* 38: 636643.

15. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422: 633637.

16. Raser JM, OShea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304: 18111814.

17. Blake WJ, Balazsi G, Kohanski MA, Isaacs FJ, Murphy KF, et al. (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell* 24: 853865.

18. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Gen* 6: 451464.

19. Maheshri N, OShea EK (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct* 36: 413434.

20. Wernet MF, Mazzoni EO, Celik A, Duncan DM, Duncan I, et al. (2006) Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* 440: 174180.

21. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* 122: 169182.

22. Ackerman M, Stecher B, Freed NE, Songhet P, Hardt W, et al. (2008) Self-destructive cooperation mediated by phenotypic noise. *Nature* 454: 987990.

23. Choi PJ, Cai L, Frieda K, Xie XS (2008) A stochastic single molecule event triggers phenotype switching of a bacterial cell. *Science* 322: 442445.

24. Losik R, Desplan C (2008) Stochasticity and cell fate. *Science* 320: 6568.

25. Singh A, Weinberger LS (2009) Stochastic gene expression as a molecular switch for viral latency. *Curr Op Microbiol* 12: 460466.

26. Austin DW, Allen MS, McCollum JM, Dar RD, Wilgus JR, et al. (2006) Gene network shaping of inherent noise spectra. *Nature* 439: 608611.

27. Cox CD, McCollum JM, Allen MS, Dat RS, Simpson ML (2008) Using noise to probe and characterize gene circuits. *Proc Natl Acad Sci* 105: 1080910814.

28. Nevozhay D, Adams RM, Murphy KF, Josic K, Balazsi G (2009) Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. *Proc Natl Acad Sci* 106: 51235128.

29. Pedraza JM, Van Oudenaarden A (2005) Noise propagation in gene networks. *Science* 307: 19651969.

30. Sanchez A, Kondev J (2008) Transcriptional control of noise in gene expression. *Proc Natl Acad Sci* 105: 50815086.

31. To TL, Maheshri N (2010) Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science* 327: 11421145.

32. Rossi FMV, Kringstein AM, Spicher A, Guicherit OM, Blau HM (2000) Transcriptional control: Rheostat converted to On/Off switch. *Mol Cell* 6: 723728.

33. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: Applications. *Curr Opin Gen Dev* 15: 125135.

34. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* 15: 116124.

35. Paulsson J (2004) Summing up the noise in gene networks. *Nature* 427: 415418.

36. Peccoud J, Ycart B (1995) Markovian modelig of gene product synthesis. *Theor Popul Biol* 48: 222234.

37. Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J* 81: 31163136.

38. Ingram PJ, Stumpf MP, Stark J (2008) Nonidentifiability of the source of intrinsic noise in gene expression from single-burst data. *PLoS Comp Biol* 4: e1000192.

39. Warmflash A, Dinner A (2008) Signatures of combinatorial regulation in intrinsic biological noise. *Proc Natl Acad Sci* 105: 1726217267.

40. Dunlop MJ, Cox III RS, Levine JH, Murray RM, Elowitz MB (2007) Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet* 40: 14931498.

41. Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda: A physical chemical model for gene regulation. *J Mol Biol* 181: 211230.

42. Wong OK, Guthold M, Erie DA, Gelles J (2008) Interconvertible lac repressor- DNA loops revealed by single-molecule experiments. *PLOS Biol* 6: e232.

43. Wang Y, Guo L, Golding I, Cox EC, Ong NP (2009) Quantitative transcription factor binding kinetics at the single-molecule level. *Biophys J* 96: 609620.

44. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci* 98: 16841689.

45. Paulsson J (2005) Models of stochastic gene expression. *Phys Life Rev* 2: 157175.

46. Höfer T, Rasch MJ (2005) On the kinetic design of transcription. *Genome Inform* 16: 7382.

47. Zenklusen D, Larson DR, Singer RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* 15: 12631271.

48. Halford SE (2009) An end of 40 years of mistakes in DNA-protein association kinetics. *Biochem Soc Trans* 37: 343348.

49. Kim HD, OShea EK (2008) A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* 15: 11921198.

50. Dodd IB (2004) Cooperativity in long-range gene regulation by the lambda cI repressor. *Genes Dev* 18: 344354.

51. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science* 307: 19621965.

52. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81: 23402361.

53. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. *Proc Natl Acad Sci* 100: 51355141.

54. Cox CD, McCollum JM, Austin DW, Allen MS, Dar RD, et al. (2006) Frequency domain analysis of noise in simple gene circuits. *Chaos* 16: 026102.

55. Pedraza JM, Paulsson J (2008) Effects of molecular memory and bursting on flucuations in gene. *Science* 319: 339343.

56. van Zon JS, Morelli MJ, TanaseNicola S, ten Wolde PR (2006) Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys J* 91: 43504367.

57. Browning DF, Busby SJW (2004) The regulation of bacterial transcription initiation. *Nat Revs Microbiol* 2: 19.

58. Koblan KS, Ackers GK (1992) Site-specific enthalpic regulation of DNA- transcription at bacteriophage-lambda Or. *Biochemistry* 31: 5765.

59. Ptashne M (2004) A Genetic Switch. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

60. Babic AC, Little, JW (2007) Cooperative DNA binding by cI repressor is dispensable in a phage-lambda variant. *Proc Natl Acad Sci* 104: 1774117746.

61. Semsey S, Geanacopoulos M, Lewis DEA, Adhya S (2002) Operator-bound GalR dimers close DNA loops by direct interaction: Tetramerization and inducer binding. *EMBO J* 21: 43494356.

62. Muller-Hill B (2004) The Lac Operon: A Short History of a Genetic Paradigm. Berlin, New York: Walter de Gruyter.

63. Vanzi F, Broggio C, Sacconi L, Pavone FS (2006) Lac repressor hinge flexibility and DNA looping: single molecule kinetics by tethered particle motion. *Nucleic Acids Res* 34: 34093420.

64. Vilar JM, Leibler S (2003) DNA looping and physical constraints on transcriptional regulation. *J Mol Biol* 331: 981989.

65. Vilar JM, Saiz L (2005) DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transriptional noise. *Curr Opin Genet Dev* 15: 136144.

66. Cameron AD, Redfield RJ (2008) CRP binding and transcription activation at CRP-S sites. *J Mol Biol* 383: 313323.

67. Gaston K, Kolb A, Busby S (1989) Binding of the Escherichia coli cyclic AMP receptor protein toDNA fragments containing consensus nucleotide sequences. *Biochem J* 261: 649653.

68. Joung JK, Koepp DM, Hochschild A (1994) Synergistic activation of transcription by bacteriophage-lambda cI-protein and escherichia coli CAMP receptor protein. *Science* 265: 18631866.

69. Joung JK, Le LU, Hochschild A (1993) Synergistic activation of transcription by Escherichia-coli CAMP Receptor Protein. *Proc Natl Acad Sci* 90: 30833087.

70. Burz BS, Rivera-Pomar R, Jackle H, Hanes SD (1998) Cooperative DNA- binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. *EMBO J* 17: 59986009.

71. Karpova TS, Kim MJ, Spriet C, Nalley K, Stasevich TJ, et al. (2008) Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter. *Science* 5862. pp 466469.

72. Shin M, Kang S, Hyun S-J, Fujita N, Ishishama A, et al. (2001) Repression of deoP2 in

Escherichia coli by CytR: Conversion of a transcription activator into a repressor. *EMBO J* 19: 53925399.

73. Segal E, Widom J (2009) From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat Rev Genet* 10: 443456.

74. Kim HD, Shay T, OShea EK, Regev A (2009) Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science* 325: 429432.

75. Kuhlman T, Zhang Z, Saier MH, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Natl Acad Sci* 104: 60436048.

76. Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457: 215218.

77. Raveh-Sadka T, Levo M, Segal E (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* 19: 14801496.

78. Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci* 105: 1725617261.

79. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci* 99: 1279512800.

80. Ou J, Furusawa C, Yomo T, Shimizu H (2009) Analysis of stochasticity in promoter activation by using a dual-fluorescence reporter system. *Biosystems* 97: 160164.

81. Schlax PJ, Capp MW, Record MT (1995) Inhibition of transcription initiation by lac repressor. *J Mol Biol* 245: 331350.

82. Saiz L, Vilar JM (2006) Stochastic dynamics of macromolecular-assembly networks. *Mol Syst Biol* 2: 2006.0024.

83. Malan TP, McClure WR (1984) Dual promoter control of the Escherichia coli lactose operon. *Cell* 39: 173180.

84. Vanzi F, Broggio C, Sacconi L, Pavone FS (2006) Lac repressor hinge flexibility and DNA looping: single molecule kinetics by tethered particle motion. *Nucleic Acids Res* 34: 34093420.

85. Moran U, Phillips R, Milo R (2010) SnapShot: Key numbers in biology. *Cell* 141: 1262.

86. Straney SB, Crothers DM (1987) Lac repressor is a transient gene-activating protein. *Cell* 51: 699707.

87. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329: 533538.

88. Gutierrez PS, Monteoliva D, Diambra L (2009) Role of cooperative binding on noise expression. *Phys Rev E* 80: 011914.

89. Müller D, Stelling J (2009) Precise regulation of gene expression dynamics favors complex promoter architectures. *PLoS Comput Biol* 5: e1000279.

90. Boeger H, Griesenbeck J, Kornberg RD (2008) Nucleosome retention and the stochastic

nature of promoter chromatin remodeling for transcription. *Cell* 133: 716726.

91. Li G, Levitus M, Bustamante C, Widom J (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol* 12: 4653.

92. Gansen A, Valeri A, Hauger F, Felekyan S, Kalinin, et al. (2009) Nucleosome disassembly intermediates characterized by single-molecule FRET. *Proc Natl Acad Sci* 106: 1530815313.

93. Klumpp S, Hwa T (2008) Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. *Proc Natl Acad Sci* 105: 1815918164.

94. Voliotis M, Cohen N, Molina-Paris C, Liverpool TB (2008) Fluctuations, pauses and backtracking in DNA transcription. *Biophys J* 94: 334348.

95. Dobrzynski M, Bruggeman F (2009) Elongation dynamics shape bursty transcription and translation. *Proc Natl Acad Sci* 106: 25832588.

96. Tkacik G, Gregor T, Bialek W (2008) The role of input noise in transcriptional regulation. *PLoS One* 3: e2774.

97. Zurla C, Manzo C, Dunlap D, Lewis DE, Adhya S, et al. (2009) Direct demonstration and quantification of long-range DNA looping by the lambda-bacteriophage repressor. *Nucleic Acids Res* 37: 27892795.

98. Müller J, Oehler S, Müller-Hill B (1996) Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *J Mol Biol* 257: 2129.

99. Kennell D, Riezman H (1977) Transcription and translation initiation frequencies of the Escherichia coli lac operon. *J Mol Biol* 114: 121.

Chapter 6

Promoter architecture dictates cell-to-cell variability in gene expression

6.1 Introduction

The molecular events underlying gene expression are inherently stochastic. Examples of such events include transcription factor (TF) binding and unbinding [1, 2]; RNA polymerase (RNAP) open complex formation, abortive transcript production, and promoter escape [3, 4]; and the formation and dissolution of transcription-factor-mediated DNA loops [5]. This stochasticity means that gene expression is in turn inherently stochastic. Over the past decade, an array of studies have investigated variability in gene expression [6–10] and the possible phenotypic consequences of transcriptional noise [6, 10–12]. It has further been postulated that noise in gene expression may affect the fitness of microbial populations by *e.g.* providing phenotypic variability in a population of genetically identical cells [13–17]. Against this backdrop, theorists have sought to elucidate how changes in molecular-kinetic parameters such as TF binding and unbinding rates affect variability in expression from their target gene [18–20], while experimentalists have measured variability in gene expression at both the mRNA and protein level in prokaryotes and eukaryotes [11, 21–25].

The theoretical efforts result in models of transcription that hinge on the molecular details of the promoter. These models make quantitative predictions for the level of variability as the details of the target gene's promoter are varied. For example, two extremely common promoter architectures are shown schematically in Figure 6.1. Each rate parameter (r, $k_{\mathrm{R}}^{\mathrm{off}}$, $k_{\mathrm{R}}^{\mathrm{on}}$ and γ) in this cartoon has a physical interpretation (Figure 6.1C) as an element that can be tuned independently by careful genetic manipulation. The effect of promoter architecture on mean levels of gene expression is well established in prokaryotes, where thermodynamic models incorporating details of promoter architecture such as the numbers, locations, and strengths of TF binding sites have been deployed successfully to predict gene expression as a function of these parameters [26–29]. However, the

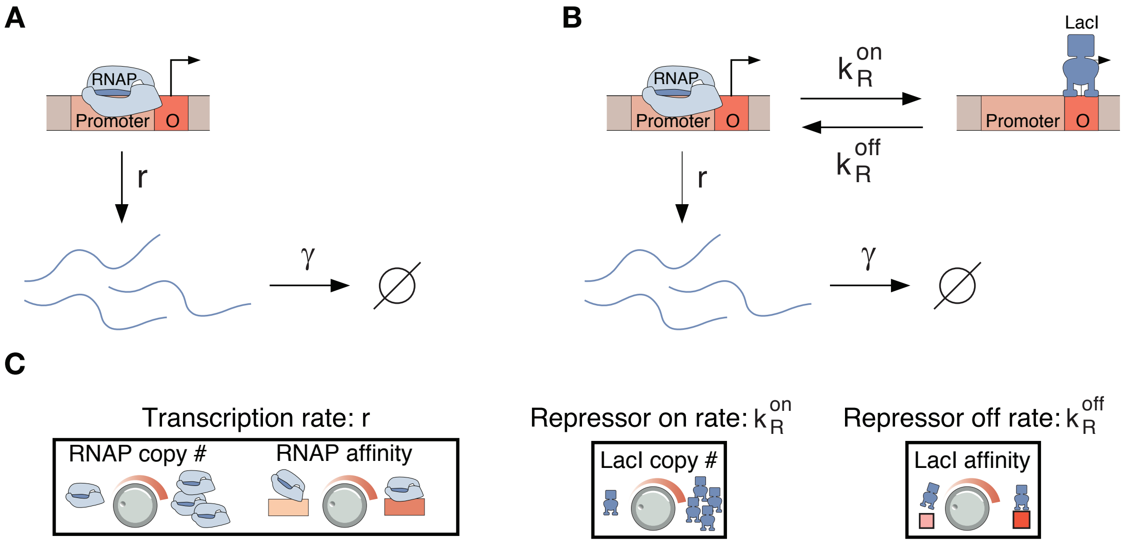


Figure 6.1: **Schematics of the kinetics of transcription for two simple regulatory architectures.** (**A**) Constitutive expression. mRNA are produced with constant probability per unit time at average rate r, and degraded at average rate γ per mRNA. (**B**) Gene controlled by binding of a repressor. The promoter takes two states: in the active state, left, the repressor is not bound and transcription and degradation occur as in constitutive expression. In the inactive state, right, the repressor is bound and transcription cannot occur. The promoter switches from the active to inactive state at rate k_{on}^{R} (the rate at which repressor binds on to the promoter) and from the inactive to active state at rate $k_{\text{R}}^{\text{off}}$ (the rate at which repressor dissociates off of the promoter). (**C**) Examples of the experimental knobs available for tuning the various model rate parameters.

associated predictions for how transcriptional noise depends on these parameters remains untested in any systematic way. In direct contrast, some high-throughput experiments have culminated in the assertion that the cell-to-cell variability in gene expression is "universal", dictated solely by the mean level of expression and insensitive to the details of the promoter driving the expression [9, 23, 24]. Thus, there is a serious divide between the experiments and theory in this field that can potentially be healed by a systematic study of transcriptional noise as any given genetically-accessible transcriptional knob is tuned.

Here, we have constructed a library of synthetic promoters driving a LacZ reporter in *Escherichia coli* and measured the resulting mRNA copy number distributions using single molecule mRNA fluorescence *in situ* hybridization (FISH) [30]. Crucially, changes in promoter sequence between constructs have clear interpretations in terms of the molecular parameters underlying transcription (*e.g.*, TF unbinding rate, basal transcription rate). This allows us to directly compare predictions of models incorporating those parameters with experimentally observed mRNA distributions, and hence to directly link the molecular events underlying transcription with observed variability in gene expression.

