

Measuring cis-regulatory energetics in living cells using allelic manifolds

**Talitha Forcier¹, Andalus Ayaz¹, Manraj S. Gill^{1,§}, Daniel Jones^{1,2,¶}, Rob Phillips²,
Justin B. Kinney^{1,*}**

*For correspondence:
jkinney@cshl.edu (JBK)

Present address: [§]Department of Biology, Massachusetts Institute of Technology, USA; [¶]Department of Cell and Molecular Biology, Uppsala University, Sweden

Abstract Gene expression in all organisms is controlled by cooperative interactions between DNA-bound transcription factors (TFs), but quantitatively measuring TF-DNA and TF-TF interactions remains difficult. Here we introduce a strategy for precisely measuring the Gibbs free energy of such interactions in living cells. This strategy centers on the measurement and modeling of “allelic manifolds”, a multidimensional generalization of the classical genetics concept of allelic series. Allelic manifolds are measured using reporter assays performed on strategically designed cis-regulatory sequences. Quantitative biophysical models are then fit to the resulting data. We used this strategy to study regulation by two *Escherichia coli* TFs, CRP and σ^{70} RNA polymerase. Doing so, we consistently obtained energetic measurements precise to ~ 0.1 kcal/mol. We also obtained multiple results that deviate from the prior literature. Our strategy is compatible with massively parallel reporter assays in both prokaryotes and eukaryotes, and should therefore be highly scalable and broadly applicable.

Introduction

Cells regulate the expression of their genes in response to biological and environmental cues. A major mechanism of gene regulation in all organisms is the binding of transcription factor (TF) proteins to cis-regulatory elements encoded within genomic DNA. DNA-bound TFs interact with one another, either directly or indirectly, forming cis-regulatory complexes that modulate the rate at which nearby genes are transcribed (*Ptashne and Gann, 2002; Courey, 2008*). Different arrangements of TF binding sites within cis-regulatory sequences can lead to different regulatory programs, but the rules that govern *which* arrangements lead to *which* regulatory programs remain largely unknown. Understanding these rules, which are often referred to as “cis-regulatory grammar” (*Spitz and Furlong, 2012*), is a major challenge in modern biology.

Measuring the quantitative strength of interactions among DNA-bound TFs is critical for elucidating cis-regulatory grammar. In particular, knowing the Gibbs free energy of TF-DNA and TF-TF interactions is essential for building biophysical models that can quantitatively explain gene regulation in terms of simple protein-DNA and protein-protein interactions (*Shea and Ackers, 1985; Bintu et al., 2005; Sherman and Cohen, 2012*). Biophysical models have proven remarkably successful at quantitatively explaining regulation by a small number of well-studied cis-regulatory sequences. Arguably, the biggest successes have been achieved in the bacterium *Escherichia coli*, particularly in the context of the *lac* promoter (*Vilar and Leibler, 2003; Kuhlman et al., 2007; Kinney et al., 2010; Garcia and Phillips, 2011; Brewster et al., 2014*) and the O_R/O_L control region of the λ phage lysogen (*Ackers et al., 1982; Shea and Ackers, 1985; Cui et al., 2013*). But in both cases, this quantitative understanding has required decades of focused study. New approaches for dissecting cis-regulatory energetics, approaches that are both systematic and scalable, will be needed before

43 a general quantitative understanding of cis-regulatory grammar can be developed.
 44 Here we address this need by describing a systematic experimental/modeling strategy for
 45 dissecting the biophysical mechanisms of transcriptional regulation in living cells. Our strategy
 46 centers on the concept of an “allelic manifold”. Allelic manifolds generalize the classical genetics
 47 concept of allelic series to multiple dimensions. An allelic series is a set of sequence variants
 48 that affect the same phenotype (or phenotypes) but differ in their quantitative strength. Here
 49 we construct allelic manifolds by measuring, in *multiple* experimental contexts, the phenotypic
 50 strength of each variant in an allelic series. Each variant thus corresponds to a data point in a
 51 multi-dimensional “measurement space”. If the measurement space is of high enough dimension,
 52 and if one’s measurements are sufficiently precise, these data should collapse to a lower-dimension
 53 manifold that represents the inherent phenotypic dimensionality of the allelic series. These data
 54 can then be used to infer quantitative biophysical models that describe the shape of the allelic
 55 manifold, as well as the location of each allelic variant within that manifold. As we show here,
 56 such inference allows one to determine *in vivo* values for important biophysical quantities with
 57 remarkable precision.

58 We demonstrate this strategy on a regulatory paradigm in *E. coli*: activation of the σ^{70} RNA
 59 polymerase holoenzyme (RNAP) by the cAMP receptor protein (CRP, also called CAP). CRP activates
 60 transcription when bound to DNA at positions upstream of RNAP (*Busby and Ebright, 1999*), and
 61 the strength of these interactions is known to depend strongly on the precise nucleotide spacing
 62 between CRP and RNAP binding sites (*Gaston et al., 1990; Ushida and Aiba, 1990*). However, the
 63 Gibbs free energies of these interactions are still largely unknown.¹ By measuring and modeling
 64 allelic manifolds, we systematically determined the *in vivo* Gibbs free energy (ΔG) of CRP-RNAP
 65 interactions that occur at a variety of different binding site spacings. These ΔG values were
 66 consistently measured to an estimated precision of ~ 0.1 kcal/mol. We also obtained ΔG values for
 67 *in vivo* CRP-DNA and RNAP-DNA interactions, again with similar estimated precision.

68 The Results section that follows is organized into three Parts, each of which describes a different
 69 use for allelic manifolds. Part 1 focuses on measuring TF-DNA interactions, Part 2 focuses on TF-TF
 70 interactions, and Part 3 shows how to distinguish different possible mechanisms of transcriptional
 71 activation. Each Part consists of three subsections: Strategy, Demonstration, and Aside. Strategy
 72 covers the theoretical basis for the proposed use of allelic manifolds. Demonstration describes how
 73 we applied this strategy to better understand regulation by CRP and RNAP. Aside describes related
 74 findings that are interesting but somewhat tangential.

75 Results

76 Part 1. Strategy: Measuring TF-DNA interactions

77 We begin by showing how allelic manifolds can be used to measure the *in vivo* strength of TF binding
 78 to a specific DNA binding site. This measurement is accomplished by using the TF of interest as a
 79 transcriptional repressor. We place the TF binding site directly downstream of the RNAP binding
 80 site in a bacterial promoter so that the TF, when bound to DNA, sterically occludes the binding
 81 of RNAP. We then measure the rate of transcription from a few dozen variant RNAP binding sites.
 82 Transcription from each variant site is assayed in both the presence and in the absence of the TF.

83 Figure 1A illustrates a thermodynamic model (*Shea and Ackers, 1985; Bintu et al., 2005; Sherman and Cohen, 2012*) for this type of simple repression. In this model, promoter DNA can be in
 84 one of three states: unbound, bound by the TF, or bound by RNAP. Each of these three states is
 85 assumed to occur with a frequency that is consistent with thermal equilibrium, i.e., with a probability
 86 proportional to its Boltzmann weight.

87 The energetics of protein-DNA binding determine the Boltzmann weight for each state. By
 88 convention we set the weight of the unbound state equal to 1. The weight of the TF-bound state is

¹To our knowledge, only the CRP-RNAP interaction at the *lac* promoter has previously been quantitatively measured (*Kuhlman et al., 2007; Kinney et al., 2010*).

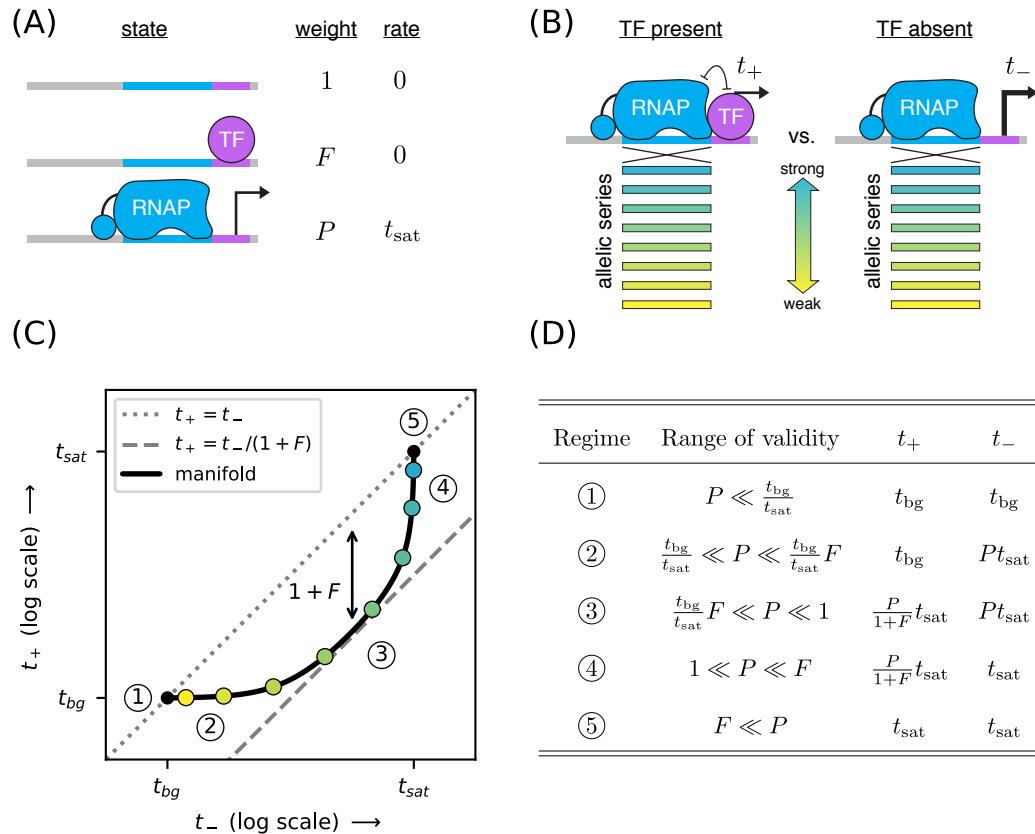


Figure 1. Strategy for measuring TF-DNA interactions. (A) A thermodynamic model of simple repression. Here, promoter DNA can transition between three possible states: unbound, bound by a TF, or bound by RNAP. Each state has an associated Boltzmann weight and rate of transcript initiation. F is the TF binding factor and P is the RNAP binding factor; see text for a description of how these dimensionless binding factors relate to binding affinity and binding energy. t_{sat} is the rate of specific transcript initiation from a promoter fully occupied by RNAP. (B) Transcription is measured in the presence (t_+) and absence (t_-) of the TF. Measurements are made for an allelic series of RNAP binding sites that differ in their binding strengths (blue-yellow gradient). (C) If the model in panel A is correct, plotting t_+ vs. t_- for the promoters in panel B (colored dots) will trace out a 1D allelic manifold. Mathematically, this manifold reflects Equation 1 and Equation 2 computed over all possible values of the RNAP binding factor P while the other parameters (F, t_{sat}) are held fixed. Note that these equations include a background transcription term t_{bg} ; it is assumed throughout that $t_{\text{bg}} \ll t_{\text{sat}}$ and that t_{bg} is independent of RNAP binding site sequence. The resulting manifold exhibits five distinct regimes (circled numbers), corresponding to different ranges for the value of P that allow the mathematical expressions in Equations 1 and 2 to be approximated by simplified expressions. In regime 3, for instance, $t_+ \approx t_-/(1+F)$, and thus the manifold approximately follows a line parallel (on a log-log plot) to the diagonal but offset below it by a factor of $1+F$ (dashed line). Data points in this regime can therefore be used to determine the value of F . (D) The five regimes of the allelic manifold, including approximate expressions for t_+ and t_- in each regime, as well as the range of validity for P .

then given by $F = [\text{TF}]K_F$ where $[\text{TF}]$ is the concentration of the TF and K_F is the affinity constant in inverse molar units. Similarly, the weight of the RNAP-bound state is $P = [\text{RNAP}]K_P$. In what follows we refer to F and P as the “binding factors” of the TF-DNA and RNAP-DNA interactions, respectively. We note that these binding factors can also be written as $F = e^{-\Delta G_F/k_B T}$ and $P = e^{-\Delta G_P/k_B T}$ where k_B is Boltzmann’s constant, T is temperature, and ΔG_F and ΔG_P respectively denote the Gibbs free energy of binding for the TF and RNAP. Note that each Gibbs free energy accounts for the entropic cost of pulling each protein out of solution. In what follows, we report ΔG values in units of kcal/mol; note that 1 kcal/mol = $1.62 k_B T$ at 37°C .