6.2 Results

6.2.1 Transcriptional Variability in Constitutive Expression

The theoretical foundation of our modeling efforts is the master equation detailing how the probability distribution of mRNA expression levels changes with respect to time [19, 20]. These changes are written in terms of molecular processes within the cell such as mRNA decay, and binding and unbinding of the proteins involved in transcription. As shown schematically in Figure 6.1A for the case of constitutive expression, mRNA transcripts are produced at average rate r and degraded at average rate γ, in both cases with constant probability per unit time. It can be shown [31] that the resulting steady-state mRNA copy number distribution is given by a Poisson distribution

$$P(m) = \frac{\lambda^m}{\lambda!} e^{-\lambda}, \tag{6.1}$$

where $P(m)$ is the probability of observing m mRNA in a cell, and $\lambda = r/\gamma$. This distribution has a mean r/γ. In the following experimental results, we will use the Fano factor, defined as the variance divided by the mean, to characterize variability in gene expression. This metric characterizes the fold change in the squared coefficient of variation (σ^2/μ^2) relative to a Poisson process, for which $\sigma^2/\mu^2 = 1/\mu$. Therefore the predicted Fano factor for constitutive expression equals one identically; namely,

$$\text{Fano} = \frac{\text{variance}}{\text{mean}} = 1. \tag{6.2}$$

This simple result is inconsistent with previous studies which have concluded that even constitutive mRNA production is non-Poissonian or "bursty," with observed Fano factors significantly greater than one [23, 32]. Moreover, we also observe Fano factor values greater than one in constitutive expression data, with a trend of higher Fano factors associated with higher expression levels (Figure 6.2A) (this data is discussed in more detail below). It is apparent that the preceding analysis encapsulated in equation 6.2 is incomplete, as we will now discuss. The schematics of Figure 6.1 represent the dynamics of the stochastic processes (TF binding/unbinding, mRNA degradation, transcription initiation) that contribute to so-called "intrinsic" variability in gene expression. However, rate parameters such as the repressor binding rate $k_{\mathrm{R}}^{\mathrm{on}}$ and transcription rate r are themselves subject to fluctuations due to cell-to-cell variability in repressor and RNAP copy numbers, respectively. Such effects, collectively termed "extrinsic variability," increase the measured variability and possibly obscure the intrinsic component of the variability arising from transcription itself [33, 34].

As we will see, the variability in gene copy number within a population due to chromosome replication is a particularly important source of extrinsic variability [35]. We thus turn to modeling its effect. In the following experiments, our reporter constructs are chromosomally integrated at the *galK* locus. At our growth rates (roughly 60 minutes) and chromosomal locus of integration [36],

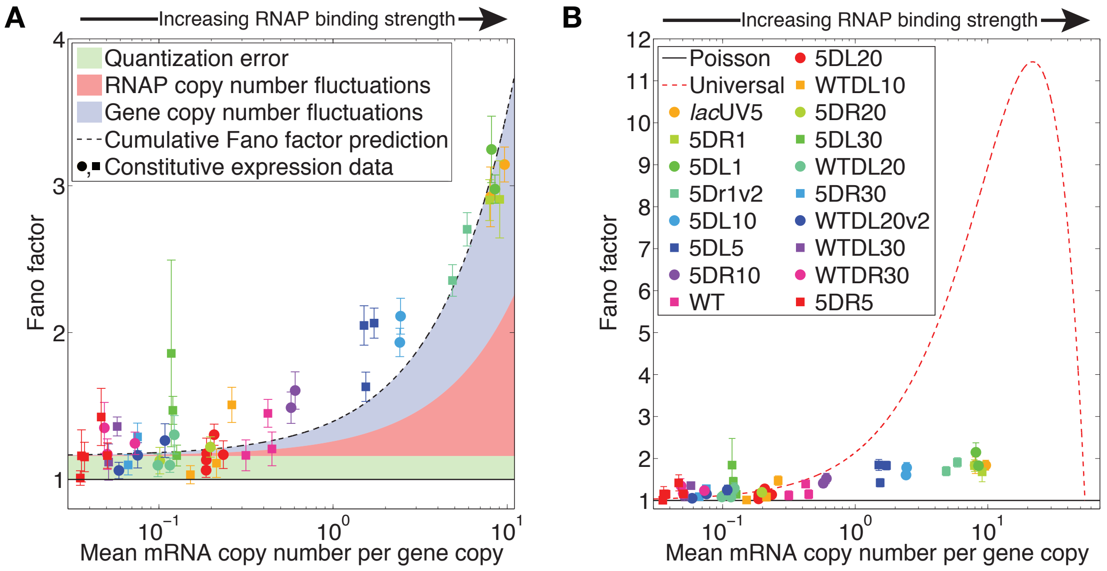


Figure 6.2: **Fano factor vs mean mRNA copy number for constitutive expression library**. (**A**) Fano factor (gene copy number variation not subtracted) vs. mean expression is plotted for each of 18 constitutive promoters along with estimates of the effects of gene copy number variation (blue), RNAP copy number fluctuations (red), and image analysis error (green). See supplementary text for details of these estimates. These factors can account for essentially the entirety of the deviation from Fano = 1. (**B**) Measured Fano factor for various promoters under constitutive expression, with gene copy number variation subtracted. Each strain is represented by a unique symbol and each instance represents repeated measurements. Error bars in Fano factor are the result of bootstrap sampling expression measurements of individual cells. For reference, the predictions of pure Poissonian production (black solid line) and the "universal noise" curve observed in [23] (red dashed) theories are shown. The data are clearly inconsistent with the "universal" curve and are reasonably well fit by the model of Poissonian transcription. Error bars in Fano factor in both (**A**) and (**B**) are the result of bootstrap sampling expression measurements of individual cells.

we expect that a cell has one copy of the reporter gene for the first $\approx 1/3$ of the cell cycle and two copies for the remainder of the cell cycle. We will characterize this with the variable $f = 2/3$, the fraction of the cell cycle for which two copies of the gene of interest exist. Then the probability distribution for mRNA copy number in our population of cells, $P(m)$, is generically

$$P(m) = (1 - f) \times p_1(m) + f \times p_2(m), \tag{6.3}$$

where $p_1(m)$ and $p_2(m)$ are the probability distributions for mRNA copy number when one or two copies of the gene are present, respectively. If the gene copies are assumed to be independent [37], the mean expression level from a population of genetically identical cells, obtained by summing equation 6.3 over all m, is $\langle m \rangle = (1 + f)\langle m \rangle_1$ where $(1 + f)$ is the mean gene copy number and $\langle m \rangle_1$ is the average expression expected from a single copy of the gene, $\langle m \rangle_1 = \sum_{m=0}^{\infty} m\, p_1(m)$. The noise in gene expression due to the variability in gene copy number can also be calculated in this general framework. It can be shown (supplementary text) that the effect of gene copy number variation on the variability in expression is independent and additive to the variability predicted from transcriptional noise such that

$$\text{Fano} = \underbrace{\frac{\langle m^2\rangle_1 - \langle m\rangle_1^2}{\langle m\rangle_1}}_{\text{Transcription}} + \underbrace{\frac{f(1-f)}{1+f}\langle m\rangle_1}_{\text{Gene copy number}}\,, \tag{6.4}$$

where $\langle m^2\rangle_1 - \langle m\rangle_1^2$ is the variance predicted from a single copy of the gene. The first term, labeled "Transcription", is simply the (promoter architecture dependent) Fano factor of a single copy of a gene. The predicted contribution to the Fano factor from copy number variation is labeled "Gene copy number". The effect is directly proportional to the mean expression per gene copy $\langle m\rangle_1$, with a proportionality constant that depends on f. As expected, if the copy number has a defined, static value ($f = 0$ or $f = 1$) there is no contribution to the Fano factor from gene copy number variation. At $f = 2/3$, corresponding to our locus of integration and growth conditions, the gene copy number contribution to the Fano factor is roughly 1.3 for the most highly expressed strain (and decreases with decreasing expression; see Figure 6.10).

With these results in hand, we return to our experimental investigation of constitutive expression. In a previous work [38], we demonstrated the ability to predictively tune the transcription rate r (and hence mean expression r/γ) by changing the RNAP binding site sequence. We designed a set of 18 constitutive promoters whose mean expression spanned nearly three orders of magnitude [38]. In order to quantitatively test the predictions of the model presented above for the dynamics of transcription under constitutive expression, we measure the resulting mRNA copy number distribution using mRNA FISH for each of these 18 unique promoters. Representative mRNA copy number histograms are shown in Figure 6.3 for each strain, with the strain names above each histogram for reference. For each strain, we plot the predicted mRNA copy number distributions both with (black lines) and without (blue dashed lines) accounting for gene copy number variation. Specifically, the blue curves are given by equation 6.1 with λ set equal to the observed mean $\langle m\rangle$ of each strain:

$$P(m) = (\langle m\rangle^m/\langle m\rangle!)\, e^{-\langle m\rangle}. \tag{6.5}$$

The black curves are given by combining the preceding equation with equation 6.3, namely,

$$P(m) = (1-f)(\langle m\rangle_1^m/\langle m\rangle_1!)\, e^{-\langle m\rangle_1} + (f)((2\langle m\rangle_1)^m/(2\langle m\rangle_1)!)\, e^{-2\langle m\rangle_1}, \tag{6.6}$$

where $\langle m\rangle_1 = \langle m\rangle/(1+f)$, $f = 2/3$, and $\langle m\rangle$ is the observed mean for each strain. It can readily be seen that accounting for gene copy number variation improves the agreement between theory and experiment without requiring additional free parameters. To quantify this improvement, we

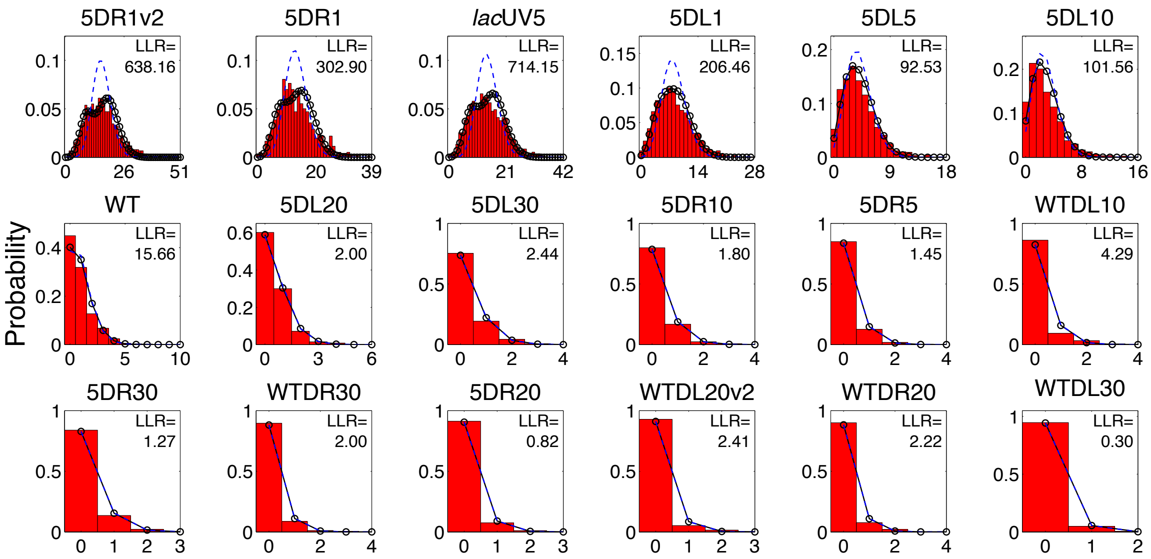


Figure 6.3: **mRNA copy number histograms for constitutive expression.** Observed mRNA copy number distributions for library of 18 constitutive promoters. For each promoter, we plot the predicted mRNA copy number distribution assuming Poissonian production and degradation, both with (black circles) and without (dashed blue lines) accounting for gene copy number variation. The log-likelihood ratio (LLR) of the observed data with and without accounting for copy number variation is shown on each histogram. Accounting for gene copy number variation substantially improves agreement between theory and data, as indicated by positive LLRs. The mean number of cells included in a histogram is 1238 ± 267 cells for each sample.

report the log-likelihood ratio (LLR) for each strain. This quantity is the logarithm of the ratio of the likelihood of the data given the variable copy number probability distribution (black circles, equation 6.6) to the likelihood of the data given by the simple Poisson prediction (blue dashed line, equation 6.5). LLR $= 0$ implies that the observed data is equally likely given either theoretical distribution, whereas LLR > 0 implies that the data is more likely to have been observed given the variable gene copy distribution of equation 6.6. We obtain positive LLR values for every strain, with the most positive values tending to occur at high mRNA expression, where the difference between equation 6.5 and equation 6.6 is most pronounced.

In Figure 6.2A, we plot the Fano factor vs. mean expression for each of this set of promoters. For clarity, each strain has a unique colored symbol (symbol key in Figure 6.2B) and the individual results of repeated measurements are shown. The solid black line is the bare Poisson prediction (equation 6.2) for the noise in constitutive expression. The shaded regions represent the effects of what we believe are the three most important additional sources of noise. The green shaded region, quantization error, is the variability introduced by our measurement and analysis process (supplementary text). The red shaded region covers the expected contribution from cell-to-cell differences in RNAP copy number (supplementary text) and the blue region is the expected contribution from gene copy number variation (described above). The data and theoretical predictions are in good accord, implying that the dynamics of constitutive transcription are, at a basic level, Poissonian with some additional extrinsic noise resulting from cell-to-cell inhomogeneities. We find no need to invoke transcriptional bursting or related hypotheses to explain the deviation from Fano = 1. Further corroborating this analysis, in Figure 6.14, we show that gene copy number noise is largely removed by gating on cell size such that all cells analyzed for a given data point have either one or two chromosome copies. Thus we feel confident that the gene copy number contribution to the overall variability is well characterized and we will subtract it from the observed data for the remainder of this work, in order to highlight the transcriptional contribution to the noise. In Figure 6.2B, we plot the Fano factor minus the predicted gene copy number contribution and observe a quantitative disagreement between the measured noise in expression to that predicted by the "universal" noise model as reported in [23]. But to conclusively demonstrate the architecture dependence of the variability we need to look at alternative regulatory architectures. To that end we now consider architectures controlled by a repressor.

6.2.2 Transcriptional variability in simple repression

We will next consider a slightly more complex architecture in which transcription can be blocked by a repressor TF. Despite the simplicity of this architecture, simple repression and activation are the most common regulatory motifs in *E. coli* after constitutive expression, and thus have direct physiological relevance [39]. As shown in Figure 6.1B, the promoter can transition between an active state in which no repressor is bound and transcription occurs at rate r (as in the constitutive case), and an inactive state in which a repressor is bound and transcription cannot occur. The promoter transitions from the active to inactive state at rate $k_{\mathrm{R}}^{\mathrm{on}}$, and from the inactive to active state at rate $k_{\mathrm{R}}^{\mathrm{off}}$. It can be shown [19] that the mean mRNA copy number is given by

$$\langle m \rangle = \frac{k_{\mathrm{R}}^{\mathrm{off}}}{k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}} \frac{r}{\gamma}, \tag{6.7}$$

where r/γ is the mean expression of the same gene constitutively expressed. The Fano factor is given by

$$\mathrm{Fano} = 1 + \frac{k_{\mathrm{R}}^{\mathrm{on}}}{\left(k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}\right)} \frac{r}{\left(\gamma + k_{\mathrm{R}}^{\mathrm{off}} + k_{\mathrm{R}}^{\mathrm{on}}\right)}. \tag{6.8}$$

An expression for the full probability distribution function can be found in reference [40]. If binding of repressor to its binding site is diffusion limited, then $k_{\mathrm{R}}^{\mathrm{on}}$ will be proportional to the concentration of repressor TF in the cell: $k_{\mathrm{R}}^{\mathrm{on}} = k_0[R]$. Similarly, $k_{\mathrm{R}}^{\mathrm{off}}$ will be a function of the interaction energy $\Delta\epsilon_R$ between the repressor and its binding site: $k_{\mathrm{R}}^{\mathrm{off}} \propto e^{\Delta\epsilon_R}$. This interaction energy is itself a function of the repressor binding site sequence. We can thus tune $k_{\mathrm{R}}^{\mathrm{on}}$ by changing the concentration of repressor in the cell (by, in this work, expressing it from an inducible promoter), and can tune

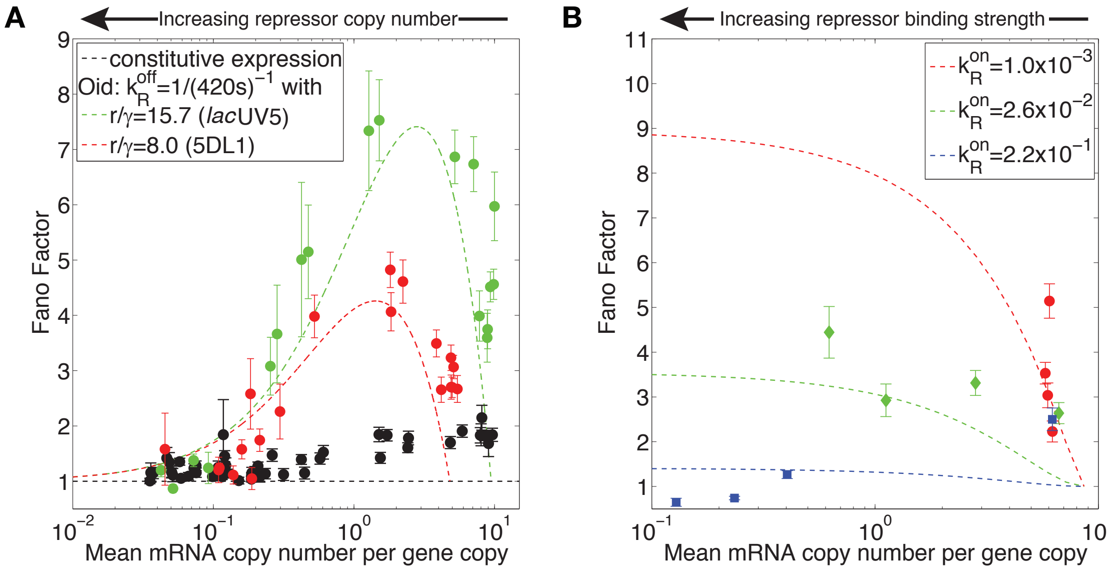


Figure 6.4: **Measurements of variability in gene expression for systematic tuning of promoter architecture**. (**A**) Fano factor vs. mean mRNA copy number for two promoters (choices of r/γ): 5DL1 (red points) and *lac*UV5 (green points) while tuning k_{on}^{R} by inducing LacI to varying levels. The parameter-free predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color holding promoter (r/γ) and repressor binding strength ($k_{\text{off}}^{\text{R}}$) constant. For reference, the black data is the constitutive data from Figure 6.2. (**B**) Fano factor vs. mean mRNA copy number for *lac*UV5 while tuning $k_{\text{off}}^{\text{R}}$ by changing repressor binding site identity at fixed repressor copy number. Each color is a different induction condition from red (lowest LacI induction) to blue (highest LacI induction). Again, the predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color. In both cases, the Fano factor at a given mean depends on the choice of molecular parameters and agrees with the expectations from theory. The effect of gene copy number variation was subtracted from all data points and error bars in Fano factor are the result of bootstrap sampling expression measurements of individual cells.