The overall rate of transcription is computed by summing the amount of transcription produced

99 by each state, weighting each state by the probability with which it occurs. In this case we assume
 100 the RNAP-bound state initiates at a rate of t_{sat} , and that the other states produce no transcripts. We
 101 also add a term, t_{bg} , to account for background transcription (e.g., from an unidentified promoter
 102 further upstream). The rate of transcription in the presence of the TF is thus given by

$$t_+ = t_{\text{sat}} \frac{P}{1 + F + P} + t_{\text{bg}}. \quad (1)$$

103 In the absence of the TF ($F = 0$), the rate of transcription becomes

$$t_- = t_{\text{sat}} \frac{P}{1 + P} + t_{\text{bg}}. \quad (2)$$

104 Our goal is to measure the TF-DNA binding factor F . To do this, we create a set of promoter
 105 sequences where the RNAP binding site is varied (thus generating an allelic series) but the TF binding
 106 site is kept fixed. We then measure transcription from these promoters in both the presence and
 107 absence of the TF, respectively denoting the resulting quantities by t_+ and t_- (Figure 1B). Our
 108 rationale for doing this is that changing the RNAP binding site sequence should, according to our
 109 model, affect only the RNAP-DNA binding factor P . All of our measurements are therefore expected
 110 to lie along a one-dimensional allelic manifold residing within the two-dimensional space of (t_-, t_+)
 111 values. Moreover, this allelic manifold should follow the specific mathematical form implied by
 112 Equations 1 and 2 when P is varied and the other parameters (t_{sat} , t_{bg} , F) are held fixed; see Figure
 113 1C.

114 The geometry of this allelic manifold is nontrivial. Assuming $F \gg 1$ and $t_{\text{bg}} \ll t_{\text{sat}}$, there are five
 115 different regimes corresponding to different values of the RNAP binding factor P . These regimes
 116 are listed in Figure 1D and derived in Appendix 4. In regime 1, P is so small that both t_+ and t_-
 117 are dominated by background transcription, i.e., $t_+ \approx t_- \approx t_{\text{bg}}$. P is somewhat larger in regime 2,
 118 causing t_- to be proportional to P while t_+ remains dominated by background. In regime 3, both t_+
 119 and t_- are proportional to P with $t_+/t_- \approx 1/(1+F)$. In regime 4, t_- saturates at t_{sat} while t_+ remains
 120 proportional to P . Regime 5 occurs when both t_+ and t_- are saturated, i.e., $t_+ \approx t_- \approx t_{\text{sat}}$.

121 Part 1. Demonstration: Measuring CRP-DNA binding

122 The placement of CRP immediately downstream of RNAP is known to repress transcription (*Morita et al., 1988*). We therefore reasoned that placing a DNA binding site for CRP downstream of
 123 RNAP would allow us to measure the binding factor of that site. Figure 2 illustrates measure-
 124 ments of the allelic manifold used to characterize the strength of CRP binding to the 22 bp site
 125 GAATGTGACCTAGATCACATT. This site contains the well-known consensus site, which comprises two
 126 palindromic pentamers (underlined) separated by a 6 bp spacer (*Gunasekera et al., 1992*). We
 127 performed measurements using this CRP site centered at two different locations relative to the
 128 transcription start site TSS: +0.5 bp and +4.5 bp.² To avoid influencing CRP binding strength, the
 129 -10 region of the RNAP site was kept fixed in the promoters we assayed while the -35 region of the
 130 RNAP binding site was varied (Figure 2A). Promoter DNA sequences are shown in Appendix 1 Figure
 131 1.

132 We obtained t_- and t_+ measurements for these constructs using a modified version of the
 133 colorimetric β -galactosidase assay of *Lederberg (1950)* and *Miller (1972)*; see Appendix 2 for details.
 134 Our measurements are largely consistent with an allelic manifold having the expected mathematical
 135 form (Figure 2B). Moreover, the measurements for promoters with CRP sites at two different
 136 positions (+0.5 bp and +4.5 bp) appear consistent with each other, although the measurements for
 137 +4.5 bp promoters have appear to lower values for P overall. A small number of data points do
 138 deviate substantially from this manifold, but the presence of such outliers is not surprising from a
 139 biological perspective (see Discussion). Fortunately, outliers appear at a rate small enough for us to
 140 identify them by inspection.

²The first transcribed base is, in this paper, assigned position 0 instead of the more conventional +1. Half-integer positions indicate centering between neighboring nucleotides.

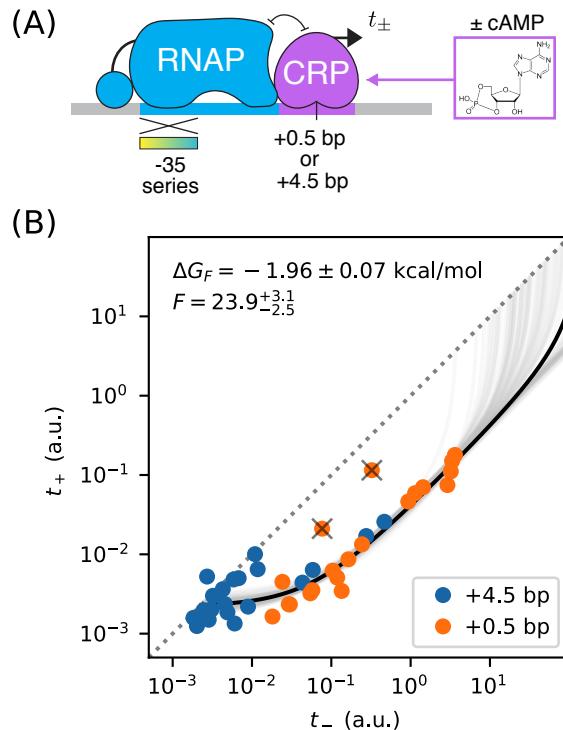


Figure 2. Precision measurement of *in vivo* CRP-DNA binding. (A) Expression measurements were performed on promoters for which CRP represses transcription by occluding RNAP. Each promoter assayed contained a near-consensus CRP binding site centered at either +0.5 bp or +4.5 bp, as well as an RNAP binding site with a partially mutagenized -35 region (gradient). t_+ (or t_-) denotes measurements made using *E. coli* strain JK10 grown in the presence (or absence) of the small molecule effector cAMP. (B) Dots indicate measurements for 41 such promoters. A best-fit allelic manifold (black) was inferred from $n = 39$ of these data points after the exclusion of 2 outliers (gray 'X's). Gray lines indicate 100 plausible allelic manifolds fit to bootstrap-resampled data points. The parameters of these manifolds were used to determine the CRP-DNA binding factor F and thus the Gibbs free energy $\Delta G_F = -k_B T \log F$. Error bars indicate 68% confidence intervals determined by bootstrap resampling. See Appendix 3 for more information about our manifold fitting procedure.

142 We quantitatively modeled the allelic manifold in Figure 2B by fitting $n + 3$ parameters to our $2n$
 143 measurements, where $n = 39$ is the number of non-outlier promoters. The $n + 3$ parameters were
 144 t_{sat} , t_{bg} , F , and P_1, P_2, \dots, P_n , where each P_i is the RNAP binding factor of promoter i . Nonlinear least
 145 squares optimization was used to infer values for these parameters. Uncertainties in t_{sat} , t_{bg} , and F
 146 were quantified by repeating this procedure on bootstrap-resampled data points.

147 These results yielded highly uncertain values for t_{sat} because none of our measurements appear
 148 to fall within regime 4 or 5 of the allelic manifold. A reasonably precise value for t_{bg} was obtained,
 149 but substantial scatter about our model predictions in regime 1 and 2 remain. This scatter likely
 150 reflects some variation in t_{bg} from promoter to promoter, variation that is to be expected since the
 151 source of background transcription is not known and the appearance of even very weak promoters
 152 could lead to such fluctuations.

153 These data do, however, determine a highly precise value for the strength of CRP-DNA binding:
 154 $F = 23.9^{+3.1}_{-2.5}$ or, equivalently, $\Delta G_F = -1.96 \pm 0.07 \text{ kcal/mol}$.³ This allelic manifold approach is thus able
 155 to measure the strength of TF-DNA binding with a precision of $\sim 0.1 \text{ kcal/mol}$. For comparison, the
 156 typical strength of a hydrogen bond in liquid water is -1.9 kcal/mol (Markovitch and Agmon, 2007).

157 We note that CRP forms approximately 38 hydrogen bonds with DNA when it binds to a consensus DNA site (Parkinson *et al.*, 1996). Our result indicates that, in living cells, the enthalpy resulting

³See Appendix 3 for a description of how these values and their uncertainties were computed.

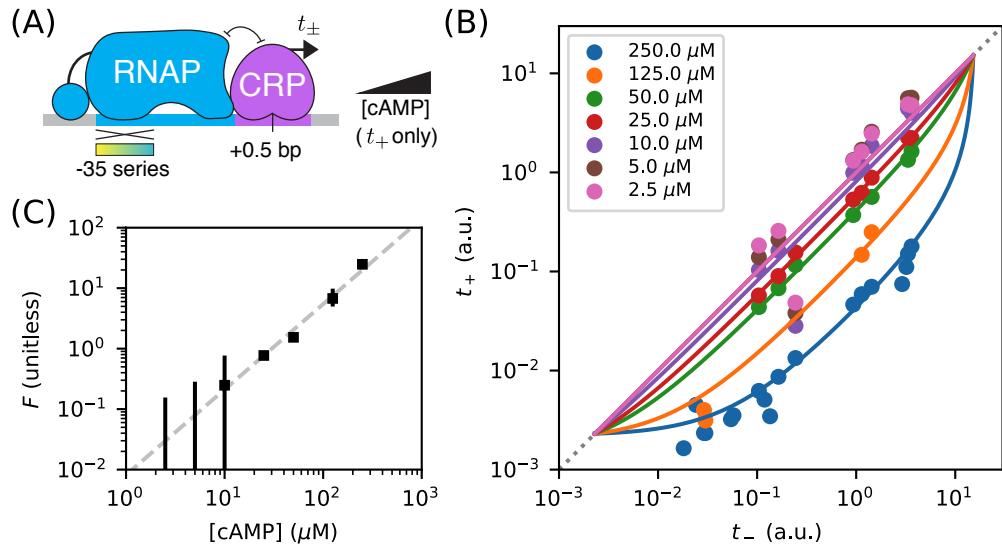


Figure 3. Measuring *in vivo* changes in TF concentration. (A) Allelic manifolds were measured for the +0.5 bp occlusion promoter architecture using seven different concentrations of cAMP (ranging from 250 μM to 2.5 μM) when assaying t_+ . (B) As expected, these data follow allelic manifolds that have cAMP-dependent values for the CRP binding factor F . (C) Values for F inferred from the data in panel B exhibit a nontrivial power law dependence on [cAMP]. Error bars indicate 68% confidence intervals determined by bootstrap resampling.

from these and other interactions is almost exactly canceled by entropic factors. We also note that our *in vivo* value for F is far smaller than expected from experiments in aqueous solution. The consensus CRP binding site has been measured *in vitro* to have an affinity constant of $K_F \sim 10^{11} \text{ M}^{-1}$ (Ebright *et al.*, 1989). There are probably about 10^3 CRP dimers per cell (Schmidt *et al.*, 2016), giving a concentration of $[\text{CPR}] \sim 10^{-6} \text{ M}$. Putting these numbers together gives a binding factor of $F \sim 10^5$. The nonspecific binding of CRP to genomic DNA and other molecules in the cell, and perhaps limited DNA accessibility as well, might be responsible for this $\sim 10^5$ -fold disagreement with our *in vivo* measurements.