$k_{\text{R}}^{\text{off}}$ by changing the repressor binding site sequence. It is important to note that these physical interpretations of kinetic parameters are for the moment only tentative assumptions; however, these assumptions will be subjected to rigorous experimental scrutiny.

The mean expression $\langle m \rangle$ can be tuned via either of the parameters k_{R}^{on} or $k_{\text{R}}^{\text{off}}$. Importantly, the predicted relationship between the mean and the Fano factor is unique depending on which of these rates is being tuned to explore the range of means (Figure 6.4A and B, dashed lines). To test the predicted effect of changing of k_{R}^{on}, we take two of the constitutive promoters from the preceding section (*lac*UV5 and 5DL1) and place them under simple repression via a LacI Oid binding site immediately downstream of the promoter [30]. The difference in transcription rate for the two constructs (*lac*UV5 and 5DL1) is reflected in different values of r/γ. When LacI is bound to its binding site, transcription is prevented via steric inhibition, as schematized in Figure 6.1B. At the same time, we introduced into our constructs a genetic circuit enabling inducible control of LacI expression via the small molecule anhydrotetracycline (aTc). In Figure 6.4A, we plot the measured Fano factor (with the effect of gene copy number variation subtracted) as a function of the mean expression over LacI concentrations ranging from ≈ 0 to 80 nM, for both promoters. In addition, we plot the zero-free-parameter theoretical prediction for the Fano factor as a function of mean using the measured value of r/γ from the constitutive data and the LacI unbinding rate from reference [41]. We find excellent agreement between model predictions and observed data.

To test the effect of changing $k_{\mathrm{R}}^{\mathrm{off}}$ on variability, we can change the LacI-DNA binding energy by altering the sequence of the LacI binding site. Holding the RNAP binding site constant (and thus $\langle m\rangle_{\max} = r/\gamma$ constant), we created constructs corresponding to four different LacI binding sites (the LacI Oid, O1, O2, and O3 sites) [30]. At constant repressor concentration (*i.e.*, constant $k_{\mathrm{R}}^{\mathrm{on}}$), tuning mean expression by altering $k_{\mathrm{R}}^{\mathrm{off}}$ is predicted to yield a characteristic curve, while different repressor concentrations (and hence $k_{\mathrm{R}}^{\mathrm{on}}$ values) correspond to distinct instances of this curve (see Figure 6.4B). We measure the mRNA distributions resulting from changing $k_{\mathrm{R}}^{\mathrm{off}}$ at each of three different repressor concentrations. Since different repressor concentrations are achieved by varying the degree of induction, we exploit the well-characterized nature of simple repression [38, 42] to infer the intracellular repressor concentration by measuring the fold-change in expression for our strains at each inducer concentration. The remainder of the molecular rates are available in the literature [41] (see also supplementary text and Table 6.2). Figure 6.4B shows the measured Fano factor (again, with gene copy variability subtracted) vs. mean expression for the three distinct repressor concentrations. Points of the same color differ in their repressor binding site. We find agreement in the trends between theory and experiment, although less good than in the case of tuning $k_{\mathrm{R}}^{\mathrm{on}}$. One possible explanation comes from recent work showing that changing TF-DNA binding affinity affects the TF-DNA association rate $k_{\mathrm{R}}^{\mathrm{on}}$ as well as the dissociation rate $k_{\mathrm{R}}^{\mathrm{off}}$ [41]. Thus, our assumption that the same value of $k_{\mathrm{R}}^{\mathrm{on}}$ obtains along each of the red, green, and blue curves is probably not strictly correct, and this is reflected in less precise agreement between theory and experiment. That being said, the most important outcome of this set of measurements is a demonstration of the qualitatively distinct variability profile when a different set of transcriptional parameters are controlled, illustrating once again the systematic dependence of variability on promoter architecture.

6.3 Discussion

We have demonstrated the ability to predict mRNA copy number distributions based on biophysical models of the underlying molecular events. Here we have shown that transcriptional noise is well predicted by molecularly detailed models for the two most common promoter architectures in *E. coli* as the various genetic knobs are tuned. Moreover, this agreement is emphatically not the result of fitting theory curves to data, since the predicted curves are generated using physical parameter values reported elsewhere in the literature and in that sense are zero-parameter predictions for the expected variability in gene expression. For instance, the repressor binding rate $k_{\mathrm{R}}^{\mathrm{on}}$ was measured using fluorescence microscopy with fluorescently tagged LacI molecules, an experiment wholly unrelated to measuring mRNA distributions [2]. This transferability between disparate experimental contexts gives us confidence that what we have constructed is not simply a "phenomenological model" but instead is a real, physical description of the molecular processes underlying transcription. Somewhat surprisingly, we are able to omit a number of potential sources of variability, including extrinsic noise due TF copy number fluctuations (although, in the case of constitutive expression, we do seem to inherit some measurable noise from RNAP fluctuations), DNA supercoiling, and DNA condensation due to nucleoid-associated proteins like HU and H-NS. Despite these omissions, we see good agreement between model predictions and observed data.

Although previous studies have observed that transcription is "bursty" even in the case of fully induced expression [23, 32], we do not find this to be the case. The claim of "bursty" constitutive transcription is based on the observation that the Fano factor is greater than one for constitutive mRNA production (as well as direct kinetic measurements). Various explanatory hypotheses have been proposed, including transcriptional silencing via DNA condensation by nucleoid proteins [43], negative supercoiling induced by transcription, or the formation of long-lived "dead-end" initiation complexes [44]. Although our data does not completely rule out these hypotheses, we find that gene copy number variation is sufficient to explain most of the deviation from Fano = 1 in our constitutive expression data. If furthermore we incorporate a simple model of the effects of RNAP copy number fluctuations (supplementary text), and account for the contribution of measurement error (supplementary text), we find that we can explain essentially the entirety of the deviation from Fano = 1 (Figure 6.2B). Thus, we find no need to invoke these alternative hypotheses to explain the observed "burstiness" of constitutive transcription. Indeed, "burstiness" would be something of a misnomer in this case, as what we are describing is not so much periods of inactivity punctuated by bursts of activity, as it is constant activity at a fluctuating average rate.

We wish also to highlight a philosophical difference between this work and other recent studies in the field. Recent work from other labs has examined the noise properties of a broad swath of naturally occurring *E. coli* promoters [9, 23]. In [23], the authors examined a number of promoters under varying induction conditions and concluded that variability as a function of the mean follows a "universal" curve described by an effective two-state model. From our perspective, these experiments make it difficult to interpret differences between promoters and induction conditions in terms of distinct physical parameters because of the wide variety of promoter architectures in play as well as the diverse mechanisms of induction. In this work, we have instead taken a "synthetic biology" approach of building promoters from the ground up. By directly controlling aspects of the promoter architecture, our goal has been to directly relate changes in promoter architecture to changes in observed gene expression variability. We believe that this work has convincingly demonstrated our ability to do so, and leads us inexorably to the conclusions that variability in gene expression does depend on promoter architecture, and that mutations in regulatory DNA can alter gene expression noise. This suggests that gene expression noise may be a tunable property subject to evolutionary selection pressure, as mutations in regulatory DNA could provide greater fitness by increasing (or decreasing) variability. Demonstrating the relevance of this hypothesis in natural environments and extending these results to the often complex and poorly characterized promoter architectures of wild-type *E. coli* and other organisms remain ongoing challenges for researchers in this field.

6.4 Materials and Methods

6.4.1 Strains

The genetic modifications used to create the strains used in the various sections of the main paper are listed below with a bold heading characterizing which parameter is tuned within that class of strains.

Constitutive expression: tuning r. As described in [38], promoter constructs consist of an RNAP binding site with a LacI O2 binding site immediately downstream of the transcription start site, as shown in Figure 6.1A. (The O2 binding site does not affect expression because *lacI* is deleted from the host strain used in gene expression measurements.) Expression of a LacZ reporter is tuned over a factor of ≈ 500 by changing the DNA sequence of the RNAP binding site. Promoter + reporter constructs are chromosomally integrated at the *gal* locus in an *E. coli* strain (HG105) in which *lacI* and *lacZYA* are deleted.

Simple repression: tuning $k_{\mathbf{R}}^{\mathbf{on}}$. In two of the constitutive promoter constructs (*lac*UV5 and 5DL1) the O2 LacI binding site is replaced with an Oid LacI binding site. This construct is integrated into an *E. coli* strain (RCB110) in which wild-type *lacIZYA* is deleted as before, but with the addition of a genetic circuit allowing inducible control of LacI expression. This circuit consists of two components: first, the *tet* repressor TetR is chromosomally integrated at the *gspI* locus under the control of the strong constitutive promoter $\mathrm{P_{N25}}$; and second, LacI is integrated at the *ybcN* locus under the control of the $\mathrm{P_{LtetO\text{-}1}}$ promoter [45], which is repressed by TetR. Expression of LacI can thus be induced by the small molecule anhydrotetracycline (aTc), which interacts with TetR and prevents it from repressing transcription of LacI. The ribosomal binding sequence of the LacI is the "1147" version from reference [42]: at full induction this produces roughly 40 LacI per cell. By varying the aTc concentration, we tune mean gene expression by tuning the intracellular concentration of LacI, yielding the curves shown in Figure 6.4A.

Simple repression: tuning $k_{\mathbf{R}}^{\mathbf{off}}$. These measurements exploit the same strain (RCB110) from the k_R^{on} tuning; however in this case we use only the *lac*UV5 promoter strain and create constructs in which the O2 LacI binding site is replaced by O1, O3 or Oid. These constructs are integrated into the *galK* locus as before, yielding four constructs total each with a different LacI binding site. These strains are measured at constant inducer concentration (to achieve equal repressor copy numbers across all samples) for three distinct induction conditions. This was done for each of three different LacI concentrations as shown in Figure 6.4B.

6.4.2 Growth

Cultures are grown overnight to saturation (at least 8 hours) in LB and diluted 1:4000 into 30 mL of M9 minimal media supplemented with 0.5% glucose in a 125mL baffled flask. Growth in minimal media continues approximately 8 hours and cells are harvested in exponential phase when OD600= 0.3 − 0.5 is reached.

6.4.3 mRNA FISH

6.4.3.1 Fixation and labeling

Our assay is based on that used in reference [23]. Once a culture reaches OD600= 0.3 − 0.5, it is immersed in ice for 15 minutes before being harvested in a large centrifuge chilled to $4°C$ for 5 minutes at 4500 g. The cells are then fixed by resuspending in 1 mL of 3.7% formaldehyde in 1x PBS which is then allowed to mix gently at room temperature for 30 minutes. Next, they are centrifuged (8 minutes at 400 g) and washed twice in 1 mL of 1x PBS twice. The cells are permeabilized by resuspension in 70% Ethanol which proceeds, with mixing, for 1 hour at room temperature. The cells are then pelleted (centrifuge at 600 g for 7 minutes) and resuspended in 1 mL of 20% wash solution (200 μL formamide, 100 μL 20x SSC, 700 μL water). This mixture is allowed to sit several minutes before centrifugation (7 minutes at 600g) and resuspended in 50 μL of DNA probes (consisting of a mix of 72 unique DNA probes; individual oligo sequences listed in Table 6.1) labeled with ATTO532 dye (Atto-tec) in hybridization solution (0.1 g dextran sulfate, 0.2 mL formamide, 1 mg *E.coli* tRNA, 0.1 mL 20x SSC, 0.2 mg BSA, 10 μL of 200 mM Ribonucleoside vanadyl complex). This hybridization reaction is allowed to proceed overnight. The hybridized product is then washed four times in 20% wash solution before imaging in 2x SSC.

6.4.3.2 FISH data acquisition

Samples are imaged on a 1.5% agarose pad made from PBS buffer. Each field of view is imaged with phase contrast at the focal plane and with 532 nm epifluorescence (Verdi V2 laser, Coherent Inc.) both at the focal plane and in 8 z-slices spaced 200 nm above and below the focal plane (for a total of 17 slices), sufficient to cover the entire depth of the *E. coli.* The images are taken with an EMCCD camera (Andor Ixon2) under 150× magnification. The phase image is used for cell segmentation and the fluorescence images are used in mRNA detection. A total of 100 unique fields of view are imaged in each sample and a typical field of view has between 5 and 15 viable cells (cells which are touching and cells that have visibly begun to divide are ignored) resulting in roughly 900 individual cells per sample on average. However, due to differences in plating density and position quality, the actual number can vary. A histogram of the sample size for all samples in this study is shown in Figure 6.5.

6.4.3.3 FISH analysis

The FISH data is analyzed in a series of Matlab (The Mathworks) routines. The overview of the workflow is as follows: identifying individual cells, segmenting the fluorescence to identify possible mRNA, quantifying the mRNA which are found (because of the small size of *E. coli*, at high copy number mRNA can be difficult to distinguish and count by eye).

Cell identification and segmentation: In phase contrast imaging, *E. coli* are easily distinguishable from the background and automated programs can identify, segment and label cells with high fidelity. The results of the phase segmentation are manually checked for accuracy: cells which are touching or overlapping other cells, misidentification of cells or their boundaries or cells which have visibly begun to undergo division are all discarded manually.