Part 1. Aside: Measuring changes in the concentration of active CRP

Varying cAMP concentrations in growth media changes the *in vivo* concentration of active CRP in the *E. coli* strain we assayed (JK10). Such variation is therefore expected to alter the CRP-DNA binding factor F . We tested whether this was indeed the case by measuring multiple allelic manifolds, each using a different concentration of [cAMP] when measuring t_+ . These measurements were performed on promoters with CRP binding sites at +0.5 bp (Figure 3A). The resulting data are shown in Figure 3B. To these data, we fit allelic manifolds having variable values for F , but fixed values for both t_{bg} and t_{sat} .⁴

This procedure allowed us to quantitatively measure changes in the RNAP binding factor F , and thus changes in the *in vivo* concentration of active CRP. Our results, shown in Figure 3C, suggest a nontrivial power law relationship between F and [cAMP]. To quantify this relationship, we performed least squares regression ($\log F$ against $\log [\text{cAMP}]$) using data for the four largest cAMP concentrations; measurements of F for the three other cAMP concentrations have large asymmetric uncertainties and were therefore excluded. We found that $F \propto [\text{cAMP}]^{1.41 \pm 0.18}$, with error bars representing a 95% confidence interval. We emphasize, however that our data do not rule out a more complex relationship between [cAMP] and F .

There are multiple potential explanations for this deviation from proportionality. One possibility

⁴ $t_{\text{bg}} = 2.30 \times 10^{-3}$ a.u. was inferred in the prior analysis for Figure 2B; $t_{\text{sat}} = 15.1$ a.u. was inferred in subsequent analysis for Figure 5C.

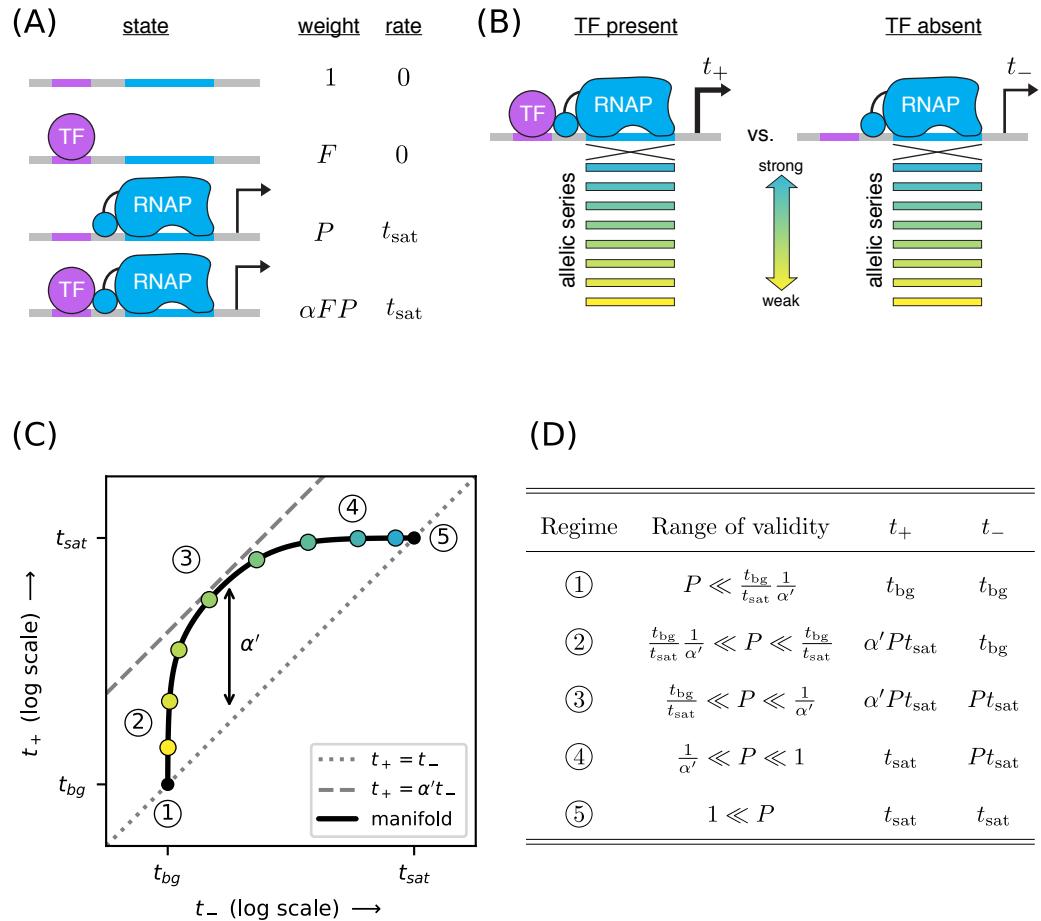


Figure 4. Strategy for measuring TF-RNAP interactions. (A) A thermodynamic model of simple activation. Here, promoter DNA can transition between four different states: unbound, bound by the TF, bound by RNAP, or doubly bound. As in Figure 1, F is the TF binding factor, P is the RNAP binding factor, and t_{sat} is the rate of transcript initiation from an RNAP-saturated promoter. The cooperativity factor α quantifies the strength of the interaction between DNA-bound TF and RNAP molecules; see text for more information on this quantity. (B) As in Figure 1, expression is measured in the presence (t_+) and absence (t_-) of the TF for promoters that have an allelic series of RNAP binding sites (blue-yellow gradient). (C) If the model in panel A is correct, plotting t_+ vs. t_- (colored dots) will reveal a 1D allelic manifold that corresponds to Equation 4 (for t_+) and Equation 2 (for t_-) evaluated over all possible values of P . Circled numbers indicate the five regimes of this manifold. In regime 3, $t_+ \approx \alpha' t_-$ where α' is the renormalized cooperativity factor given in Equation 5; data in this regime can thus be used to measure α' . Separate measurements of F , using the strategy in Figure 1, then allow one to compute α from knowledge of α' . (D) The five regimes of the allelic manifold in panel C. Note that these regimes differ from those in Figure 1D.

is cooperative binding of cAMP to the two binding sites within each CRP dimer. Such cooperativity could, for instance, result from allosteric effects like those described in *Einav et al. (2018)*. Alternatively, this power law behavior might reflect unknown aspects of how cAMP is imported and exported from *E. coli* cells. It is worth comparing and contrasting this result to those reported in *Kuhlman et al. (2007)*. JK10, the *E. coli* strain used in our experiments, is derived from strain TK310, which was developed in *Kuhlman et al. (2007)*. In that work, the authors concluded that $F \propto [\text{cAMP}]$, whereas our data leads us to reject this hypothesis. This illustrates one way in which using allelic manifolds to measure how *in vivo* TF concentrations vary with growth conditions can be useful.

192 **Part 2. Strategy: Measuring TF-RNAP interactions**

193 Next we discuss how to measure an activating interaction between a DNA-bound TF and DNA-bound
 194 RNAP. A common mechanism of transcriptional activation is “stabilization” (also called “recruitment”;
 195 see [Ptashne \(2003\)](#)). This occurs when a DNA-bound TF stabilizes the RNAP-DNA closed complex.
 196 Stabilization effectively increases the RNAP-DNA binding affinity K_p , and thus the binding factor P .
 197 It does not affect t_{sat} , the rate of transcript initiation from RNAP-DNA closed complexes.

198 A thermodynamic model for activation by stabilization is illustrated in Figure 4A. Here promoter
 199 DNA can be in four states: unbound, TF-bound, RNAP-bound, or doubly bound. In the doubly bound
 200 state, a “cooperativity factor” α contributes to the Boltzmann weight. This cooperativity factor is
 201 related to the TF-RNAP Gibbs free energy of interaction, ΔG_α , via $\alpha = e^{-\Delta G_\alpha/k_B T}$. Activation occurs
 202 when $\alpha > 1$ (i.e., $\Delta G_\alpha < 0$). The resulting activated transcription rate is given by

$$t_+ = t_{\text{sat}} \frac{P + \alpha F P}{1 + F + P + \alpha F P} + t_{\text{bg}}. \quad (3)$$

203 This can be rewritten as

$$t_+ = t_{\text{sat}} \frac{\alpha' P}{1 + \alpha' P} + t_{\text{bg}}, \quad (4)$$

204 where

$$\alpha' = \frac{1 + \alpha F}{1 + F} \quad (5)$$

205 is a renormalized cooperativity that accounts for the strength of TF-DNA binding. As before, t_- is
 206 given by Equation 2. Note that $\alpha' \leq \alpha$ and that $\alpha' \approx \alpha$ when $F \gg 1$ and $\alpha \gg 1/F$.

207 As before, we measure both t_+ and t_- for an allelic series of RNAP binding sites (Figure 4B).
 208 These measurements will, according to our model, lie along an allelic manifold resembling the one
 209 shown in Figure 4C. This allelic manifold exhibits five distinct regimes (when $t_{\text{sat}}/t_{\text{bg}} \gg \alpha' \gg 1$) listed
 210 in Figure 4D.

211 **Part 2. Demonstration: Measuring class I CRP-RNAP interactions**

212 CRP activates transcription at the *lac* promoter and at other promoters by binding to a 22 bp site
 213 centered at -61.5 bp relative to the TSS. This is an example of class I activation, which is mediated
 214 by an interaction between CRP and the C-terminal domain of one of the two RNAP α subunits (the
 215 α CTDs) ([Busby and Ebright, 1999](#)). *In vitro* experiments have shown this class I CRP-RNAP interaction
 216 to activate transcription by stabilizing the RNAP-DNA closed complex.

217 We measured t_+ and t_- for 47 variants of the *lac** promoter (see Appendix 1 Figure 1 for
 218 sequences). These promoters have the same CRP binding site assayed for Figure 2, but positioned
 219 at -61.5 bp relative to the TSS (Figure 5A). They differ from one another in the -10 or -35 regions of
 220 their RNAP binding sites. Figure 5B shows the resulting measurements. With the exception of 3
 221 outlier points, these measurements appear consistent with stabilizing activation via a Gibbs free
 222 energy of $\Delta G_\alpha = -4.05 \pm 0.08$ kcal/mol, corresponding to a cooperativity of $\alpha = 712^{+102}_{-83}$. We note that,
 223 with $F = 23.9$ determined in Figure 2B, $\alpha' = \alpha$ to 4% accuracy.