Fluorescence segmentation: First we perform several steps to process the raw intensity images. The images are flattened, a process to correct for any uneven elements in the illumination profile, using a flattening image. The flattening image is an average over $10-15$ images of an agarose pad coated with a small drop of fluorescein (such that the drop spreads evenly across most of the pad); the resulting image is a map of illumination intensity at any given pixel I_{flat}. Each pixel of every fluorescence image is scaled such that the corresponding pixel in the flattening image would be of a uniform brightness (typically each pixel is scaled up to the level of the brightest pixel). This can be achieved by renormalizing each pixel in the data images and dividing by the ratio of the intensity of the corresponding pixel in the flattening image to the intensity of the brightest pixel. In other words the raw images, I, are renormalized such that for pixel i, j with raw intensity $I^{(i,j)}$,

$$I^{(i,j)}_{\text{corrected}} = \left(I^{(i,j)} - I^{(i,j)}_{\text{dark}}\right) \times \left(\frac{\max_{i,j}(I^{(i,j)}_{\text{flat}} - I^{(i,j)}_{\text{dark}})}{I^{(i,j)}_{\text{flat}} - I^{(i,j)}_{\text{dark}}}\right), \tag{6.9}$$

where I_{dark} corresponds to an image taken with no illumination (mostly these counts are from camera offset). In the preceding equation, the first term in parentheses is the signal from the i, jth pixel, while the second term in parentheses corrects the signal for nonuniform illumination using the flattening image. We then subtract from every pixel the contribution to our signal associated with autofluorescence. The value for the autofluorescence is obtained by averaging over the fluorescence of every pixel in a control sample (one which underwent the entire FISH protocol but did not possess the *lacZ* gene). Finally, all local 3D maxima (where $x - y$ is the image plane) in fluorescence are identified. We require that the maxima be above a threshold in fluorescence (typically $300 - 400\%$ above the background autofluorescence signal). This threshold eliminates all fluorescence maxima in the control sample, which does not contain the *lacZ* gene.

mRNA quantification: Each identified maximum pixel is dilated in the image plane to a 5×5 box of surrounding pixels. These 5×5 boxes are referred to as "spots". If multiple spots overlap, the pixels which make up each overlapping spot are merged into one larger spot to avoid double counting the signal from any one pixel. Since, due to the small size of the *E. coli* we cannot guarantee that every spot corresponds to exactly one mRNA, we must have a way to quantify the relation between signal and mRNA copy number. An example histogram of the intensity of identified spots is shown in Figure 6.6A. The histogram has two clear peaks in probability: one corresponding to background noise, at approximately zero intensity, and the other corresponding to the intensity of a single mRNA. The low intensity peak, corresponding to background noise, is removed by thresholding the spots and rejecting spots that are less bright than the threshold. The threshold is selected to eliminate spots in a control sample that does not contain the target mRNA. However, we find choice of this threshold does not alter our results significantly since these spots are already significantly dimmer than an mRNA. To determine the calibration between signal intensity and mRNA copy number we take an average over all remaining isolated spots (meaning no merge events with other, nearby spots) in very low expression samples (where the mean $\ll 1$ and mRNA are statistically unlikely to overlap); see Figure 6.7A and B for examples of these thresholded histograms. Once this single mRNA intensity value is identified, when possible we also verify in other low expression strains that as we increase the mean expression it simply increases the frequency of spots with the single mRNA intensity but does not increase the mean intensity of each spot. An example of this is shown in Figure 6.7A where the single spot intensity histograms of seven unique strains are shown with each histogram normalized by the number of cells in each sample. The growing peak at 1 mRNA shows that as we transition from low expression towards having upwards of 1 mRNA per cell, the increased signal is primarily due to an increased number of identified single mRNA, although some brighter spots begin to appear corresponding to multiple mRNA per spot. Normalizing these same histograms by the total number of identified spots, as shown in fig 6.7B, demonstrates that the identified spots have the same character in each sample regardless of mean. The dashed black line shows the result of a Gaussian fit to the combined data from all seven samples in the figure. Finally, the day-to-day variability in these histograms is shown in Figure 6.7C for five different acquisitions on two distinct constitutive expression strains.

With this calibration in hand, we sum the signal from all identified spots in a given cell and determine how many mRNA are in that cell by dividing by the single mRNA intensity calibration found previously. A different technique, used in some studies, is to first quantify every individual identified spot then determine the copy number by summing the total number of identified spots in each cell. Figure 6.8 shows a direct comparison of these methods on simulated data (crosses) and real data (circles) with red corresponding to the “whole cell” method used in our study, where signal is summed over the whole cell and black corresponding to the single spot analysis where each spot is quantized and rounded to the nearest whole number of mRNA. The methods gives roughly identical values for the mean; however, the corresponding Fano factor is very slightly systematically higher using the single spot analysis, probably due to the rounding performed on each spot.

6.4.4 Miller LacZ assay

Concurrent with the mRNA FISH protocol, each sample also has LacZ activity measured by Miller assay. The protocol is identical to that described previously [38, 42], which is a slightly modified from that described in Ref [46]. Once cultures are ready for measurement, the OD600 of each sample is recorded. Next, a volume of cells between 5 μL and 200 μL is added to Z-buffer (60mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol, pH 7.0) to reach a total of 1 mL in a 1.5 mL Eppendorf tube. The time of the enzymatic reaction is inversely proportional to the volume of cells, and thus low expression samples require more volume and high expression samples require less to ensure that the time of reaction is reasonable ($\sim 1-10$'s of hours) to avoid measurement uncertainty, and to ensure that the yellow color is easily distinguishable from a blank sample of 1 mL of Z-buffer. The cells are lysed with 25 μL of 0.1% SDS and 50 μL of chloroform, mixed by a 10s vortex. To begin the reaction, 200 μL of 4mg/mL 2-nitrophenyl β-D-galactopyranoside (ONPG) in Z-buffer is added to the Eppendorf tube. The tube is monitored for the development of yellow color and once sufficient yellow has developed in a sample (sufficient absorbance at 420nm, without saturating the reading), the reaction is stopped through the addition of 200 μL of 2.5 M Na_2CO_3. Once all samples have been stopped, cell debris is removed from the supernatant by centrifugation at $> 13,000$ g for 3 min. 200 μL of each sample, including the blank which contains no cells, are loaded into a 96 well plate and absorbance at 420nm and 550nm is measured for each well with a Tecan Safire2. The LacZ activity in Miller units is then,

$$MU = 1000\frac{OD420 - 1.75 \times OD550}{t \times v \times OD600}0.826, \tag{6.10}$$

where t is the reaction time in minutes, v is the volume of cells used in mL and OD refers to the optical density measurements obtained from the plate reader. The factor of 0.826 accounts for the use of 200 muL Na_2CO_3 as opposed to 500 μL which changes the concentration of ONPG in the final solution. While some alternative protocols involve time-resolved measurements of LacZ activity over a range of cell densities, we believe the protocol used here is a simple and accurate method for providing a consistent relative calibration for our mRNA measurements that has been shown to be equivalent in terms of accuracy and reproducibility to more complicated and time-consuming measurement protocols [42].

6.5 Supplementary text

6.5.1 Calibration of mRNA FISH data versus Miller assay

As a test of our ability to accurately measure mean mRNA copy number with mRNA FISH, we directly compare our results to simultaneously acquired measurements of mean LacZ enzymatic activity (the protein produced by the mRNA targeted in our FISH labelling). In Figure 6.9, we show the mean mRNA expression vs. mean LacZ activity in Miller units for every measurement of every strain in this study. These two measurements of expression give consistent results as demonstrated by direct proportionality between these two measurement techniques over several orders of magnitude of expression. Error bars represent the standard deviation from repeated measurements.

6.5.2 Estimation of additional noise sources

6.5.2.1 Quantification error in image analysis

As described in the main text, the intensity of a single mRNA molecule is determined using a low mRNA expression sample such that detected spots are most likely to contain either zero (*i.e.* the fluorescence maxima is due to background noise) or one mRNA. A representative histogram of detected spots is shown in Figure 6.6A. However, this identification and quantification process is not without uncertainty of its own. While ideally each spot should have the same clear value for its integrated intensity, the intensity of a given identified mRNA varies significantly, and we attribute this to factors such as fluctuations in probe hybridization efficiency and non-specifically bound probes. We wish to estimate the effect that variability in the single mRNA intensity has on the overall observed variability. In order to do so, we will make the following assumptions:

- mRNA copy numbers are distributed with mean μ_m and variance σ_m^2. Aside from this, we make no assumptions about the specific form of the mRNA distribution.
- Integrated intensities for a single mRNA are distributed with mean $\langle I_1 \rangle$ and variance σ_I^2.
- Intensities for more than one mRNA are independent and additive. For cells containing k mRNA, integrated intensities have mean $k\langle I_1 \rangle$ and variance $k\sigma_I^2$ (since variances add for independent random variables).

In this work we sought to measure the Fano factor for various mRNA copy number distributions. However, what we actually measure experimentally (as described above) is the following:

$$\mathrm{Fano}_{\mathrm{exp}} = \mathrm{Fano}\left(\frac{I}{\langle I_1 \rangle}\right) = \frac{1}{\langle I_1 \rangle}\frac{\mathrm{var}(I)}{\langle I \rangle}, \tag{6.11}$$

where I is a random variable denoting the integrated intensity of a cell, $\langle I_1 \rangle$ is the mean intensity of a single mRNA. That is, we measure the Fano factor of a set of observed integrated intensities divided by the single mRNA intensity. We will now use the assumptions listed above to further investigate equation 6.11, and proceed by computing the mean and variance of I.

The random variable I is distributed according to:

$$P(I) = \sum_{k=0}^{\infty} p_m(k) p_I(I|k) \tag{6.12}$$

where $p_m(k)$ is the probability that a cell contains k mRNA, and $p_I(I|k)$ is the probability of obtaining intensity I given that a cell contains k mRNA. We will denote the conditional expectation of the nth moment of I given k as $\langle I^n|k\rangle$: *i.e.*, $\langle I^n|k\rangle = \int I^n p_I(I|k)dI$. Then the expected value of I is given by

$$\langle I\rangle = \int_{-\infty}^{\infty} IP(I)dI, \tag{6.13}$$

$$\langle I\rangle = \int_{-\infty}^{\infty} I \sum_{k=0}^{\infty} p_m(k) p_I(I|k)dI, \tag{6.14}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k) \int Ip_I(I|k)dI, \tag{6.15}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k)\langle I|k\rangle, \tag{6.16}$$

$$\langle I\rangle = \sum_{k=0}^{\infty} p_m(k)k\langle I_1\rangle = \langle k\rangle\langle I_1\rangle, \tag{6.17}$$

$$\langle I\rangle = \mu_m\langle I_1\rangle, \tag{6.18}$$

which is exactly what one would naively expect (the mean integrated intensity equals the mean number of mRNA times the mean single mRNA intensity). To compute the variance of I, we next need to compute the expected value of I^2:

$$\langle I^2\rangle = \int_{-\infty}^{\infty} I^2 P(I)dI, \tag{6.19}$$

$$\langle I^2\rangle = \int_{-\infty}^{\infty} I^2 \sum_{k=0}^{\infty} p_m(k) p_I(I|k)dI, \tag{6.20}$$

$$\langle I^2\rangle = \sum_{k=0}^{\infty} p_m(k) \int_{-\infty}^{\infty} I^2 p_I(I|k)dI, \tag{6.21}$$

$$\langle I^2\rangle = \sum_{k=0}^{\infty} p_m(k)\langle I^2|k\rangle. \tag{6.22}$$

According to the assumptions above,

$$\text{var}(I|k) = \langle I^2|k\rangle - \langle I|k\rangle^2 = k\sigma_I^2 \tag{6.23}$$

and hence,

$$\langle I^2|k\rangle = k^2\langle I_1\rangle^2 + k\sigma_I^2. \tag{6.24}$$

Plugging this back into equation 6.22, we obtain:

$$\langle I^2\rangle = \langle I_1\rangle^2 \sum_{k=0}^{\infty} k^2 p_m(k) + \sigma_I^2 \sum_{k=0}^{\infty} k p_m(k), \tag{6.25}$$

$$\langle I^2\rangle = \langle k^2\rangle\langle I_1\rangle^2 + \langle k\rangle\sigma_I^2, \tag{6.26}$$

$$\langle I^2\rangle = (\mu_m^2 + \sigma_m^2)\langle I_1\rangle^2 + \mu_m\sigma_I^2, \tag{6.27}$$

where we have used the fact that $\langle k^2\rangle = \sigma_m^2 + \mu_m^2$. Hence

$$\text{var}(I) = (\mu_m^2 + \sigma_m^2)\langle I_1\rangle^2 + \mu_m\sigma_I^2 - \mu_m^2\langle I_1\rangle^2, \tag{6.28}$$

$$\text{var}(I) = \sigma_m^2\langle I_1\rangle^2 + \mu_m\sigma_I^2, \tag{6.29}$$

and

$$\text{Fano}_{\text{exp}} = \frac{1}{\langle I_1\rangle}\frac{\sigma_m^2\langle I_1\rangle^2 + \mu_m\sigma_I^2}{\mu_m\langle I_1\rangle}, \tag{6.30}$$

$$\text{Fano}_{\text{exp}} = \frac{\sigma_m^2}{\mu_m} + \frac{\sigma_I^2}{\langle I_1\rangle^2}. \tag{6.31}$$

The two terms in equation 6.31 have simple interpretations. The first term is the Fano factor of the actual underlying mRNA distribution. The second term reflects uncertainty in the intensity of a single mRNA spot and is essentially the squared coefficient of variation of the intensity of a single mRNA spot. This value depends slightly on the conditions of the specific acquisition. For instance, the single mRNA peaks from one experiment are shown in Figure 6.7B. For this acquisition, one can fit a Gaussian to the observed single spot mRNA intensity distribution (dashed black line) and make a measurement of both the mean and standard deviation of this distribution to calculate the expected contribution to the Fano factor from quantization error, as in equation 6.31. For this acquisition, we see that $\sigma_I^2/\langle I_1\rangle^2 = 0.16$. This result is typical (as demonstrated in Figure 6.7C) and therefore the green shaded region in Figure 6.2A has a height equal to 0.16. Of course, this value is not static and depends on the particular acquisition conditions and could be calculated for each separate acquisition independently. However, these values are small enough relative to the range of Fano factors (≈ one to eight) observed in our experiments that this effect will not change the qualitative conclusions reached in this work.

As a complementary test of the performance of our image analysis routines, we created simulated FISH data sets at a variety of mRNA expression levels. Our goal was as much as possible to faithfully reproduce the images coming off of our microscope. We acquire raw microscopy data by spotting FISHed *E. coli* cells on agarose pads, mounting the cells on the microscope, and running an automated acquisition script. The script generates a grid of ≈ 100 positions on each pad; at each position, a phase contrast image is taken for segmentation purposes, followed by a fluorescence z stack (separated by 0.2 μm) to image mRNAs. Our simulated data thus also consisted of sets of ≈ 100 "positions", with each position consisting of a simulated phase contrast image and a simulated fluorescence z stack. The data generation algorithm at each position can be roughly described as follows:

1. Generate a phase contrast image. 25 "*E. coli*" cells are placed at random in a field of view. Cells are modeled as ellipsoids 22 pixels in length, 10 pixels wide, and 4 pixels tall.

2. Determine the number of mRNA copies in each cell. For each cell, the number of mRNA it will contain is drawn from the appropriate probability distribution (for instance, from a Poisson distribution with a given mean).

3. Determine the spatial distribution of mRNAs within a cell. For each cell, mRNA are distributed uniformly at random within the cell. For instance, if a cell has four mRNA assigned to it, then four pixels within the cell are chosen at random, with each one corresponding to the center of an mRNA.

4. Determine the intensity of each mRNA. As seen in Figure 6.6, the integrated fluorescence intensity of a single mRNA can vary substantially. We choose the intensity of each mRNA from a Gaussian distribution with mean 0.4 and standard deviation 0.16 (thus $\sigma_I^2/\langle I_1\rangle^2 = 0.16$ as in Figure 6.7B). These values were chosen as reasonable representations of our physical data sets. Each mRNA pixel (as determined in the previous step) is assigned a fluorescence intensity drawn from this distribution.

5. Convolve with point-spread-function. In reality mRNA do not show up as single bright pixels but rather as diffraction-limited spots. To simulate this we convolve the fluorescence stack with a Gaussian point-spread function with a standard deviation of 0.875 pixels. This value was chosen as a reasonable representation of the point spread functions observed in our actual data.

6. Generate background and noise. In addition to the signal from actual mRNA molecules, our images are subject to background from cellular autofluorescence and the agarose pad, as well as noise from unbound or non-specifically bound fluorescent probes. To simulate this, a random fluorescence background is generated for each cell by drawing pixel values from a geometric distribution with mean 466 (reflecting typical mean background fluorescences encountered in experimental data), convolving with a Gaussian with mean 1.0 pixels (to reflect spatially correlated noise from e.g. unbound probes), then adding these values to the "signal" as determined in the previous step. This background is added to a constant offset of 1080 counts to mimic a typical camera offset.

In Figure 6.8 we show the measured Fano factor for simulated data for a population of cells with Poisson distributed mRNA copy number (circles) using two distinct mRNA quantification schemes as described in the methods. As expected, even at low mRNA expression, the measured Fano factor is greater than the correct value of one, due to the variability in intensity measured for a single mRNA. The single mRNA intensity distribution is a Gaussian with $\sigma_I^2/\langle I_1\rangle^2 = 0.16$ and is designed to mimic our experimental data (see Figure 6.6B). Equation 6.31 thus predicts that the measured Fano factor will be 1.16, this prediction is shown as the green bar in Figure 6.8 for comparison to the Fano factor in the simulated data. For reference, our Fano factor measurements (with gene copy number noise subtracted) for the constitutive expression strains are shown as crosses. While the quantification noise matches at low means, at higher means RNAP fluctuations in the real data are likely also contributing to the measured noise and pushing the experimental noise above the simulations' noise.