224 This observed cooperativity is substantially stronger than suggested by previous work. Early *in*
 225 *vivo* experiments suggested a much lower cooperativity value, e.g. 50-fold ([Beckwith et al., 1972](#)), 20-
 226 fold ([Ushida and Aiba, 1990](#)), or even 10-fold ([Gaston et al., 1990](#)). These previous studies, however,
 227 only measured the ratio t_+/t_- for a specific choice of RNAP binding site. This ratio is (by Equation
 228 4) always less than α and the differences between these quantities can be substantial. However,
 229 even studies that have used explicit biophysical modeling have determined lower cooperativity
 230 values: [Kuhlman et al. \(2007\)](#) reported a cooperativity of $\alpha \approx 240$ ($\Delta G_\alpha \approx -3.4$ kcal/mol), while
 231 [Kinney et al. \(2010\)](#) reported $\alpha \approx 220$ ($\Delta G_\alpha \approx -3.3$ kcal/mol). Both of these studies, however, relied
 232 on the inference of complex biophysical models with many parameters. The allelic manifold in
 233 Figure 4, by contrast, is characterized by only three parameters (t_{sat} , t_{bg} , α'), all of which can be
 234 approximately determined by visual inspection.

235 To test the generality of this approach, we measured allelic manifolds for 11 other potential
 236 class I promoter architectures. At every one of these positions we clearly observed the collapse of

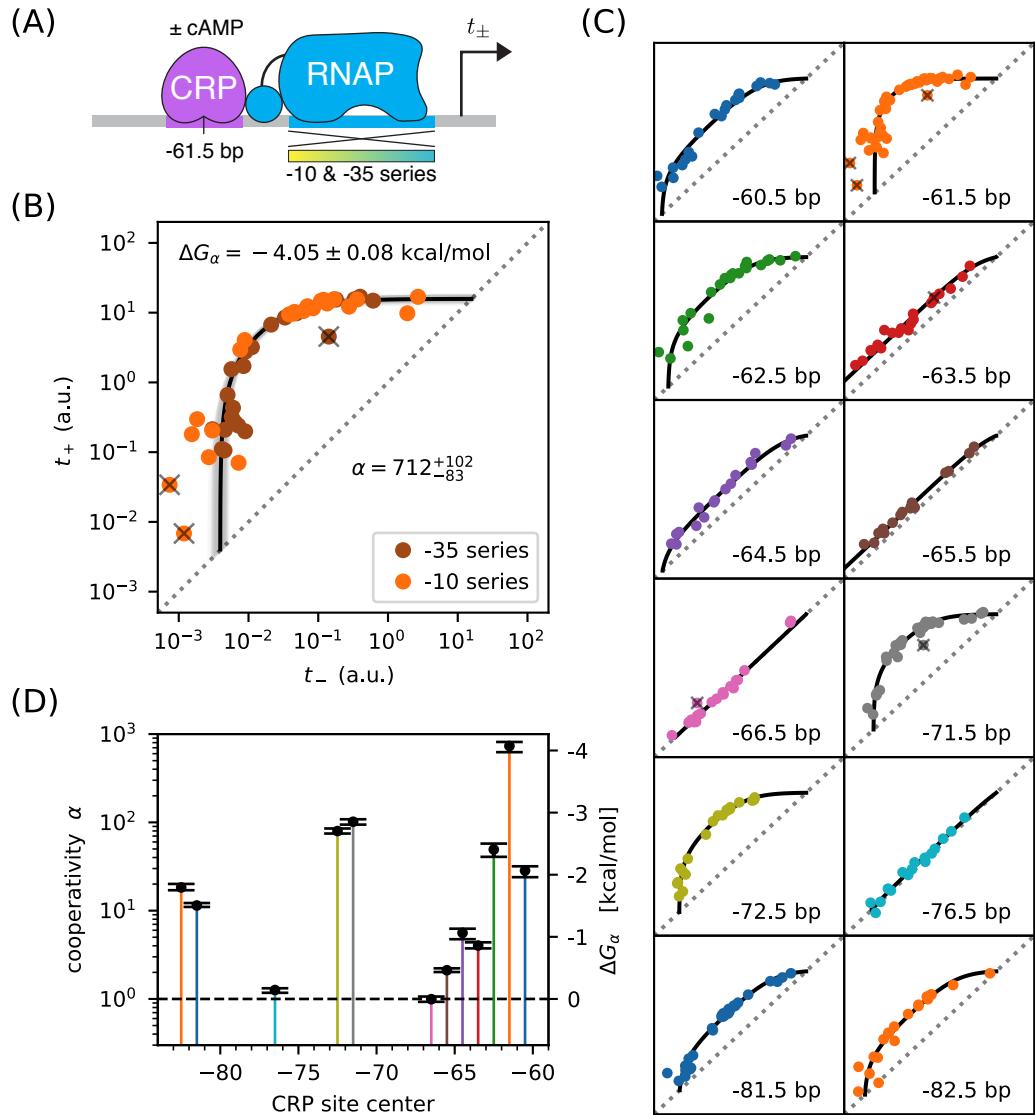


Figure 5. Precision measurement of class I CRP-RNAP interactions. (A) t_+ and t_- were measured for promoters containing a CRP binding site centered at -61.5 bp. The RNAP sites of these promoters were mutagenized in either their -10 or -35 regions (gradient), generating two allelic series. As in Figure 2, t_+ and t_- correspond to expression measurements respectively made in the presence and absence of cAMP. (B) Data obtained for 47 variant promoters having the architecture shown in panel A. Three data points designated as outliers are indicated by 'X's. The allelic manifold that best fits the $n = 44$ non-outlier points is shown in black; 100 plausible manifolds, estimated from bootstrap-resampled data points, are shown in gray. The resulting values for α and $\Delta G_\alpha = -k_B T \log \alpha$ are also shown, with 68% confidence intervals indicated. (C) Allelic manifolds obtained for promoters with CRP binding sites centered at a variety of class I positions. (D) Inferred values for the cooperativity factor α and corresponding Gibbs free energy ΔG_α for the 12 different promoter architectures assayed in panel C. Error bars indicate 68% confidence intervals. Numerical values for α and ΔG_α at all of these class I positions are provided in Table 1.

237 data to a 1D allelic manifold of the expected shape (Figure 5C). We then modeled these data using
 238 values of α and t_{bg} that depend on CRP binding site location, as well as a single overall value for t_{sat} .
 239 The resulting values for α (and equivalently ΔG_α) are shown in Figure 5D and reported in Table 1. As
 240 first shown by **Gaston et al. (1990)** and **Ushida and Aiba (1990)**, α depends strongly on the spacing
 241 between the CRP and RNAP binding sites. In particular, α exhibits a strong ~ 10.5 bp periodicity
 242 reflecting the helical twist of DNA. However, as with the measurement in Figure 5B, the α values we

measure are far larger than the t_+/t_- ratios previously reported by **Gaston et al. (1990)** and **Ushida and Aiba (1990)**; see Table 1. We also find $t_{\text{sat}} = 15.1^{+0.6}_{-0.5}$ a.u.. The single-cell observations of **So et al. (2011)** suggest that this corresponds to 13.8 ± 6.6 transcripts per minute.⁵

Part 2. Aside: Difficulties predicting binding affinity from DNA sequence.

The measurement and modeling of allelic manifolds sidesteps the need to parametrically model how protein-DNA binding affinity depends on DNA sequence. In modeling the allelic manifolds in Figure 5C, we obtained values for the RNAP binding factor, $P = [\text{RNAP}]K_p$, for each variant RNAP binding site from the position of the corresponding data point along the length of the manifold.

RNAP has a very well established sequence motif (**McClure et al., 1983**). Indeed, its DNA binding requirements were among the first characterized for any DNA-binding protein (**Pribnow, 1975**). More recently, a high-resolution model for RNAP-DNA binding energy was determined using data from a massively parallel reporter assay called Sort-Seq (**Kinney et al., 2010**). This position-specific affinity matrix (PSAM)⁶ assumes that the nucleotide at each position contributes additively to the overall binding energy (Figure 6). This model is consistent with previously described RNAP binding motifs but, unlike those motifs, it can predict binding energy in physically meaningful energy units (i.e., kcal/mol). In what follows we denote these binding energies as $\Delta\Delta G_p$, because they describe differences in the Gibbs free energy of binding between two DNA sites.

There is good reason to believe this PSAM to be the most accurate current model of RNAP-DNA binding. However, subsequent work has suggested that the predictions of this model might still have substantial inaccuracies (**Brewster et al., 2012**). To investigate this possibility, we compared our measured values for the Gibbs free energy of RNAP-DNA binding ($\Delta G_p = -k_B T \log P$) to binding energies ($\Delta\Delta G$) predicted using the PSAM from **Kinney et al. (2010)**. These values are plotted against one another in Figure 6B. Although there is a strong correlation between the predictions of the

⁵By pure coincidence, the “arbitrary unit” (a.u.) units we use in this paper correspond very closely to “transcripts per minute”.

⁶The term PSAM comes from **Foat et al. (2006)**. These models are called “energy matrices” in **Kinney et al. (2010)** and **Belliveau et al. (2018)**.

Table 1. Summary of results for class I activation by CRP. The α and ΔG_α values listed here correspond to the values plotted in Figure 5D. The corresponding value inferred for the saturated transcription rate is $t_{\text{sat}} = 15.1^{+0.6}_{-0.5}$ a.u.. Error bars indicate 68% confidence intervals; see Appendix 3 for details. n is the number of data points used to infer these values, while “outliers” is the number of data points excluded in this analysis. For comparison we show the fold-activation measurements (i.e., t_+/t_-) reported in **Gaston et al. (1990)** and **Ushida and Aiba (1990)**; ‘-’ indicates that no measurement was reported for that position.

position (bp)	n	outliers	ΔG_α (kcal/mol)	α	t_+/t_- (Gaston)	t_+/t_- (Ushida)
-60.5	21	0	-2.09 ± 0.08	$29.6^{+4.7}_{-3.5}$	3.85	-
-61.5	44	3	-4.10 ± 0.08	763^{+113}_{-84}	9.05	20.6
-62.5	23	0	-2.43 ± 0.11	$51.4^{+9.0}_{-8.5}$	4.22	-
-63.5	20	1	-0.88 ± 0.05	$4.15^{+0.30}_{-0.37}$	-	-
-64.5	17	0	-1.08 ± 0.08	$5.80^{+0.89}_{-0.67}$	-	-
-65.5	17	0	-0.48 ± 0.03	$2.16^{+0.10}_{-0.11}$	-	-
-66.5	19	1	0.00 ± 0.04	$0.99^{+0.07}_{-0.07}$	0.78	0.84
-71.5	35	1	-2.88 ± 0.04	105^{+7}_{-7}	2.50	16.4
-72.5	20	0	-2.73 ± 0.04	$83.0^{+5.2}_{-5.8}$	3.49	-
-76.5	16	0	-0.15 ± 0.04	$1.27^{+0.09}_{-0.06}$	0.54	-
-81.5	32	0	-1.53 ± 0.03	$11.9^{+0.4}_{-0.8}$	-	-
-82.5	20	0	-1.82 ± 0.05	$19.0^{+1.3}_{-1.8}$	-	6.99

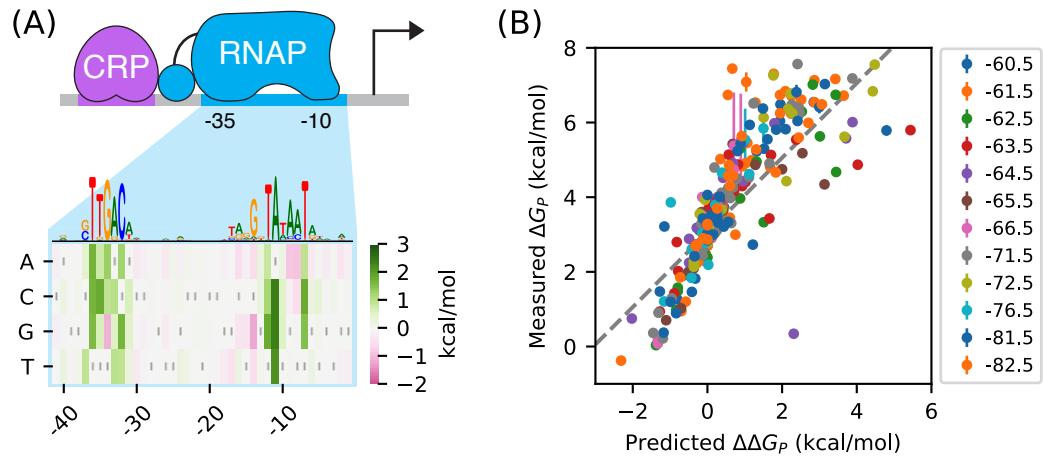


Figure 6. RNAP-DNA binding energy cannot be accurately predicted from sequence. (A) The PSAM for RNAP-DNA binding inferred by *Kinney et al. (2010)*. This model assumes that the DNA base pair at each position in the RNAP binding site contributes independently to ΔG_P . Shown are the $\Delta \Delta G_P$ values assigned by this model to mutations away from the lac* RNAP site. The sequence of the lac* RNAP site is indicated by gray vertical bars; see also Appendix 1 Figure 1. A sequence logo representation for this PSAM is provided for reference. (B) PSAM predictions plotted against the values of $\Delta G_P = -k_B T \log P$ inferred by fitting the allelic manifolds in Figure 5C. Error bars on these measurements represent 68% confidence intervals. Note that measured ΔG_P values are absolute, whereas the $\Delta \Delta G_P$ predictions of the PSAM are relative to the lac* RNAP site, which thus corresponds to $\Delta \Delta G_P = 0$ kcal/mol.

model and our measurements, deviations of 1 kcal/mol or larger (corresponding to variations in P of 5-fold or greater) are not uncommon. Model predictions also systematically deviate from the diagonal, suggesting inaccuracy in the overall scale of the PSAM.

This finding is sobering: even for one of the best understood DNA-binding proteins in biology, our best sequence-based predictions of *in vivo* protein-DNA binding affinity are still quite crude. When used in conjunction with thermodynamic models, as in *Kinney et al. (2010)*, the inaccuracies of these models can have major effects on predicted transcription rates. The measurement and modeling of allelic manifolds sidesteps the need to parametrically model such binding energies, enabling the direct inference of Gibbs free energy values for each assayed RNAP binding site.

Part 3. Strategy: Distinguishing mechanisms of transcriptional activation

E. coli TFs can regulate multiple different steps in the transcript initiation pathway (*Lee et al., 2012; Browning and Busby, 2016*). For example, instead of stabilizing RNAP binding to DNA, TFs can activate transcription by increasing the rate at which DNA-bound RNAP initiates transcription (*Roy et al., 1998*), a process we refer to as “acceleration”. CRP, in particular, has previously been reported to activate transcription in part by acceleration when positioned appropriately with respect to RNAP (*Niu et al., 1996; Rhodius et al., 1997*).

We investigated whether allelic manifolds might be used to distinguish activation by acceleration from activation by stabilization. First we generalized the thermodynamic model in Figure 4A to accommodate both α -fold stabilization and β -fold acceleration (Figure 7A). This is accomplished by using the same set of states and Boltzmann weights as in the model for stabilization, but assigning a transcription rate βt_{sat} (rather than just t_{sat}) to the TF-RNAP-DNA ternary complex. The resulting activated rate of transcription is given by

$$t_+ = t_{\text{sat}} \frac{P}{1 + F + P + \alpha FP} + \beta t_{\text{sat}} \frac{\alpha FP}{1 + F + P + \alpha FP} + t_{\text{bg}}. \quad (6)$$

This simplifies to

$$t_+ = \beta' t_{\text{sat}} \frac{\alpha' P}{1 + \alpha' P} + t_{\text{bg}}, \quad (7)$$

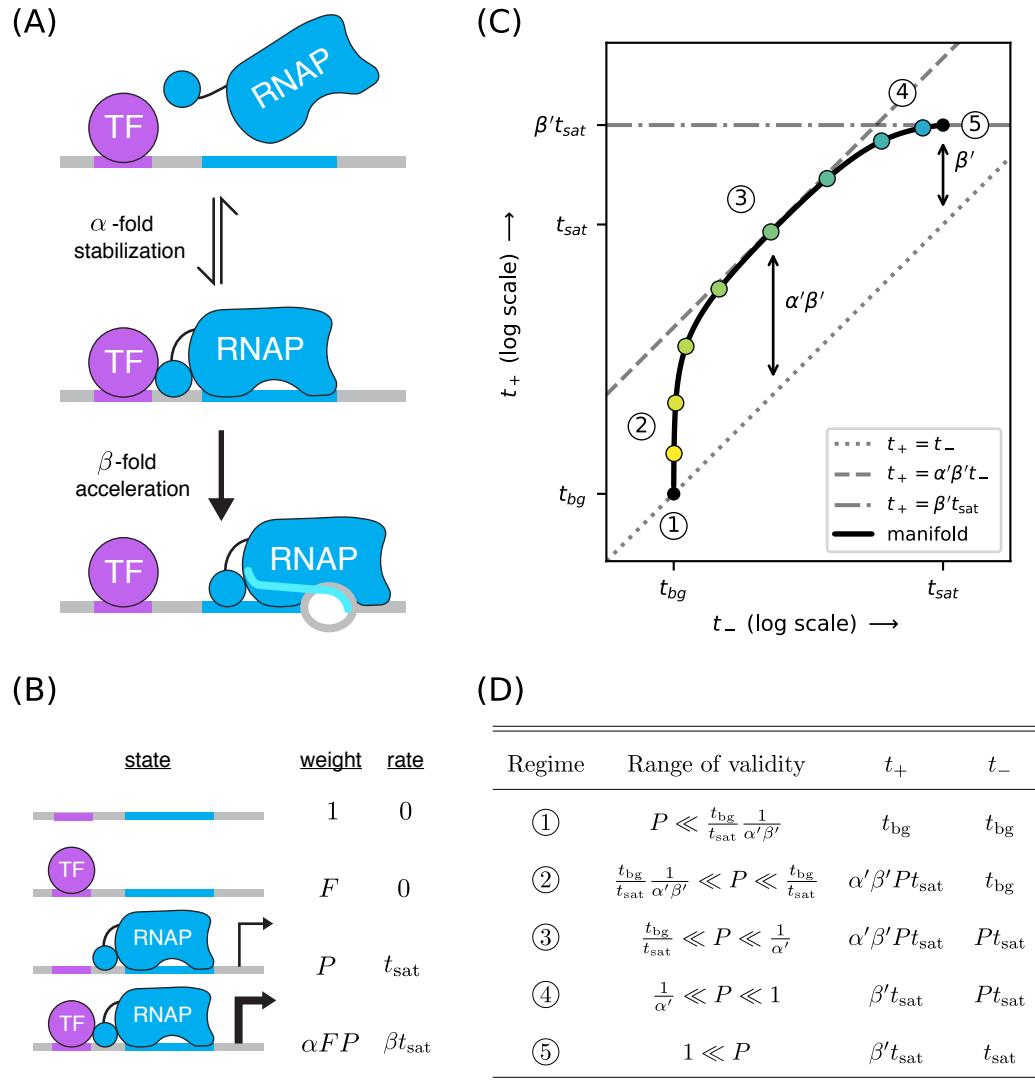


Figure 7. A strategy for distinguishing two different mechanisms of transcriptional activation. (A) A TF can activate transcription in two ways: (i) by stabilizing the RNAP-DNA complex or (ii) by accelerating the rate at which this complex initiates transcripts. (B) A thermodynamic model for the dual mechanism of transcriptional activation illustrated in panel A. Note that α multiplies the Boltzmann weight of the doubly bound complex, whereas β multiplies the transcript initiation rate of this complex. (C) Data points measured as in Figure 4C will lie along a 1D allelic manifold having the form shown here. This manifold is computed using t_+ values from Equation 7 and t_- values from Equation 2. Note that regime 5 occurs at a point positioned β' -fold above the diagonal, where β' is related to β through Equation 8. Measurements in or near the strong promoter regime ($P \gtrsim 1$) can thus be used to determine the value of β' and, consequently, the value of β . (D) The five regimes of this allelic manifold are listed.

where α' is the same as in Equation 5 and

$$\beta' = \frac{1 + \alpha \beta F}{1 + \alpha F} \quad (8)$$

is a renormalized version of the acceleration rate β . The resulting allelic manifold is illustrated in Figure 7C. Like the allelic manifold for stabilization, this manifold has up to five distinct regimes corresponding to different values of P (Figure 7D). Unlike the stabilization manifold however, $t_+ \neq t_-$ in the strong RNAP binding regime (regime 5); rather, $t_+ \approx \beta' t_{\text{sat}}$ while $t_- \approx t_{\text{sat}}$.

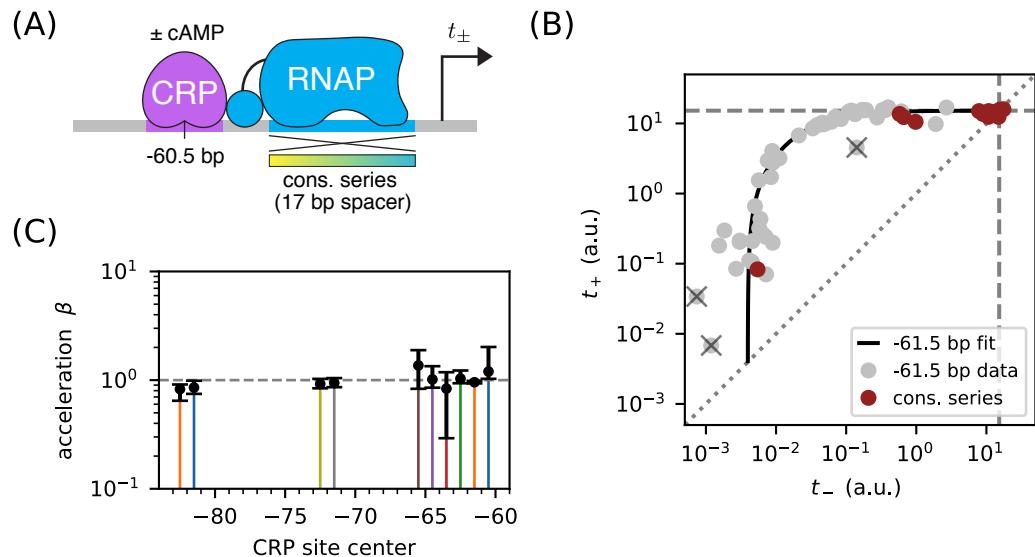


Figure 8. Class I activation by CRP occurs exclusively through stabilization. (A) t_+ and t_- were measured for promoters containing variants of the consensus RNAP binding site as well as a CRP binding site centered at -60.5 bp. Because the consensus RNAP site is 1 bp shorter than the RNAP site of the lac* promoter, CRP at -60.5 bp here corresponds to CRP at -61.5 bp in Figure 5. (B) $n = 18$ data points obtained for the constructs in panel A, overlaid on the measurements from Figure 5B (gray). The value $t_{\text{sat}} = 15.1$ a.u., inferred for Figure 5C, is indicated by dashed lines. (C) Values for β inferred using the data in Figure 5 for the 10 CRP positions that exhibited greater than 2-fold inducibility; β values at the two other CRP positions (-66.5 bp and -76.5 bp) were highly uncertain and are not shown. Error bars indicate 68% confidence intervals.