6.5.2.2 Gene copy number variation

As cells grow and divide, their chromosomes are replicated, causing the copy number of a given gene to change over the course of the cell cycle. This effect can potentially obscure our measurements of transcriptional noise. To that end, we wish to calculate the effect of changes in gene copy number on variability in gene expression. Under the growth conditions of our experiments, *E. coli* cells contain one or two chromosomes, and hence one or two copies of the gene of interest. Let $1 - f$ denote the fraction of the cell cycle for which one copy of the gene of interest is present. Then f is the fraction for which two gene copies are present. The probability that m mRNA are present in a cell is given by

$$p(m) = (1-f)\, p_1(m) + f\, p_2(m), \tag{6.32}$$

where $p_1(m)$ is the probability of m mRNA given one gene copy, and $p_2(m)$ is the probability of m mRNA given two gene copies. We will assume that, when two gene copies are present, expression from the two copies is statistically independent. Thus, we can use well-known properties of sums of independent random variables to calculate properties of $p_2(m)$.

We will proceed by computing the mean and variance of $p(m)$ given in equation 6.32.

$$\begin{aligned} \langle m\rangle = \sum_{m=0}^{\infty} mp(m) = (1-f) \sum_{m=0}^{\infty} mp_1(m) + f \sum_{m=0}^{\infty} mp_2(m) && (6.33)\\ = (1-f)\langle m\rangle_1 + f\langle m\rangle_2. && (6.34) \end{aligned}$$

It can easily be shown that $\langle m\rangle_2 = 2\langle m\rangle_1$ and hence

$$\langle m\rangle = (1+f)\langle m\rangle_1. \tag{6.35}$$

Similarly for $\langle m^2\rangle$, we have:

$$\langle m^2\rangle = (1-f)\langle m^2\rangle_1 + f\langle m^2\rangle_2. \tag{6.36}$$

It can be shown that $\langle m^2\rangle_2 = 2\langle m^2\rangle_1 + 2\langle m\rangle_1^2$ (this follows from the fact that the variance of a sum of independent random variables is equal to the sum of the variances). Thus we obtain

$$\langle m^2\rangle = (1-f)\langle m^2\rangle_1 + f\left[2\langle m^2\rangle_1 + 2\langle m\rangle_1^2\right]. \tag{6.37}$$

Putting these expressions together, we find that

$$\begin{aligned}
\mathrm{var}(m) &= \langle m^2\rangle - \langle m\rangle^2 && (6.38)\\
&= (1+f)\langle m^2\rangle_1 + 2f\langle m\rangle_1^2 - (1+f)^2\langle m\rangle_1^2 && (6.39)\\
&= (1+f)\langle m^2\rangle_1 - (1+f^2)\langle m\rangle_1^2 && (6.40)
\end{aligned}$$

The Fano factor is then

$$\begin{aligned}
F &= \mathrm{var}(m)/\langle m\rangle\\
&= \frac{(1+f)\langle m^2\rangle_1 - (1+f^2)\langle m\rangle_1^2}{(1+f)\langle m\rangle_1}\\
&= \frac{\langle m^2\rangle_1}{\langle m\rangle_1} - \frac{(1+f)\langle m\rangle_1^2 - f(1-f)\langle m\rangle_1^2}{(1+f)\langle m\rangle_1}\\
F &= \underbrace{\frac{\langle m^2\rangle_1 - \langle m\rangle_1^2}{\langle m\rangle_1}}_{\text{Transcription}} + \underbrace{\frac{f(1-f)}{1+f}\langle m\rangle_1}_{\text{Gene copy number}}, && (6.41)
\end{aligned}$$

reproducing equation 2 from the main text. The two terms of this expression each have straightforward interpretations. The first term is simply the (architecture-dependent) Fano factor of a single copy of a gene. The second term is the contribution from copy number variation. We can make two observations. First, the contributions to overall noise from promoter architecture and from gene copy number change are independent and additive. This is unsurprising since the two processes are (by assumption) independent and uncorrelated. Second, the contribution due to gene copy number increases linearly with expression. The predicted contribution to the Fano factor from copy number variation, the second term in equation 6.41, is shown in Figure 6.10 as a function of the average gene copy number ($= 2-f$). As expected, if the copy number has a defined, static value ($f=0$ or $f=1$) there is no contribution to the Fano factor from variation in copy number. However, between these two minima, the variance reaches a maximum at $f = 0.5$, when the cell spends equal time with one or two copies, and thus the contribution to Fano factor has a maximum shifted towards slightly lower means (which appears in the denominator of Fano factor). In a section to follow and in Figure 6.14, we show that as predicted, if the cells are binned into a population expected (based on physical size) to have only one or only two copies of the measured gene, the Fano factor deceases by roughly the same magnitude as that expected from copy number variations.

6.5.2.3 Extrinsic noise due to repressor copy number fluctuations

In addition to the "intrinsic" variability characterized by our modeling efforts, extrinsic sources of variability including (e.g.) changes in TF copy number can also contribute to overall variability in gene expression. To investigate the potential effects of fluctuations in repressor copy number, we performed numerical studies in which the repressor copy number was allowed to vary across a population of cells. Let $P_{\text{arc}}(m|R=k)$ denote the promoter architecture-dependent probability that a cell contains m mRNA given k copies of a repressor TF. Let $P_{TF}(R=k)$ denote the probability that a cell contains k repressor copies. (Our analysis of "intrinsic" cell-to-cell variability implicitly assumes that all cells have the same repressor copy number - that is, $P_{TF}(R=k)=\delta_{kk'}$ for some k'.) Then the overall probability of observing m mRNA in a cell is given by

$$P(m)=\sum_{k=0}^{\infty} P_{\text{arc}}(m|R=k)\cdot P_{TF}(R=k). \tag{6.42}$$

The quantity $P_{\text{arc}}(m|R=k)$ can be computed numerically as described in [19], and thus we can compute $P(m)$ numerically for any repressor copy number distribution $P_{TF}(R=k)$.

Here, we analyzed a population of cells in which the repressor copy numbers of individual cells are distributed according to a negative binomial distribution. The negative binomial distribution

$$P_{TF}(k;n,p)=\binom{n+k-1}{n-1}p^n(1-p)^k \tag{6.43}$$

gives the probability that the nth success occurs on the $(k+n)$th trial, where the probability of success on any single trial is p. It is often used to model a more dispersed or long-tailed distribution than the Poisson distribution, and has been shown to correspond to constitutive mRNA production with a geometrically distributed number of proteins translated from each mRNA [18]. The degree of dispersal can be tuned via the parameter n as shown in Figure 6.11A. For a range of different n values, we tuned mean repressor expression via the parameter p while holding n constant. The resulting Fano vs. mean curves for the target gene are shown in Figure 6.11B. We observe that, despite substantial variability in repressor copy number, the overall variability is predominantly contributed by intrinsic sources. This conclusion is robust across both relatively narrow and relatively dispersed repressor copy number distributions.

6.5.2.4 Extrinsic noise due to RNAP copy number fluctuations

In addition to repressor copy number fluctuations, RNAP copy number fluctuations also have the potential to contribute to the overall observed variability. We will follow an approach similar to the one outlined in the previous section. The distribution of RNAP copy numbers will be estimated from sources in the literature. The average RNAP copy number is reported as $\approx 10,000$ per cell [47]. In [9], the authors report that for a typical protein with 10,000 copies per cell, the standard deviation in protein copy number is approximately 3200. We will thus model the RNAP copy number distribution as a negative binomial distribution with mean equal to 10,000 and standard deviation equal to 3200, show in Figure 6.12A. We assume that the transcription rate r is proportional to the RNAP copy number in the cell. The resulting Fano vs. mean curves are plotted in Figure 6.12B for both the constitutive expression and simple repression architectures. In both cases, we see that extrinsic variability due to RNAP fluctuations increases with increasing mean expression. In the case of constitutive expression, this increasing trend in the Fano vs. mean curve is markedly similar to the increasing trend we observe in our constitutive expression data. In the case of simple repression, the addition of extrinsic variability does not change the overall qualitative features of the predicted curve. However, it does lead to the prediction that the overall observed Fano factor will not fall all the way back down to one in the absence of repressor, which is consistent with our experimental observations.

In the case of constitutive expression, it is possible to derive an informative analytical expression for the extrinsic noise contributed by variation in r. While this is a general approach to variations in r, later we will relate this to the specific circumstance of RNAP fluctuations expected in our constitutive expression measurements.

Let the probability distribution for values of r be denoted by $P_{\text{ext}}(r)$. As in the main text (equation 6.1), the steady-state probability distribution for mRNA copy number, m, given a particular value of r is a Poisson distribution with mean r/γ, such that

$$P_{\text{arc}}(m|r) = \frac{(r/\gamma)^m}{m!} e^{-r/\gamma}. \tag{6.44}$$

We assume that r changes on a timescale sufficiently long compared to $1/\gamma$ that we can use the steady-state probability distribution. Then the overall probability distribution for mRNA copy number, integrated over all possible values of r, is given by

$$P(m) = \int P_{\text{arc}}(m|r) P_{\text{ext}}(r) dr. \tag{6.45}$$

To compute the Fano factor for this overall distribution, we will as usual proceed by computing $\langle m \rangle$

and $\langle m^2 \rangle$.

$$\langle m \rangle = \sum_{m=0}^{\infty} m \int P_{\text{arc}}(m|r) P_{\text{ext}}(r) dr, \tag{6.46}$$

$$\langle m \rangle = \int P_{\text{ext}}(r) \sum_{m=0}^{\infty} m P_{\text{arc}}(m|r) dr, \tag{6.47}$$

$$\langle m \rangle = \int P_{\text{ext}}(r) \frac{r}{\gamma} dr, \tag{6.48}$$

$$\langle m \rangle = \frac{\langle r \rangle}{\gamma}, \tag{6.49}$$

where $\langle r \rangle$ is the mean value of r. Similarly, for $\langle m^2 \rangle$,

$$\langle m^2 \rangle = \sum_{m=0}^{\infty} m^2 \int P_{\text{arc}}(m|r) P_{\text{ext}}(r) dr, \tag{6.50}$$

$$\langle m^2 \rangle = \int P_{\text{ext}}(r) \sum_{m=0}^{\infty} m^2 P_{\text{arc}}(m|r) dr, \tag{6.51}$$

$$\langle m^2 \rangle = \int P_{\text{ext}}(r) \left(\frac{r^2}{\gamma^2} + \frac{r}{\gamma} \right) dr, \tag{6.52}$$

$$\langle m^2 \rangle = \frac{\langle r \rangle}{\gamma} + \frac{\langle r^2 \rangle}{\gamma^2}. \tag{6.53}$$

Hence, the Fano factor is given by

$$\text{Fano} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle}, \tag{6.54}$$

$$\text{Fano} = \frac{\frac{\langle r \rangle}{\gamma} + \frac{\langle r^2 \rangle}{\gamma^2} - \frac{\langle r \rangle^2}{\gamma^2}}{\frac{\langle r \rangle}{\gamma}}, \tag{6.55}$$

$$\text{Fano} = 1 + \frac{1}{\gamma} \frac{\langle r^2 \rangle - \langle r \rangle^2}{\langle r \rangle}, \tag{6.56}$$

$$\text{Fano} = 1 + \frac{1}{\gamma} \text{Fano}(r). \tag{6.57}$$

Thus far, we have assumed nothing about the specific mechanism causing fluctuations in r. Let's now explore the case in which fluctuations in r are caused by fluctuations in RNAP copy number. We will assume that the transcription rate r is proportional to the RNAP copy number, so that $r = r_0 E$ where E is the RNAP polymerase copy number and r_0 is a constant of proportionality that can be thought of as roughly the transcription rate per RNAP molecule. The constant of proportionality r_0 is assumed to depend on the strength of the promoter, so that when we tune mean expression by tuning promoter strength, it is really (by assumption) the parameter r_0 that we are changing.

Under this assumption, equation 6.57 becomes:

$$\text{Fano} = 1 + \frac{1}{\gamma}\,\text{Fano}(r_0 E), \tag{6.58}$$

$$\text{Fano} = 1 + \frac{r_0}{\gamma}\,\text{Fano}(E), \tag{6.59}$$

$$\text{Fano} = 1 + \frac{\langle m\rangle}{\langle E\rangle}\,\text{Fano}(E), \tag{6.60}$$

$$\text{Fano} = 1 + \langle m\rangle\frac{\sigma_E^2}{\langle E\rangle^2}, \tag{6.61}$$

$$\text{Fano} = 1 + \frac{1}{10}\langle m\rangle, \tag{6.62}$$

where we have used the fact that by assumption $\langle m\rangle = \langle r\rangle/\gamma = r_0\langle E\rangle/\gamma$, and used the result of [9] that the squared coefficient of variation $\sigma_E^2/\langle E\rangle^2 \approx 10^{-1}$ for a protein with $\approx 10^4$ copies per cell. Equation 6.62 tells us simply that fluctuations in RNAP copy number contribute an additional term to the Fano factor that increases linearly with mean gene expression, with a slope equal to the squared coefficient of variation of the RNAP copy number.

However, it is worth noting that RNAP copy number fluctuations are by no means the only extrinsic mechanism capable of generating this linear relationship between Fano factor and mean expression. For instance, one could imagine that DNA supercoiling renders the promoter inaccessible and thus silences transcription some fraction s of the time. We could model this scenario by saying that the effective RNAP copy number is zero for a fraction s of the time, and E_0 for the remainder (*i.e.*, $1 - s$) of the time, where E_0 is some number of order 10^4. As before, we assume that the transcription rate is proportional to the RNAP copy number. We can thus proceed from equation 6.59. It can easily be shown that $\text{Fano}(E) = sE_0$ in this case, and hence equation 6.59 becomes

$$\text{Fano} = 1 + \frac{r_0}{\gamma}sE_0. \tag{6.63}$$

If we use the fact that $\langle m\rangle = r_0\langle E\rangle/\gamma = r_0(1-s)E_0/\gamma$, we obtain finally

$$\text{Fano} = 1 + \frac{s}{1-s}\langle m\rangle. \tag{6.64}$$

So again, we have an extrinsic noise term that increases linearly with mean expression, here with a slope that depends on the fraction of time s for which expression is silenced. The implications of this result will be discussed in the following section.

6.5.2.5 Extrinsic sources of noise: Concluding remarks

To conclude this exploration of sources of extrinsic noise, we will offer a few observations concerning model selection and the interpretation of experimental evidence. In a 2011 paper, Huh and Paulsson pointed out that protein partitioning at cell division yields the same $1/\langle x \rangle$ overall scaling in the cell-to-cell variability in protein levels as does Poissonian transcription [48]. Thus, experimental observation of $1/\langle x \rangle$ noise scaling does not in itself provide a basis for distinguishing between these two mechanisms. Although this specific point is not relevant in the case of our experiments, since mRNA lifetimes are sufficiently short compared with division times that partitioning effects are negligible, the overall spirit of Huh and Paulsson's argument is relevant.

In particular, we found that the effect of gene copy number variation on the Fano factor increases linearly with mean gene expression. However, in the case of constitutive expression, we also find that the effect of RNAP fluctuations on the Fano factor increases linearly with mean expression. Furthermore, the same would be true (in the case of constitutive expression) if one postulated that a mechanism such as DNA supercoiling causes transcriptional silencing *e.g.* 25% of the time: one would find a linear relationship between Fano factor and mean expression. So how can we have any confidence in the breakdown of noise sources in Figure 6.2? In our view, this discussion highlights the importance of independent corroboration of each of the pieces shown in Figure 6.2. The quantization error is corroborated through both theoretical calculation and analysis of simulated data (above). The gene copy number variation effect is corroborated below by using cell size as a proxy for gene copy number (older, larger cells will have two chromosome copies, while younger, smaller cells will have one). The RNAP fluctuation effect is the most speculative. Although we believe it is defensible both in terms of the underlying assumption that expression is proportional to RNAP copy number [49], and in the magnitude of RNAP fluctuations taken from literature sources [9, 50], our data does not provide us with a means to independently corroborate this effect. (To do so, one might perform an experiment in which fluorescently tagged RNAP molecules are used to quantify RNAP fluctuations.) Thus, it is possible that the RNAP fluctuation effect is instead something else entirely, such as transcriptional silencing by DNA supercoiling. This is the reasoning behind our statement in the Discussion of the main text that "our data does not completely rule out these [alternative] hypotheses."