294 Part 3. Demonstration: Mechanisms of class I activation by CRP

295 We asked whether class I activation by CRP has an acceleration component. Previous *in vitro* work
 296 had suggested that the answer is ‘no’ (*Malan et al., 1984; Busby and Ebright, 1999*), but our allelic
 297 manifold approach allows us to address this question *in vivo*. We proceeded by assaying promoters
 298 containing variant alleles of the consensus RNAP binding site (Figure 8A). Note that the consensus
 299 RNAP site is 1 bp shorter than the lac* RNAP site (Appendix 1, Figure 1C versus Figure 1B). We
 300 therefore positioned the CRP binding site at -60.5 bp in order to realize the same spacing between
 301 CRP and the -35 element of the RNAP binding site that was realized in -61.5 bp non-consensus
 302 promoters.

303 The resulting data (Figure 8B) are seen to largely fall along the previously measured all-stabilization
 304 allelic manifold in Figure 5B. In particular, many of these data points lie at the intersection of this
 305 manifold with the $t_+ = t_-$ diagonal. We thus find that $\beta \approx 1$ for CRP at -61.5 bp. To further quantify
 306 possible β values, we fit the acceleration model in Figure 7 to each dataset shown in Figure 5B,
 307 assuming a fixed value of $t_{\text{sat}} = 15.1$ a.u.. The resulting inferred values for β , shown in Figure 8C,
 308 indicate little if any deviation from $\beta = 1$. Our high-precision *in vivo* results therefore substantiate
 309 the previous *in vitro* results of *Malan et al. (1984)* regarding the mechanism of class I activation.

310 Part 3. Aside: Surprises in class II regulation by CRP

311 Many *E. coli* TFs participate in what is referred to as class II activation (*Browning and Busby, 2016*).
 312 This type of activation occurs when the TF binds to a site that overlaps the -35 element (often com-
 313 pletely replacing it) and interacts directly with the main body of RNAP. CRP is known to participate
 314 in class II activation at many promoters (*Keseler et al., 2011; Salgado et al., 2013*), including the
 315 galP1 promoter, where it binds to a site centered at position -41.5 bp (*Adhya, 1996*). *In vitro* studies
 316 have shown CRP to activate transcription at -41.5 bp relative to the TSS through a combination of
 317 stabilization and acceleration (*Niu et al., 1996; Rhodius et al., 1997*).

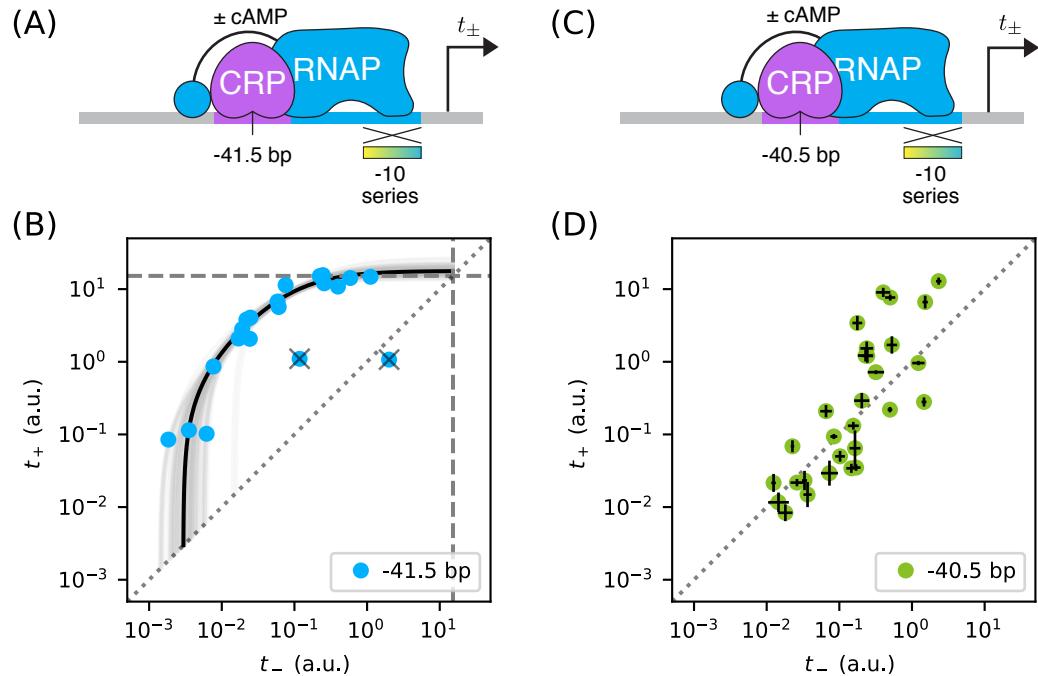


Figure 9. Surprises in class II regulation by CRP. (A) Regulation by CRP centered at -41.5 bp was assayed using an allelic series of RNAP binding sites that have variant -10 elements (gradient). (B) The observed allelic manifold plateaus at the value of $t_{\text{sat}} = 15.1$ a.u. (dashed lines) determined for Figure 5B, thus indicating no detectable acceleration by CRP. This lack of acceleration is at odds with prior *in vitro* studies (*Niu et al., 1996; Rhodius et al., 1997*). (C) Regulation by CRP centered at -40.5 bp was assayed in an analogous manner. (D) Unexpectedly, data from the promoters in panel C do not collapse to a 1D allelic manifold. This finding falsifies the biophysical models in Figures 4A and 7B and indicates that CRP can either activate or repress transcription from this position, depending on as-yet-unidentified features of the RNAP binding site. Error bars in panel D indicate 95% confidence intervals estimated from replicate experiments.

318 We sought to reproduce this finding *in vivo* by measuring allelic manifolds. We therefore placed
 319 a consensus CRP site at -41.5 bp, replacing much of the -35 element in the process, and partially
 320 mutated the -10 element of the RNAP binding site (Figure 9A). Surprisingly, we observed that the
 321 resulting allelic manifold saturates at the same t_{sat} value shared by all class I promoters. Thus,
 322 CRP appears to activate transcription *in vivo* solely through stabilization, and not at all through
 323 acceleration, when located at -41.5 bp relative to the TSS (Figure 9B).

324 The genome-wide distribution of CRP binding sites suggests that CRP also participates in class
 325 II activation when centered at -40.5 bp (*Keseler et al., 2011; Salgado et al., 2013*). When assaying
 326 this promoter architecture, however, we obtained a scatter of 2D points that did not collapse to
 327 any discernible 1D allelic manifold (Figure 9D). Some of these promoters exhibit activation, some
 328 exhibit repression, and some exhibit no regulation by CRP.

329 These observations complicate the current understanding of class II regulation by CRP. Our *in*
 330 *vivo* measurements of CRP at -41.5 bp call into question the mechanism of activation previously
 331 discerned using *in vitro* techniques. The scatter observed when CRP is positioned at -40.5 bp
 332 suggests that, at this position, the -10 region of the RNAP binding site influences the values of
 333 at least two relevant biophysical parameters (not just P , as our model predicts). A potential
 334 explanation for both observations is that, because CRP and RNAP are so intimately positioned at
 335 class II promoters, even minor changes in their relative orientation caused by differences between
 336 *in vivo* and *in vitro* conditions or by changes in RNAP site sequence could have a major effect on
 337 CRP-RNAP interactions. Such sensitivity would not be expected to occur in class I activation, due to

338 the flexibility with which the RNAP α CTDs are tethered to the core complex of RNAP.

339 Discussion

340 We have shown how the measurement and quantitative modeling of allelic manifolds can be used
 341 to dissect cis-regulatory biophysics in living cells. This approach was demonstrated in *E. coli* in
 342 the context of transcriptional regulation by two well-characterized TFs: RNAP and CRP. Here we
 343 summarize our primary findings. We then address some caveats and limitations of the work
 344 reported here. Finally, we elaborate on how future studies might be able to scale up this approach
 345 using massively parallel reporter assays (MPRAs), including for studies in eukaryotic systems.

346 Summary

347 In each of our experiments, we quantitatively measured transcription from an allelic series of
 348 variant RNAP binding sites, each site embedded in a fixed promoter architecture. Two expression
 349 measurements were made for each variant promoter: t_+ was measured in the presence of the
 350 active form of CRP, while t_- was measured in the absence of active CRP. This yielded a data point,
 351 (t_-, t_+) , in a two-dimensional measurement space. We had expected the data points thus obtained
 352 for each allelic series to collapse to a 1D curve (the allelic manifold), with different positions along
 353 this manifold corresponding to different values of RNAP-DNA binding affinity. Such collapse was
 354 indeed observed in all but one of the promoter architectures we studied. By fitting the parameters
 355 of quantitative biophysical models to these data, we obtained *in vivo* values for the Gibbs free
 356 energy (ΔG) of a variety of TF-DNA and TF-TF interactions.

357 In Part 1, we showed how measuring allelic manifolds for promoters in which a DNA-bound TF
 358 occludes RNAP can allow one to precisely measure the ΔG of TF-DNA binding. We demonstrated
 359 this strategy on promoters where CRP occludes RNAP, thereby obtaining the ΔG for a CRP binding
 360 site that was used in subsequent experiments. As an aside, we demonstrated how performing such
 361 measurements in different concentrations of the small molecule cAMP allowed us to quantitatively
 362 measure *in vivo* changes in active CRP concentration.

363 In Part 2, we showed how allelic manifolds can be used to measure the ΔG of TF-RNAP inter-
 364 actions. We used this strategy to measure the stabilizing interactions by which CRP up-regulates
 365 transcription at a variety of class I promoter architectures. Our strategy consistently yielded ΔG
 366 values with an estimated precision of ~ 0.1 kcal/mol. As an aside, we showed how ΔG values for
 367 RNAP-DNA binding could also be obtained from these data. Notably, these ΔG measurements for
 368 RNAP-DNA binding were seen to deviate substantially from sequence-based predictions using an
 369 established position-specific affinity matrix (PSAM) for RNAP. This highlights just how difficult it can
 370 be to accurately predict TF-DNA binding affinity from DNA sequence.

371 In Part 3, we showed how allelic manifolds can allow one to distinguish between two potential
 372 mechanisms of transcriptional activation: "stabilization" (a.k.a. "recruitment") and "acceleration".
 373 Applying this approach to the data from Part 2, we confirmed (as expected) that class I activation by
 374 CRP does indeed occur through stabilization and not acceleration. As an aside, we pursued this
 375 approach at two class II promoters. In contrast to prior *in vitro* studies (Niu et al., 1996; Rhodius
 376 et al., 1997), no acceleration was observed when CRP was positioned at -41.5 bp relative to the TSS.
 377 Even more unexpectedly, no 1D allelic manifold was observed at all when CRP was positioned at
 378 -40.5 bp. This last finding indicates that the variant RNAP binding sites we assayed control at least
 379 one functionally important biophysical quantity in addition to RNAP-DNA binding affinity.