6.5.3 Error bars in Fano factor measurements

In the main text Figures 6.2 and 6.4 as well as SI Figures 6.13 and 6.14B contain experimental measurements for the Fano factor. Typically there are at least three repeats of any given condition and each individual data point represents an individual experiment; all available data points are plotted on every figure. Error bars are determined by bootstrapping the single cell copy number distribution 1000 times and calculating the standard deviation in the Fano factor for those 1000 independent bootstrapped data sets. In other words, the single cell mRNA copy number distribution, which typically contains roughly 900 entries of the number of mRNA in a given cell, is randomly resampled with replacement (the same cell may appear multiple times in a given bootstrapped data set) to create a new data set with the equivalent number of entries. This is repeated 1000 times and the standard deviation of the Fano factor in these measurements is used as an error bar for the measurement.

6.5.4 Copy number variation: Uncorrected figures

In the main text, our focus was on the promoter architecture-dependent component of gene expression variability, and thus we subtracted the gene copy number-dependent term (as defined by the second term in equation 6.41) from the measured Fano factor in Figures 6.2 and 6.4 of the main text. However, doing so does not change the qualitative conclusion that variability is promoter architecture dependent. In Figure 6.13, we plot the data from Figures 6.2 and 6.4 without subtracting the gene copy number-dependent term.

6.5.5 Testing gene copy number noise by cell size segregation

In the main text we claim that one significant contribution to the measured cell-to-cell mRNA copy number variability stems from the fact that our measurements contain a mix of cells with one or two copies of the gene of interest. To test this claim, we take the data from each measurement of our constitutive strains and divide the data into two subsets based on their physical size. The idea is to use our knowledge of the cell cycle, based on growth rate and gene position in the chromosome, to divide each data set into one set with cells likely to have a single copy of the target gene (referred to as "small cells") and cells likely to have two copies of the target gene (referred to as "large cells"). As mentioned in the main text, at 60 minute growth rate we expect that the galK locus has a copy number of 1.66 [36], which implies that we should set our division line at roughly 1/3 of the way through the cell cycle. We determine this point by plotting the cumulative probability distribution of the cell area of every cell in every sample and identifying the area value where 1/3 of the cells are smaller and 2/3 are larger. For our cells this is at an area of roughly 3.75 μm^2. To help ensure that this division is "clean" we discard cells 1/8 above and below the division line so that our small cells contain the smallest 21% of cells and the large cells are the 54% largest cells.

In Figure 6.14A, we show the result of plotting the mean mRNA copy number of the small cells bin versus the mean mRNA copy number of the large cells bin for every measurement of every constitutive strain (black points). As stated previously while deriving gene copy number noise, we expect that the large cells, binned to have two copies of the reporter gene, should have twice the transcriptional activity of the small cells. This is precisely what is observed, the red line is a line of slope two and intercept zero. While we do not expect this method to achieve a perfect division of the total population, this test indicates that these subsets of data contain primarily cells with one and two copies of the reporter gene.

In Figure 6.14B, we show the Fano factor of each of these two data subsets (black squares for small cells, black circles for large cells) as well as the Fano factor of the full data sets as red diamonds. Once again, the relevant sources of intrinsic and quantization noise are shaded in green (quantization error), red (RNAP fluctuation error) and blue (gene copy number noise). First, when the mean is small (< 1 per cell per gene copy) the expected contribution from copy number variations is small (the blue shaded region is small) and thus the two subsets and the full data set give similar results for the Fano factor as expected. Above this threshold we begin to see that the Fano factor of either subset of data (both large cells and small cells) falls below the corresponding Fano factor of the full data set. Furthermore, we see that the reduction in Fano factor causes the subsets to fall approximately on the interface between the shaded blue and red region; the subset data is now consistent with a Poisson process with quantization error and RNAP fluctuations but without gene copy number fluctuations.

6.5.6 Determination of rate parameter values

The values of $k_{\mathrm{R}}^{\mathrm{off}}$ used in this work are taken from [51] and [19], and are shown in Table 6.2. More specifically, reference [51] used a single molecule *in vitro* assay to measure the dissociation rate from the LacI Oid operator. Reference [19] used this rate, along with knowledge of the dissociation constants of the Oid, O1, O2, and O3 operators (reported in [26]), to estimate the dissociation rates for the three additional operators, using the assumption that the ratio of the dissociation rates for two particular operators is equal to the ratio of their dissociation constants. In order to determine the three different values of $k_{\mathrm{R}}^{\mathrm{on}}$ used in Figure 6.4B of the main text, slightly more work was required. Recall that we are assuming a diffusion-limited on rate for which $k_{\mathrm{R}}^{\mathrm{on}} = k_0[R]$. Reference [2] reports that $k_0 = 2.7 \times 10^{-3} (\mathrm{s\ nM})^{-1}$. To determine $k_{\mathrm{R}}^{\mathrm{on}}$ for each of the three aTc concentrations, we must determine the repressor concentration $[R]$ at each aTc concentration. Unfortunately we do not independently possess the exact input-output relation between aTc concentration and repressor copy number, but we can estimate the repressor concentration at each aTc concentration by looking at how strongly gene expression is repressed at each aTc concentration.

More specifically, in a recent work [42], the authors defined the “repression” as the ratio between gene expression in the absence of repressor transcription factors (TFs) and gene expression in the presence of repressor TFs. They showed that, for the type of “simple repression” promoter architecture used in this paper, the repression is given by

$$\text{Repression} = 1 + \frac{2R}{N_{\mathrm{NS}}}\, e^{-\Delta\epsilon_{\mathrm{rd}}/k_{\mathrm{B}}T}, \tag{6.65}$$

where R is the repressor copy number, N_{NS} is the number of non-specific binding sites (taken to be equal to the size of the *E. coli* genome, or $2 \times 5 \times 10^6$), and $\Delta\epsilon_{rd}$ is the repressor-DNA binding energy ($-17.3\,k_BT$) for the Oid LacI binding site. This expression can be solved to determine the repressor copy number R as a function of the repression

$$R = (\text{Repression} - 1) \times \frac{N_{NS}}{2}\, e^{\Delta\epsilon_{rd}/k_BT}. \tag{6.66}$$

Garcia and Phillips [42] used this equation to determine the effective repressor copy number R, and verified their results using quantitative Western blot analysis. We used a similar approach by computing the repression at each of the aTc concentrations for the Oid operator construct, then using equation 6.66 coupled with the fact that the volume of an *E. coli* cell is approximately 1 fL to determine the repressor concentration at each aTc concentration. Finally, we multiplied these concentrations by k_0 to determine the appropriate value of $k_{\mathrm{R}}^{\mathrm{on}}$. The results are summarized in Table 6.3.

6.5.7 Supplemental Figures

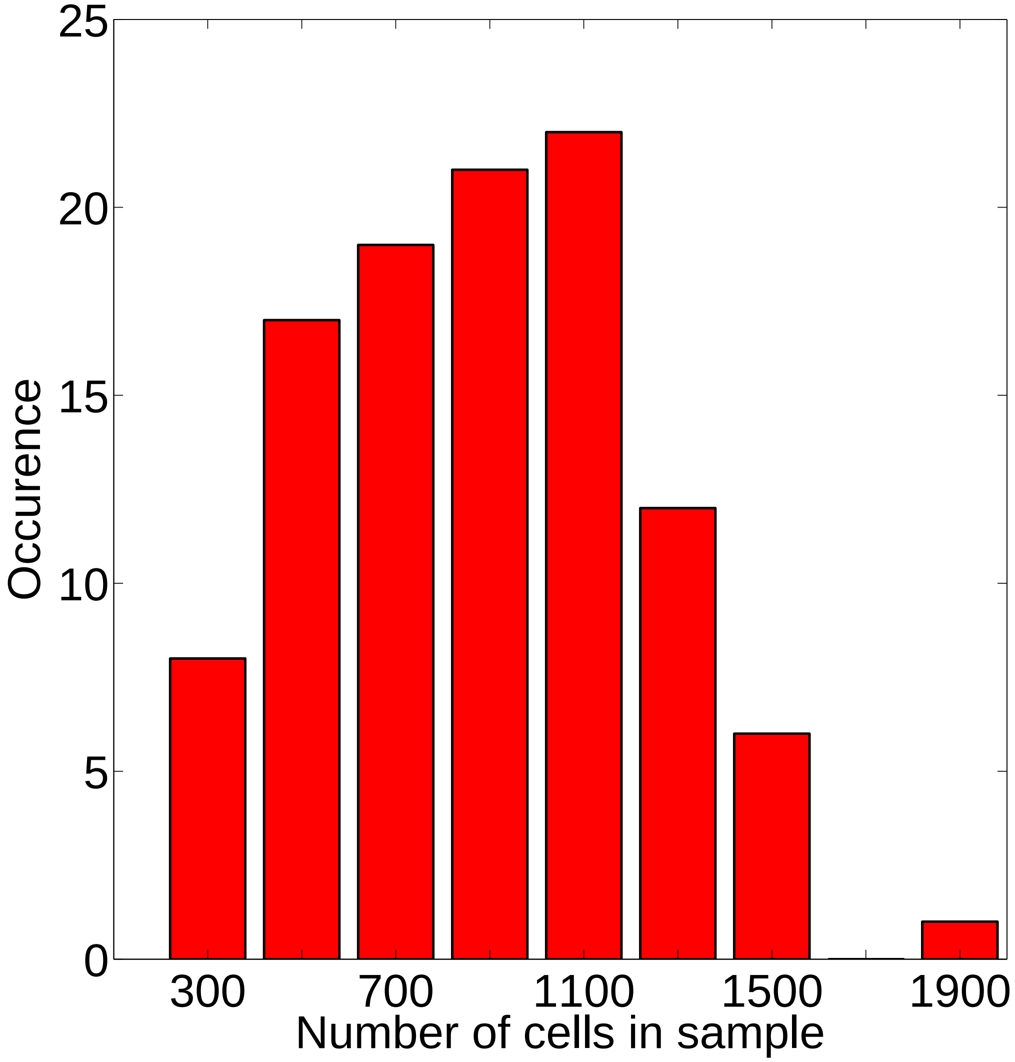


Figure 6.5: **Histogram of the number of cells per FISH sample.**
Each sample has 100 unique positions imaged. Due to differences in cell density and position quality (positions are chosen in an automated process), samples range in size and have roughly 900 cells on average per sample.

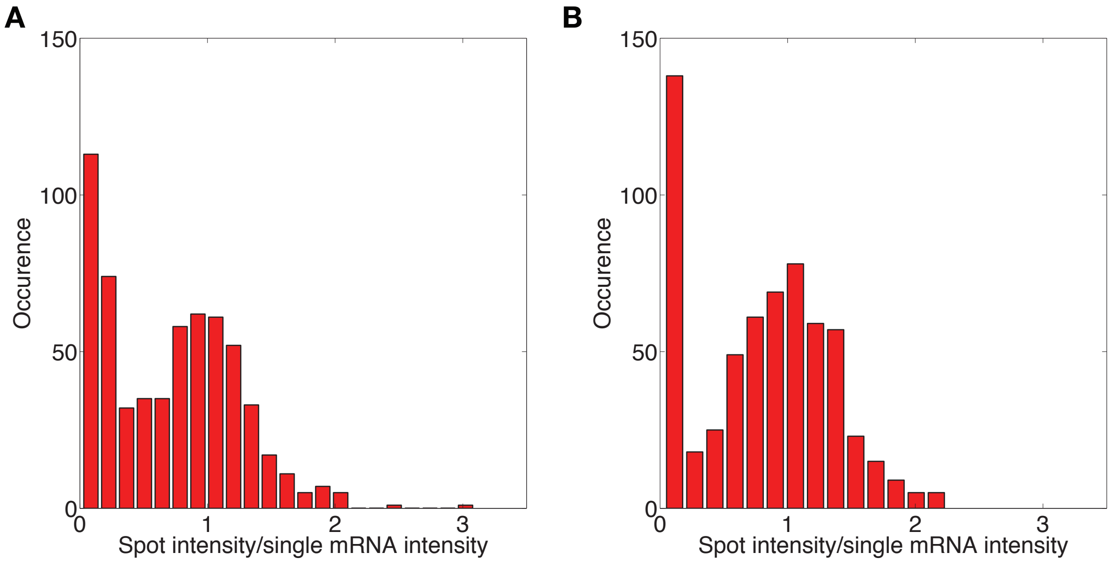


Figure 6.6: **Histograms of detected spot intensities for low expression FISH data.** Detected spots (local maxima in fluorescence signal), in principle, correspond to either zero or one mRNA. Part (**A**) shows a representative histogram from FISH experiments, while part (**B**) shows a histogram from simulated data. The signal of each spot has been normalized by the "single mRNA intensity". For both histograms, we identify a "noise" or background peak at fluorescence intensity ≈ 0. This peak corresponds to unbound or nonspecifically bound probes. In both cases, although we can discern distinct peaks, distinguishing between zero and one mRNA is not completely unambiguous.

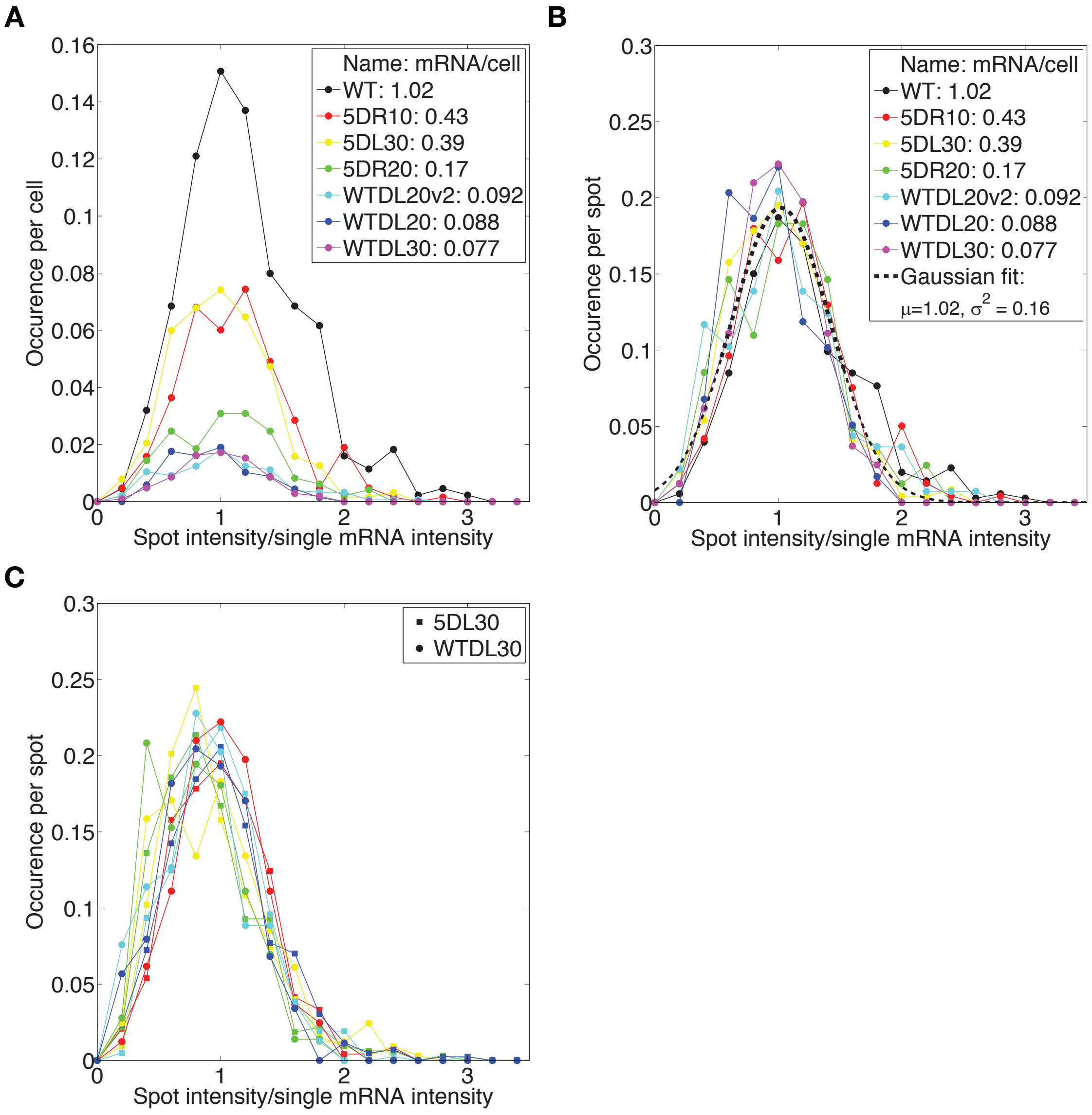


Figure 6.7: **Comparison of identified single spot intensities across different samples and experimental acquisitions.**

(**A**) Histograms of single spot intensities for seven samples, normalized by the number of cells in each sample. The value on the *y*-axis corresponds to the probability of finding a spot with a given intensity in any one cell in the sample. Mean expression in the samples ranges between 0.077 mRNA per cell and 1.02 mRNA per cell. When expression is low, increases in the mean expression level increase the probability of finding a spot with intensity equal to a single mRNA, rather than, for instance, increasing the intensity of identified spots. (**B**) Histograms of single spot intensity values for the same seven samples, normalized by the total number of identified spots in each sample. In this case, the value on the *y*-axis corresponds to the probability that a given identified spot has a particular intensity. The spots have roughly the same properties in each of these samples, although in the highest expression samples, we begin to see increased probability of having spots with intensity corresponding to more than 1 mRNA. The day-to-day reproducibility in this identification process is shown in part (**C**) where two different strains (5DL30 and WTDL30) are shown measured across five different acquisitions.