380 Caveats and limitations

381 An important caveat is that our ΔG measurements assume that the *true* transcription rates (of which
 382 we obtain only noisy measurements) exactly fall along a 1D allelic manifold of the hypothesized
 383 mathematical form. These assumptions are well-motivated by the data collapse that we observed
 384 for all except one promoter architecture. But for some promoter architectures, there were a small
 385 number of "outlier" data points that we judged (by eye) to deviate substantially from the inferred

386 allelic manifold. The presence of a few outliers makes sense biologically: The random mutations
 387 we introduced into variant RNAP binding sites will, with some nonzero probability, either shift
 388 the position of the RNAP site or create a new binding site for some other TF. However, even for
 389 promoters that exhibit clear clustering of 2D data around a 1D curve, the deviations of individual
 390 non-outlier data points from our inferred allelic manifold were often substantially larger than the
 391 experimental noise that we estimated from replicates. It may be that the biological cause of outliers
 392 is not qualitatively different from what causes these smaller but still detectable deviations from our
 393 assumed model.

394 The low-throughput experimental approach we pursued here also has important limitations.
 395 Each of the 448 variant promoters for which we report data was individually catalogued, sequenced,
 396 and assayed in at least three replicate experiments for both ι_+ and ι_- . We opted to use a low-
 397 throughput colorimetric assay of β -galactosidase activity (*Lederberg, 1950; Miller, 1972*) because
 398 this approach is well established in *E. coli* to produce a quantitative measure of transcription with
 399 high precision and high dynamic range. Such assays have also been used by other groups to
 400 develop sophisticated biophysical models of transcriptional regulation (*Kuhlman et al., 2007; Cui*
 401 *et al., 2013*). However, this low-throughput approach has limited utility because it cannot be readily
 402 scaled up.

403 Our reliance on cAMP as a small molecule effector of CRP presents a second limitation. In
 404 our experiments, we controlled the *in vivo* activity of CRP by growing a specially designed strain
 405 of *E. coli* in either the presence (for ι_+) or absence (ι_-) of cAMP. This mirrors the strategy used by
 406 *Kuhlman et al. (2007)*, and the validity of this approach is attested to by the calibration data shown
 407 in Appendix 2 Figure 1. However, controlling *in vivo* TF activity using small molecules has many
 408 limitations. Most TFs cannot be quantitatively controlled with small molecules, and those that
 409 can often require special host strains (e.g., see *Kuhlman et al. (2007)*). Moreover, varying the *in*
 410 *vivo* concentration of a TF can affect cellular physiology in ways that can confound quantitative
 411 measurements.

412 **Outlook**

413 MPRAs performed on array-synthesized promoter libraries should be able to overcome both of
 414 these experimental limitations. Current MPRA technology is able to quantitatively measure gene
 415 expression for $\gtrsim 10^4$ transcriptional regulatory sequences in parallel. We estimate that this would
 416 enable the simultaneous measurement of $\sim 10^2$ highly resolved allelic manifolds, each manifold
 417 representing a different promoter architecture. Moreover, by using array-synthesized promoters in
 418 conjunction with MPRAs, one can measure ι_+ and ι_- by systematically altering the DNA sequence of
 419 TF binding sites, rather than relying on small molecule effectors of each TF. This capability would,
 420 among other things, enable biophysical studies of promoters that have multiple binding sites for
 421 the same TF; in such cases it might make sense to use measurement spaces having more than two
 422 dimensions.

423 Will allelic manifolds be useful for understanding transcriptional regulation in eukaryotes?
 424 Both FACS-based MPRAs (*Sharon et al., 2012; Weingarten-Gabbay et al., 2017*) and RNA-Seq-based
 425 MPRAs (*Melnikov et al., 2012; Kwasnieski et al., 2012; Patwardhan et al., 2012*) are well established
 426 in eukaryotes so, on a technical level, experiments analogous to those described here should be
 427 feasible. The bigger question, we believe, is whether the results of such experiments would
 428 be interpretable. Eukaryotic transcriptional regulation is far more complex than transcriptional
 429 regulation in bacteria. Still, we believe that pursuing the measurement and modeling of allelic
 430 manifolds in this context is worthwhile. Despite the underlying complexities, simple “effective”
 431 biophysical models might work surprisingly well. Similar approaches might also be useful for
 432 studying other eukaryotic regulatory processes that are compatible with MPRAs, such as alternative
 433 splicing (*Wong et al., 2018*).

434 Based on these results, we advocate a very different approach to dissecting *cis*-regulatory
 435 grammar than has been pursued by other groups. Rather than attempting to identify a single

436 quantitative model that can explain regulation by many different arrangements of TF binding sites
 437 (**Gertz et al., 2009; Sharon et al., 2012; Mogno et al., 2013; Smith et al., 2013; Levo and Segal, 2014;**
 438 **White et al., 2016**), we suggest focused studies of the biophysical interactions that result from
 439 specific TF binding site arrangements. The measurement and modeling of allelic manifolds provides
 440 a systematic and stereotyped way of doing this. By coupling this approach with MPRA, it should
 441 be possible to perform such studies on hundreds of systematically varied regulatory sequence
 442 architectures in parallel. General rules governing cis-regulatory grammar might then be identified
 443 empirically. We suspect that this bottom-up strategy to studying cis-regulatory grammar is likely to
 444 reveal regulatory mechanisms that would be hard to anticipate in top-down studies.

445 Materials and Methods

446 Appendix 1 describes the media, strains, plasmids, and promoters assayed in this work. Appendix
 447 2 describes the colorimetric β -galactosidase activity assay, adapted from **Lederberg (1950)** and
 448 **Miller (1972)**, that was used to measure expression levels. Appendix 3 provides details about how
 449 quantitative models were fit to these measurements, as well as how uncertainties in estimated
 450 parameters were computed. Supplemental File 1 is an Excel spreadsheet containing the DNA
 451 sequences of all assayed promoters, all t_+ and t_- measurements used in this work, and all of the
 452 parameter values fit to these data, both with and without bootstrap resampling.

453 Author contributions

454 JBK conceived of this study. TF and JBK designed this study. JBK, TF, AA, MSG, and DJ carried out the
 455 experiments. TF and JBK carried out the computational analysis. JBK wrote the manuscript with
 456 input from MSG, RP, DJ, TF, and AA. JBK funded this study.

457 Acknowledgments

458 We thank Stirling Churchman, Barak Cohen, David McCandlish, Bryce Nickels, and Saurabh Sinha
 459 for helpful discussions. We also thank Naama Barkai, Ulrich Gerland, Richard Neher, and one
 460 anonymous referee for reviewing this manuscript and providing helpful feedback. This work was
 461 supported by a CSHL/Northwell Health Alliance grant to JBK and by NIH Cancer Center Support
 462 Grant 5P30CA045508.

463 References

- 464 **Ackers G, Johnson A, Shea M.** Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci U S A. 1982 Feb; 79(4):1129–1133.
- 466 **Adhya S.** The lac and gal operons today. Regulation of Gene Expression in *Escherichia coli*. 1996; p. 1–20.
- 467 **Beckwith J, Grodzicker T, Arditti R.** Evidence for two sites in the lac promoter region. J Mol Biol. 1972 Aug; 69(1):155–160.
- 469 **Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, Hess S, Kinney JB, Phillips R.** Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc Natl Acad Sci USA. 2018 May; 115(21):E4796–E4805.
- 472 **Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R.** Transcriptional regulation by the numbers: models. Curr Opin Genet Dev. 2005 Apr; 15(2):116–124.
- 474 **Brewster RC, Jones DL, Phillips R.** Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. PLoS Comput Biol. 2012; 8(12):e1002811.
- 476 **Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R.** The transcription factor titration effect dictates level of gene expression. Cell. 2014 Mar; 156(6):1312–1323.
- 478 **Browning DF, Busby SJW.** Local and global regulation of transcription initiation in bacteria. Nat Rev Microbiol. 2016 Oct; 14(10):638–650.

- 480 **Busby S**, Ebright RH. Transcription activation by catabolite activator protein (CAP). *J Mol Biol.* 1999 Oct; 481 293(2):199–213.
- 482 **Courey AJ**. Mechanisms in transcriptional regulation. Malden, MA: Blackwell; 2008.
- 483 **Cui L**, Murchland I, Shearwin KE, Dodd IB. Enhancer-like long-range transcriptional activation by λ CI-mediated 484 DNA looping. *Proc Natl Acad Sci USA.* 2013 Feb; 110(8):2922–2927.
- 485 **Ebright RH**, Ebright YW, Gunasekera A. Consensus DNA site for the Escherichia coli catabolite gene activator 486 protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA 487 site. *Nucl Acids Res.* 1989 Dec; 17(24):10295–10305.
- 488 **Einav T**, Duque J, Phillips R. Theoretical analysis of inducer and operator binding for cyclic-AMP receptor protein 489 mutants. *PLoS ONE.* 2018; 13(9):e0204275.
- 490 **Foat B**, Morozov A, Bussemaker H. Statistical mechanical modeling of genome-wide transcription factor 491 occupancy data by MatrixREDUCE. *Bioinformatics.* 2006 Jul; 22(14):e141–9.
- 492 **Garcia HG**, Phillips R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci 493 USA.* 2011 Jul; 108(29):12173–12178.
- 494 **Gaston K**, Bell A, Kolb A, Buc H, Busby S. Stringent spacing requirements for transcription activation by CRP. 495 *Cell.* 1990 Aug; 62(4):733–743.
- 496 **Gertz J**, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. 497 *Nature.* 2009 Jan; 457(7226):215–218.
- 498 **Gunasekera A**, Ebright Y, Ebright R. DNA sequence determinants for binding of the Escherichia coli catabolite 499 gene activator protein. *J Biol Chem.* 1992 Jul; 267(21):14713–14720.
- 500 **Keseler IM**, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, Bonavides- 501 Martinez C, Paley S, Krummenacker M, Altman T, Kaipa P, Spaulding A, Pacheco J, Latendresse M, Fulcher 502 C, Sarker M, Shearer AG, Mackie A, Paulsen I, Gunsalus RP, et al. EcoCyc: a comprehensive database of 503 Escherichia coli biology. *Nucl Acids Res.* 2011; 39(Database issue):D583–90.
- 504 **Kinney JB**, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of 505 a transcriptional regulatory sequence. *Proc Natl Acad Sci USA.* 2010 May; 107(20):9158–9163.
- 506 **Kuhlman T**, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of Escherichia 507 coli. *Proc Natl Acad Sci USA.* 2007 Apr; 104(14):6043–6048.
- 508 **Kwasnieski JC**, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian 509 cis-regulatory element. *Proc Natl Acad Sci USA.* 2012 Nov; 109(47):19498–19503.
- 510 **Lederberg J**. The beta-d-galactosidase of Escherichia coli, strain K-12. *J Bacteriol.* 1950; 60(4):381–392.
- 511 **Lee DJ**, Minchin SD, Busby SJW. Activating transcription in bacteria. *Annu Rev Microbiol.* 2012; 66(1):125–152.
- 512 **Levo M**, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet.* 2014 Jul; 15(7):453–468.
- 513 **Malan T**, Kolb A, Buc H, McClure W. Mechanism of CRP-cAMP activation of lac operon transcription initiation 514 activation of the P1 promoter. *J Mol Biol.* 1984 Dec; 180(4):881–909.
- 515 **Markovitch O**, Agmon N. Structure and energetics of the hydronium hydration shells. *J Phys Chem A.* 2007 516 Mar; 111(12):2253–2256.
- 517 **McClure WR**, Hawley DK, Youderian P, Susskind MM. DNA determinants of promoter selectivity in Escherichia 518 coli. *Cold Spring Harb Symp Quant Biol.* 1983; 47 Pt 1:477–481.
- 519 **Melnikov A**, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnrke A, Callan CG, Kinney JB, Kellis M, 520 Lander ES, Mikkelsen TS. Systematic dissection and optimization of inducible enhancers in human cells using 521 a massively parallel reporter assay. *Nat Biotechnol.* 2012 Feb; 30(3):271–277.
- 522 **Miller J**. Experiments in Molecular Genetics. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 523 1972.
- 524 **Mogno I**, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of 525 binding site variants. *Genome Res.* 2013 Nov; 23(11):1908–1915.