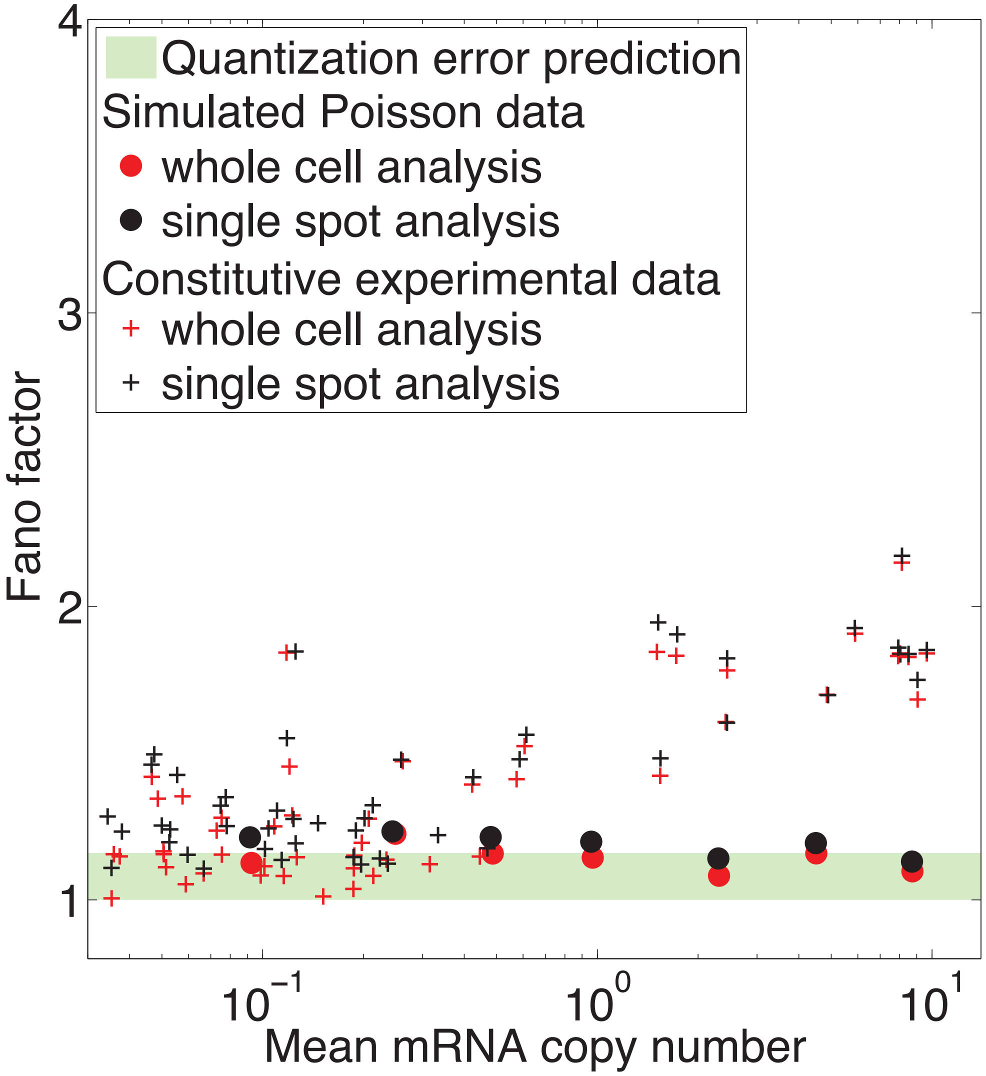


Figure 6.8: **Fano factor vs mean plot for simulated Poisson distributed data.** Simulated mRNA FISH data with Poisson distributed mRNA copy numbers (circles) is analyzed over a range of mean mRNA levels to evaluate our analysis code. Since the simulated data is Poisson distributed, the true value of the Fano factor is one. However, we see here that the measured Fano factor is always slightly greater than one. This persistent noise represents the "quantization error" discussed in the text; the expected contribution to the Fano factor from quantization error is indicated by the height of the green bar. For comparison, the crosses are the Fano factors (corrected for gene copy number noise) from our constitutive expression strains (data from Figure 6.2A). The different colors (black and red) represent two distinct methods for quantifying the resulting mRNA signal. For the black symbols, individual mRNA spots are identified and quantified (divided by the intensity of a single mRNA) and rounded to the nearest whole number of mRNA and the copy number in a cell is the sum of the number of mRNA in each identified spot for a given cell. The red symbols correspond to summing the signal of all identified spots in a cell and determining a cell's copy number by dividing the summed signal by the single mRNA intensity (and, in this case, not rounding). This second method (red symbols) is used in this work, but this choice does not significantly influence the outcome.

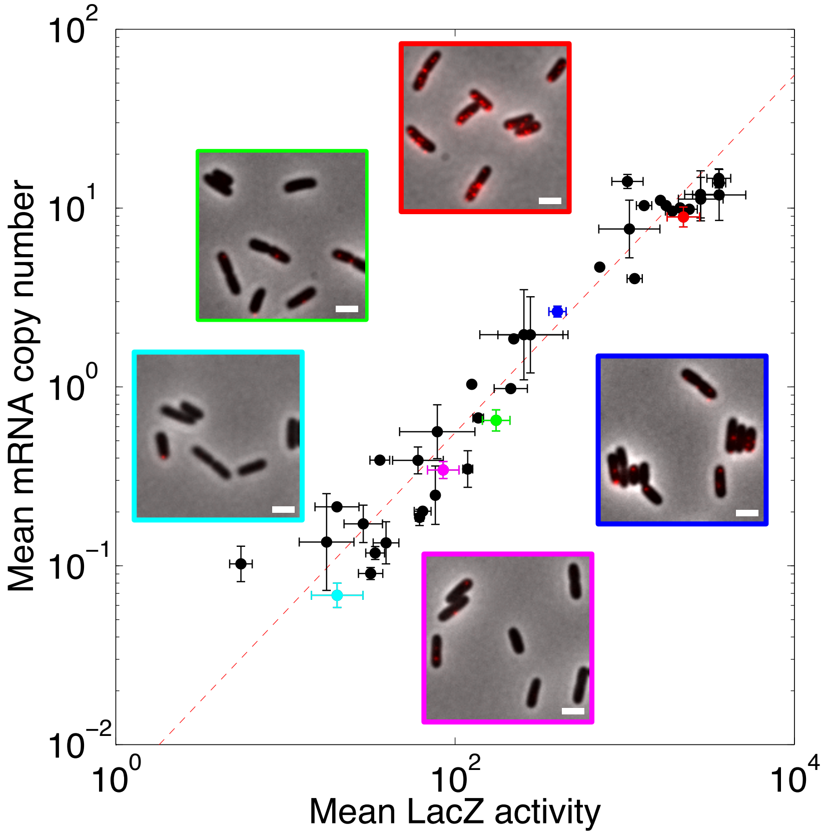


Figure 6.9: **Experimental comparison of mean mRNA FISH measurements to enzymatic assay.**

Direct comparison of the average mRNA copy number to the average enzymatic activity of the encoded protein for every data strain and condition used in the text. The red line is a linear fit to the data. Error bars are standard deviation from multiple measurements.

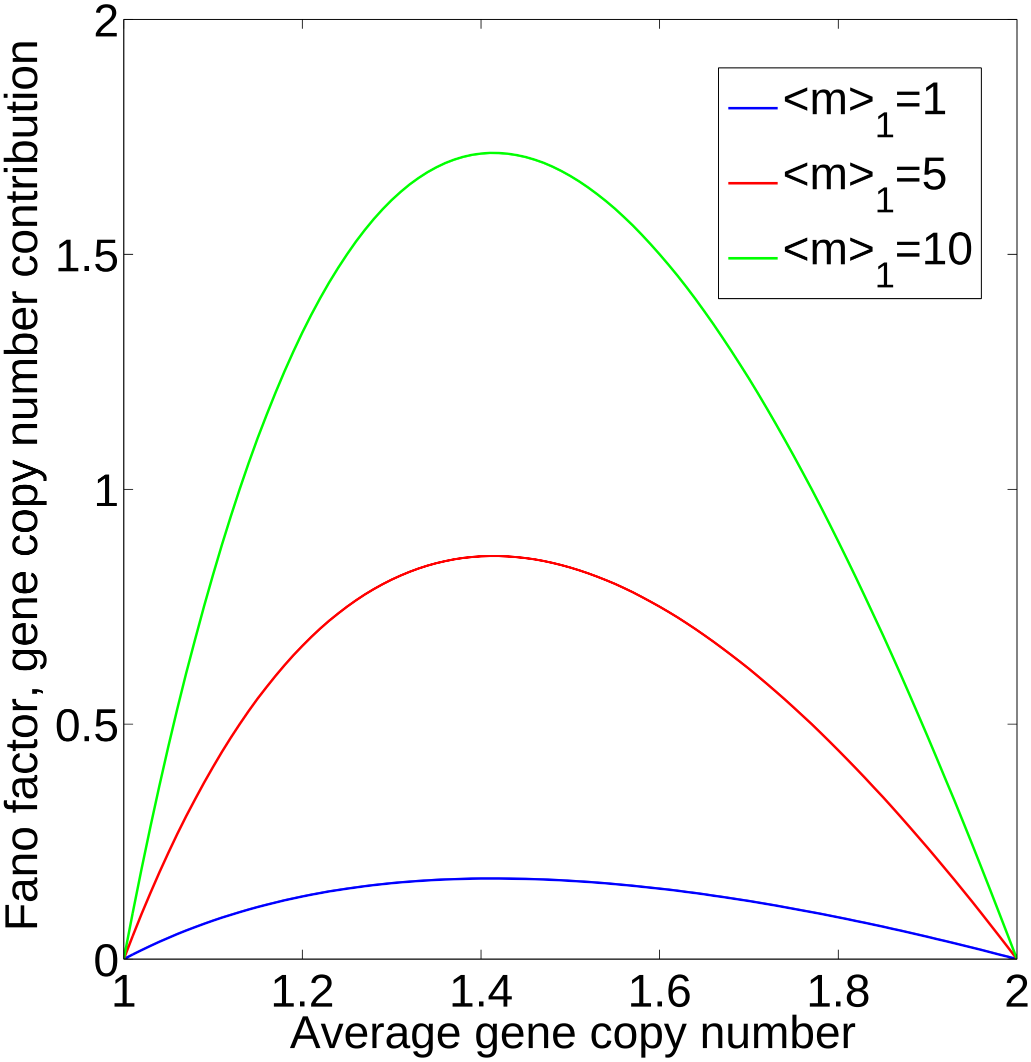


Figure 6.10: **Fano factor contribution from gene copy number variation.** Predicted contribution to the Fano factor from gene copy number variation for three distinct mean expression levels, 1 (blue curve), 5 (red curve) and 10 (green curve) mRNA copies per cell per gene copy. The effect increases with transcription rate and is largest when the gene spends approximately half the cell cycle with 1 copy and the other half with 2 copies.

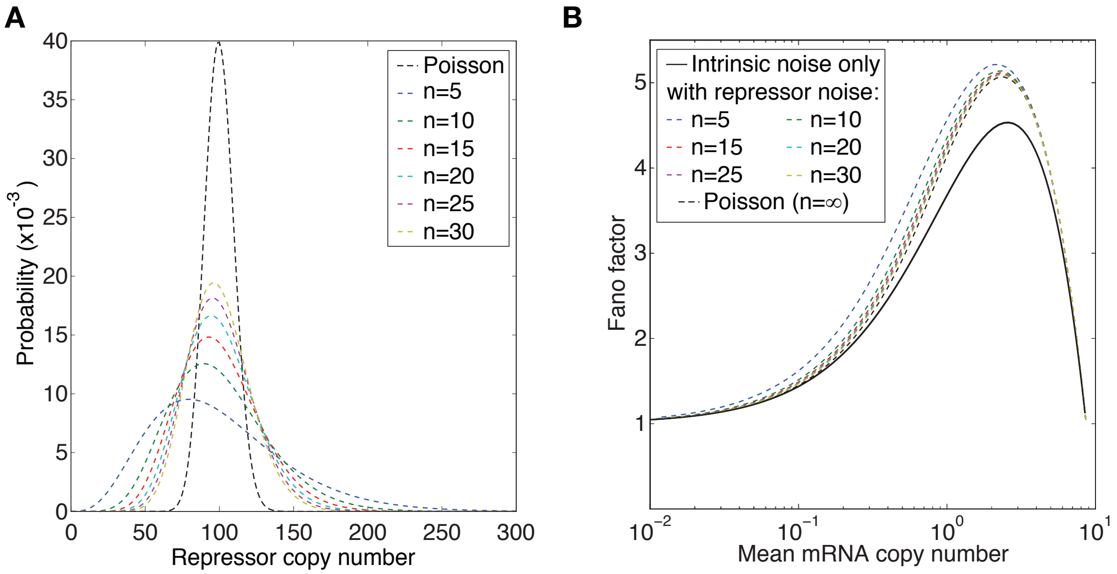


Figure 6.11: **Quantifying the extrinsic noise contribution of repressor copy number variation.**

(**A**) Single-cell repressor distribution for the negative binomial distribution with various choices for the parameter n and for a Poisson distribution. (**B**) Predicted Fano factor for simple repression with a static value for the repressor copy number without distribution (solid black line) along with the Fano factor for the distributions shown in (A) of this figure. Even when the distribution is quite wide, the added noise above the intrinsic piece is relatively small.

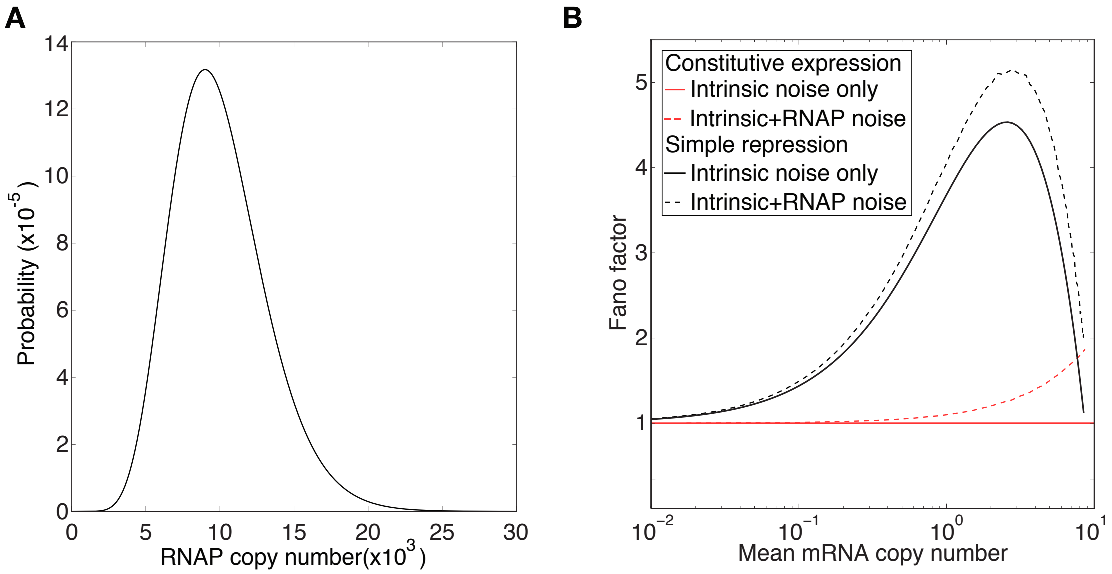


Figure 6.12: **Quantifying the extrinsic noise contribution of RNAP copy number variation.**

(**A**) Negative binomial model of RNAP copy number distribution with width chosen to coincide with reported literature values [9]. (**B**) The resulting contribution to the Fano factor from extrinsic noise in RNAP copy number. The solid lines are the theoretical predictions without any source of extrinsic noise for constitutive (red solid line) and simple repression (black solid line) and with RNAP fluctuation noise (corresponding dashed lines).

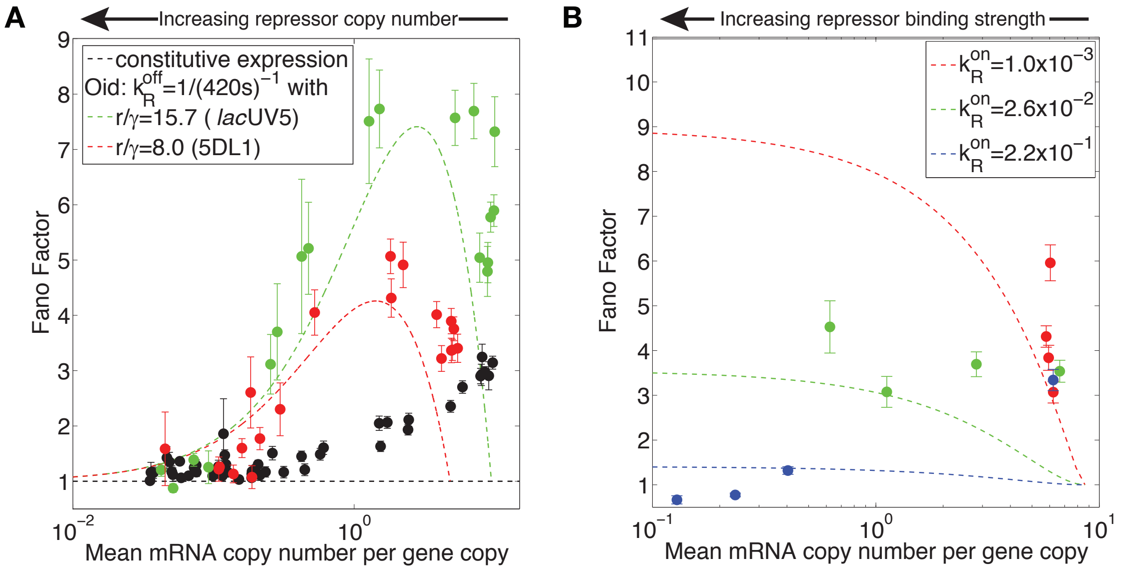


Figure 6.13: **Fano factor vs. mean mRNA copy number.** The data from Figure 6.4 of the main text are plotted without subtracting the effect of gene copy number variation. (**A**) Fano factor vs. mean mRNA copy number for two promoters (choices of r/γ): 5DL1 (red points) and *lac*UV5 (green points) while tuning k_{on}^{R} by inducing LacI to varying levels. The parameter-free predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color holding promoter (r/γ) and repressor binding strength ($k_{\text{off}}^{\text{R}}$) constant. For reference, the black data is the constitutive data from Figure 6.2. (**B**) Fano factor vs. mean mRNA copy number for *lac*UV5 while tuning $k_{\text{off}}^{\text{R}}$ by changing repressor binding site identity at fixed repressor copy number, each color is a different induction condition from red (lowest LacI induction) to blue (highest LacI induction). Again, the predictions from the kinetic theory of transcription are shown as dashed lines in the corresponding color. For both panels, not subtracting gene copy number variation slightly worsens the fit between theory and data, but the overall conclusion that variability is promoter architecture dependent is not affected. Error bars are the result of bootstrap sampling of the expression measurements in each sample

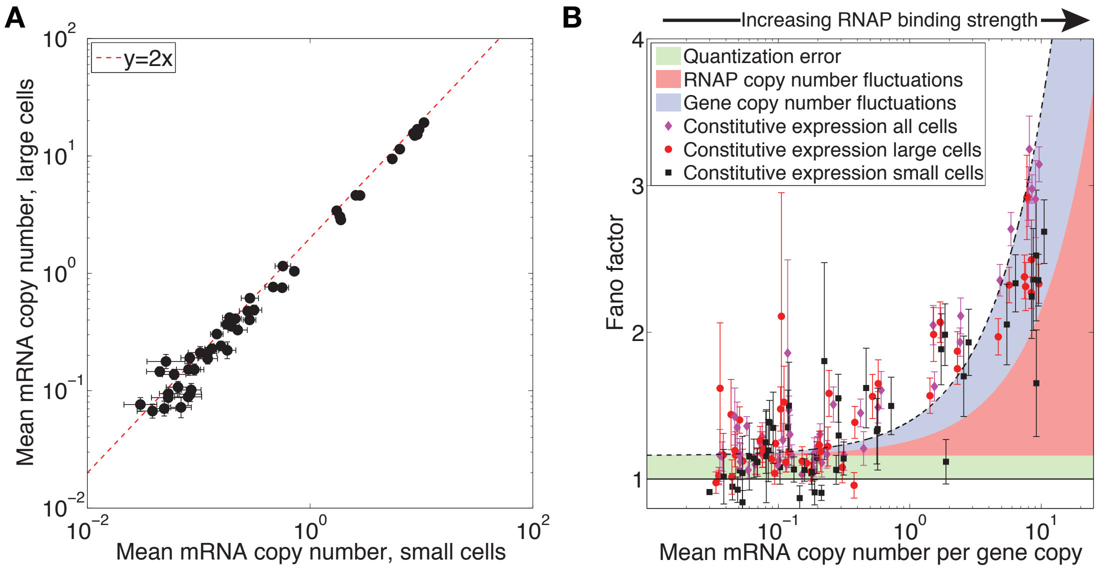


Figure 6.14: **Analysis of small cell and large cell data subsets in constitutive expression.** Each individual constitutive expression sample measurement (multiple measurements of all 18 strains) is divided into two subsets of "large" and "small" cells based on cell area. The division line between these sets is chosen such that small cells are expected to have one copy of the reporter gene while large cells are expected to contain two copies. (**A**) Mean mRNA copy number of large cells vs. mean mRNA copy number of small cells within the same sample. The mean copy number of the large cells is double the mean copy number of the small cells, supporting the assertion that the data sets are correctly divided based on gene copy number. (**B**) Fano factor vs. mean mRNA copy number for the full data sets (red points, as from Figure 6.2B), large cells (black circles) and small cells (black squares). The Fano factor for the full data samples agrees with the noise prediction including quantization noise, RNAP fluctuation noise and gene copy number noise. The subsets, divided to remove gene copy number variation in a sample, are described best without the gene copy number noise term. All error bars are the result of bootstrap sampling of the expression measurements in each sample.

6.5.8 Supplementary Tables

Name	Sequence	Name	Sequence
lacZ1	gtgaatccgtaatcatggtc	lacZ37	gatcgacagatttgatccag
lacZ2	tcacgacgttgtaaaacgac	lacZ38	aaataatatcggtggccgtg
lacZ3	attaagttgggtaacgccag	lacZ39	tttgatggaccatttcggca
lacZ4	tattacgccagctggcgaaa	lacZ40	tattcgcaaaggatcagcgg
lacZ5	attcaggctgcgcaactgtt	lacZ41	aagactgttacccatcgcgt
lacZ6	aaaccaggcaaagcgccatt	lacZ42	tgccagtatttagcgaaacc
lacZ7	agtatcggcctcaggaagat	lacZ43	aaacggggatactgacgaaa
lacZ8	aaccgtgcatctgccagttt	lacZ44	taatcagcgactgatccacc
lacZ9	taggtcacgttggtgtagat	lacZ45	gggttgccgttttcatcata
lacZ10	aatgtgagcgagtaacaacc	lacZ46	tcggcgtatcgccaaaatca
lacZ11	gtagccagctttcatcaaca	lacZ47	ttcatacagaactggcgatc
lacZ12	aataattcgcgtctggcctt	lacZ48	tggtgttttgcttccgtcag
lacZ13	agatgaaacgccgagttaac	lacZ49	acggaactggaaaaactgct
lacZ14	aattcagacggcaaacgact	lacZ50	tattcgctggtcacttcgat
lacZ15	tttctccggcgcgtaaaaat	lacZ51	gttatcgctatgacggaaca
lacZ16	atcttccagataactgccgt	lacZ52	tttaccttgtggagcgacat
lacZ17	aacgagacgtcacggaaaat	lacZ53	gttcaggcagttcaatcaac
lacZ18	gctgatttgtgtagtcggtt	lacZ54	ttgcactacgcgtactgtga
lacZ19	ttaaagcgagtggcaacatg	lacZ55	agcgtcacactgaggttttc
lacZ20	aactgttacccgtaggtagt	lacZ56	atttcgctggtggtcagatg
lacZ21	ataatttcaccgccgaaagg	lacZ57	acccagctcgatgcaaaaat
lacZ22	tttcgacgttcagacgtagt	lacZ58	cggttaaattgccaacgctt
lacZ23	atagagattcgggatttcgg	lacZ59	ctgtgaaagaaagcctgact
lacZ24	ttctgcttcaatcagcgtgc	lacZ60	ggcgtcagcagttgtttttt
lacZ25	accattttcaatccgcacct	lacZ61	tacgccaatgtcgttatcca
lacZ26	ttaacgcctcgaatcagcaa	lacZ62	taaggttttcccctgatgct
lacZ27	atgcagaggatgatgctcgt	lacZ63	atcaatccggtaggttttcc
lacZ28	tctgctcatccatgacctga	lacZ64	gtaatcgccatttgaccact
lacZ29	ttcatcagcaggatatcctg	lacZ65	agttttcttgcggccctaat
lacZ30	cacggcgttaaagttgttct	lacZ66	atgtctgacaatggcagatc
lacZ31	tggttcggataatgcgaaca	lacZ67	ataattcaattcgcgcgtcc
lacZ32	ttcatccaccacatacaggc	lacZ68	tgatgttgaactggaagtcg
lacZ33	tgccgtgggtttcaatattg	lacZ69	tcagttgctgttgactgtag

lacZ34	atcggtcagacgattcattg	lacZ70	attcagccatgtgccttctt
lacZ35	tgatcacactcgggtgatta	lacZ71	aatccccatatggaaaccgt
lacZ36	atacagcgcgtcgtgattag	lacZ72	agaccaactggtaatggtag

Table 6.1: Names and sequences of LacZ mRNA probes.

Operator	$k_{\mathrm{R}}^{\mathrm{off}}(s^{-1})$
Oid	0.0023
O1	0.0069
O2	0.091
O3	2.1

Table 6.2: **Repressor dissociation rates**. These rates are taken directly from refs. [19] and [51]. The dissociation rate of the Oid operator was directly measured *in vitro*, while the O1, O2, and O3 dissociation rates were computed using the ratios of these binding sites' equilibrium occupancies to that of Oid.

aTc concentration	R (copy number)	[R] (nM)	$k_{\mathrm{R}}^{\mathrm{on}}(s^{-1})$
0.5 ng/mL	0.21	0.35	0.0010
2 ng/mL	5.9	9.8	0.026
10 ng/mL	50	83	0.22

Table 6.3: **Repressor binding rates**. As described in the supplementary text, these rates were computed by combining the association rate per repressor reported in reference [2] with an estimate of the repressor copy number at each aTc concentration. The overall association rate is then the product of the estimated repressor copy number with the association rate per repressor molecule.

Bibliography

[1] P H von Hippel and O G Berg. Facilitated target location in biological systems. *Journal of Biological Chemistry*, 264(2):675–8, January 1989.

[2] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.

[3] H. Buc and W. R. McClure. Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the *lac* UV5 promoter. Evidence for a sequential mechanism involving three steps. *Biochemistry*, 24(11):2712–23, 1985.

[4] Achillefs N Kapanidis, Emmanuel Margeat, Sam On Ho, Ekaterine Kortkhonjia, Shimon Weiss, and Richard H Ebright. Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science*, 314(5802):1144–7, November 2006.

[5] Carlo Manzo, Chiara Zurla, David D. Dunlap, and Laura Finzi. The effect of nonspecific binding of lambda repressor on DNA looping dynamics. *Biophysical Journal*, 103(8):1753 – 1761, 2012.

[6] H. Maamar, A. Raj, and D. Dubnau. Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science*, 317(5837):526–9, 2007.

[7] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–26, 2008.

[8] D. Zenklusen, D. R. Larson, and R. H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural and Molecular Biology*, 15(12):1263–71, 2008.

[9] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533, 2010.

[10] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73, 2010.

[11] W. J. Blake, G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell*, 24(6):853–65, 2006.

[12] Gürol M Süel, Rajan P Kulkarni, Jonathan Dworkin, Jordi Garcia-Ojalvo, and Michael B Elowitz. Tunability and noise dependence in differentiation dynamics. *Science*, 315(5819):1716–9, March 2007.

[13] Mukund Thattai and Alexander van Oudenaarden. Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530, 2004.

[14] Edo Kussell and Stanislas Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.

[15] Denise M. Wolf, Vijay V. Vazirani, and Adam P. Arkin. Diversity in times of adversity: Probabilistic strategies in microbial survival games. *Journal of Theoretical Biology*, 234(2):227 – 253, 2005.

[16] Hubertus J. E. Beaumont, Jenna Gallie, Christian Kost, Gayle C. Ferguson, and Paul B. Rainey. Experimental evolution of bet hedging. *Nature*, 462(7269):90–93, Nov 2009.

[17] Mark Viney and Sarah E. Reece. Adaptive noise. *Proceedings of the Royal Society B: Biological Sciences*, 280(1767), 2013.

[18] J Paulsson and M Ehrenberg. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review Letters*, 84(23):5447–50, June 2000.

[19] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Computational Biology*, 7(3):e1001100, 2011.

[20] B. Munsky, G. Neuert, and A. van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–7, 2012.

[21] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, 2002.

[22] J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4, 2004.

[23] L. H. So, A. Ghosh, C. Zong, L. A. Sepulveda, R. Segev, and I. Golding. General properties of transcriptional time series in *Escherichia coli*. *Nature Genetics*, 43(6):554–60, 2011.

[24] Hanna Salman, Naama Brenner, Chih-kuan Tung, Noa Elyahu, Elad Stolovicki, Lindsay Moore, Albert Libchaber, and Erez Braun. Universal protein fluctuations in populations of microorganisms. *Physical Review Letters*, 108:238105, Jun 2012.

[25] Maya Dadiani, David van Dijk, Barak Segal, Yair Field, Gil Ben-Artzi, Tali Raveh-Sadka, Michal Levo, Irene Kaplow, Adina Weinberger, and Eran Segal. Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise. *Genome Research*, 23(6):966–976, 2013.

[26] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: Applications. *Current Opinion in Genetics and Development*, 15(2):125–35, 2005.

[27] T. Kuhlman, Z. Zhang, Jr. Saier, M. H., and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6043–8, 2007.

[28] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20):9158–63, May 2010.

[29] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *Journal of Molecular Biology*, 331(5):981–9, 2003.

[30] Materials and methods are available as supplementary material on *Science* Online.

[31] A. Sanchez and J. Kondev. Transcriptional control of noise in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(13):5081–6, 2008.

[32] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[33] P.S. Swain, M.B. Elowitz, and E.D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99:12795., 2002.

[34] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–72, July 2011.

[35] C. J. Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput Biol*, 9(7):e1003161, 07 2013.

[36] H. Bremer and P. P. Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In Frederick C. Neidhardt et al., editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1553–1569. ASM Press, Washington DC, 1996.

[37] Robert C. Brewster, Franz M. Weinert, Hernan G. Garcia, Dan Song, Mattias Rydenfelt, and Rob Phillips. The transcription factor titration effect dictates level of gene expression. *Cell*, 156:1–12, Mar 2014.

[38] R.C. Brewster, D.L. Jones, and R. Phillips. Tuning promoter strength through rna polymerase binding site design in *Escherichia coli*. *PLoS Computational Biology*, 8(12):e1002811, 2012.

[39] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides. RegulonDB version 7.0: Transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*, 39(Database issue):D98–105, 2011.

[40] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

[41] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf. The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598, Jun 2012.

[42] H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12174–82, 2011.

[43] Alvaro Sanchez, Sandeep Choubey, and Jane Kondev. Regulation of noise in gene expression. *Annual Review of Biophysics*, 42:469–91, January 2013.

[44] Namiko Mitarai, Ian B Dodd, Michael T Crooks, and Kim Sneppen. The generation of promoter-mediated transcriptional noise in bacteria. *PLoS Computational Biology*, 4(7):e1000109, January 2008.

[45] R. Lutz and H. Bujard. 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):1203–10, 1997.

[46] J. H. Miller. *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1972.

[47] I. L. Grigorova, N. J. Phleger, V. K. Mutalik, and C. A. Gross. Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 103(14):5332–7, 2006.

[48] Dann Huh and Johan Paulsson. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100, February 2011.

[49] S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–75, 2009.

[50] S. Bakshi, A. Siryaporn, M. Goulian, and J. C. Weisshaar. Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Mol. Microbiol.*, 85(1):21–38, Jul 2012.

[51] O. K. Wong, M. Guthold, D. A. Erie, and J. Gelles. Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biology*, 6(9):e232, 2008.