- 526 **Morita T**, Shigesada K, Kimizuka F, Aiba H. Regulatory effect of a synthetic CRP recognition sequence placed
527 downstream of a promoter. *Nucl Acids Res.* 1988 Aug; 16(15):7315–7322.
- 528 **Neidhardt FC**, Bloch PL, Smith DF. Culture medium for enterobacteria. *J Bacteriol.* 1974 Sep; 119(3):736–747.
- 529 **Niu W**, Kim Y, Tau G, Heyduk T, Ebright RH. Transcription activation at class II CAP-dependent promoters: two
530 interactions between CAP and RNA polymerase. *Cell.* 1996 Dec; 87(6):1123–1134.
- 531 **Parkinson G**, Wilson C, Gunasekera A, Ebright YW, Ebright RH, Ebright RE, Berman HM. Structure of the CAP-DNA
532 complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J Mol Biol.* 1996 Jul;
533 260(3):395–408.
- 534 **Patwardhan RP**, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, Ahituv N,
535 Pennacchio LA, Shendure J. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat
Biotechnol.* 2012 Feb; 30(3):265–270.
- 537 **Pribnow D**. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad
Sci USA.* 1975 Mar; 72(3):784–788.
- 539 **Ptashne M**, Gann A. Genes and signals. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2002.
- 540 **Ptashne M**. Regulated recruitment and cooperativity in the design of biological regulatory systems. *Philos
Transact A Math Phys Eng Sci.* 2003 Jun; 361(1807):1223–1234.
- 542 **Reznikoff WS**. The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol.* 1992 Sep;
543 6(17):2419–2422.
- 544 **Rhodius VA**, West DM, Webster CL, Busby SJ, Savery NJ. Transcription activation at class II CRP-dependent
545 promoters: the role of different activating regions. *Nucl Acids Res.* 1997 Jan; 25(2):326–332.
- 546 **Roy S**, Garges S, Adhya S. Activation and repression of transcription by differential contact: two sides of a coin. *J
Biol Chem.* 1998 Jun; 273(23):14059–14062.
- 548 **Salgado H**, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-
549 Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K,
550 López-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martínez YI, Pannier L, Olvera
551 M, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold
552 standards and more. *Nucl Acids Res.* 2013 Jan; 41(Database issue):D203–13.
- 553 **Schmidt A**, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heine-
554 mann M. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol.* 2016 Jan;
555 34(1):104–110.
- 556 **Sharon E**, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. Infer-
557 ring gene regulatory logic from high-throughput measurements of thousands of systematically designed
558 promoters. *Nat Biotechnol.* 2012 May; 30(6):521–530.
- 559 **Shea MA**, Ackers GK. The OR control system of bacteriophage lambda. A physical-chemical model for gene
560 regulation. *J Mol Biol.* 1985 Jan; 181(2):211–230.
- 561 **Sherman MS**, Cohen BA. Thermodynamic state ensemble models of cis-regulation. *PLoS Comput Biol.* 2012;
562 8(3):e1002407.
- 563 **Smith RP**, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. Massively parallel
564 decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet.* 2013
565 Sep; 45(9):1021–1028.
- 566 **So Lh**, Ghosh A, Zong C, Sepúlveda LA, Segev R, Golding I. General properties of transcriptional time series in
567 *Escherichia coli*. *Nature Genetics.* 2011 Jun; 43(6):554–560.
- 568 **Spitz F**, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet.*
569 2012 Sep; 13(9):613–626.
- 570 **Thompson MG**, Sedaghatian N, Barajas JF, Wehrs M, Bailey CB, Kaplan N, Hillson NJ, Mukhopadhyay A, Keasling
571 JD. Isolation and characterization of novel mutations in the pSC101 origin that increase copy number. *Sci Rep.*
572 2018 Jan; 8(1):1590.

- 573 **Ushida C**, Aiba H. Helical phase dependent action of CRP: effect of the distance between the CRP site and the
574 -35 region on promoter activity. *Nucl Acids Res.* 1990 Nov; 18(21):6325–6330.
- 575 **Vilar JMG**, Leibler S. DNA looping and physical constraints on transcription regulation. *J Mol Biol.* 2003 Aug;
576 331(5):981–989.
- 577 **Weingarten-Gabbay S**, Nir R, Lubliner S, Sharon E, Kalma Y, Weinberger A, Segal E. Deciphering transcriptional
578 regulation of human core promoters. *bioRxiv*. 2017 Aug; .
- 579 **White MA**, Kwasnieski JC, Myers CA, Shen SQ, Corbo JC, Cohen BA. A Simple Grammar Defines Activating and
580 Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep.* 2016 Oct; 17(5):1247–1254.
- 581 **Wong MS**, Kinney JB, Krainer AR. Quantitative Activity Profile and Context Dependence of All Human 5' Splice
582 Sites. *Mol Cell.* 2018 Aug; 71(6):1012–1026.e3.

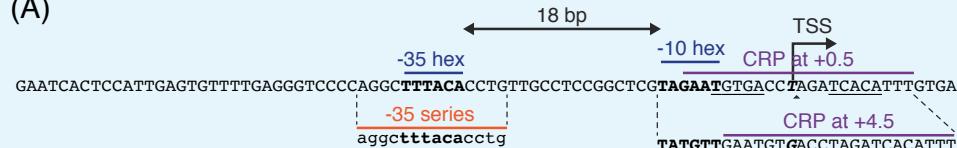
Appendix 1

583

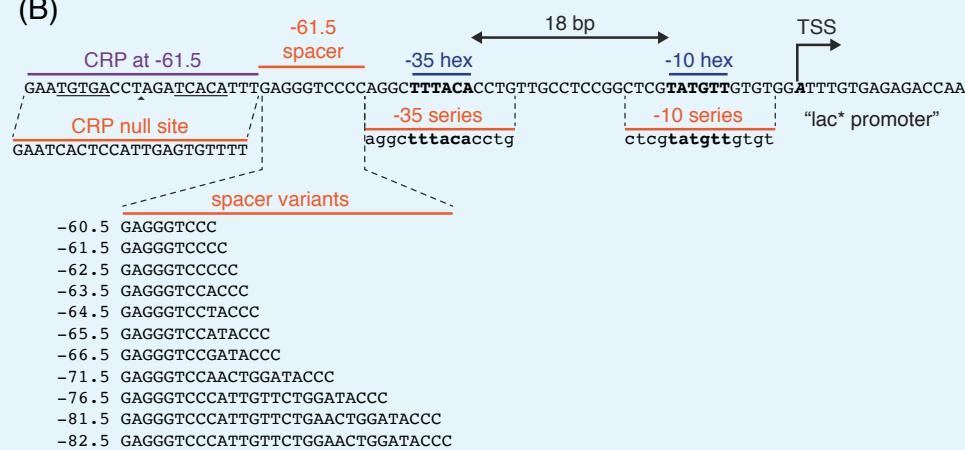
Media, Strains, Plasmids, and Promoters

584

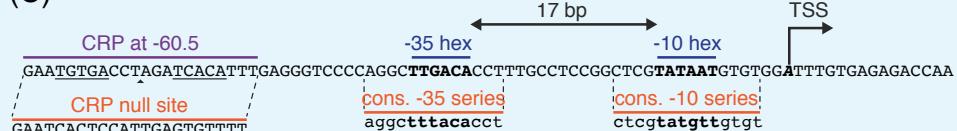
(A)



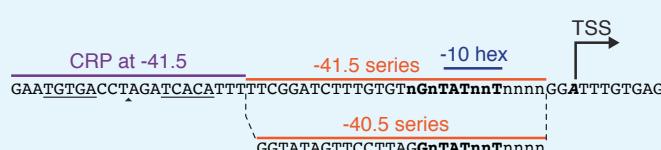
(B)



(C)



(D)



585

586

587

588

589

590

591

592

593

594

595

596

Appendix 1 Figure 1. Promoter sequences used in this study. In all panels, the -35 and -10 hexamers of the RNAP binding site are in bold. CRP binding site centers are indicated by small triangles. The palindromic pentamers of the core CRP binding site in each construct are underlined. The transcription start site (TSS) is bold and italicized. Lowercase bases ('a', 'c', 'g', and 't') indicate positions synthesized with a 24% mutation rate. The lowercase character 'n' indicates completely randomized positions. (A) Occlusion promoters assayed for Main Text Figure 2. (B) Class I promoters assayed for Main Text Figure 5. In the main text we refer to the wild-type promoter with CRP at -61.5 bp as the lac* promoter. The lac* promoter served as the template for all of the promoters shown here. (C) Strong class I promoters assayed for Main Text Figure 8. (D) Class II promoters assayed for Main Text Figure 9.

Expression measurements were performed on cells grown in rich defined media (RDM; purchased from Teknova) (Neidhardt *et al.*, 1974) supplemented with 10 mM NaHCO₃, 1 mM IPTG (Sigma), and 0.2% glucose. We refer to this media as RDM'. RDM' was further supplemented with 50 µg/ml kanamycin (Sigma) when growing cells, as well as 250 µM cAMP

(Sigma) when measuring t_+ .

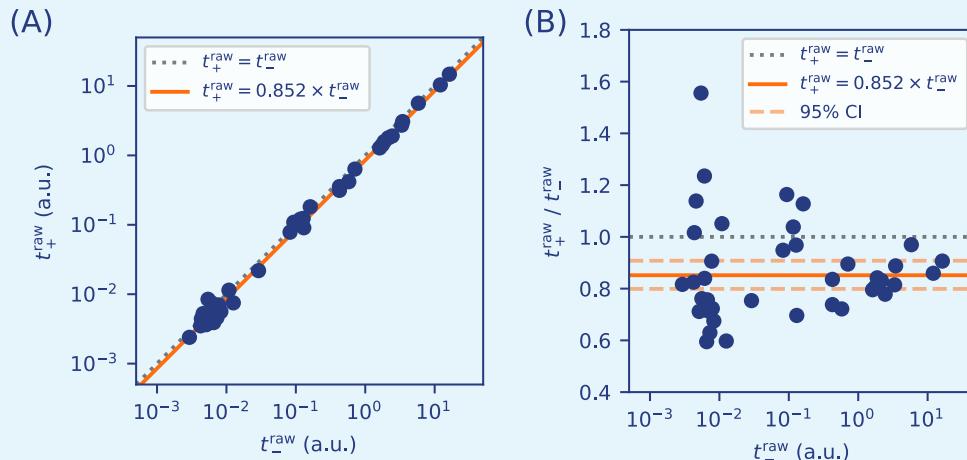
602 Expression measurements were performed in *E. coli* strain JK10, which has genotype $\Delta cyaA$
 603 $\Delta cpdA \Delta lacY \Delta lacZ \Delta dksA$. JK10 is derived from strain TK310 (**Kuhlman et al., 2007**), which
 604 is $\Delta cyaA \Delta cpdA \Delta lacY$. The $\Delta cyaA \Delta cpdA$ mutations prevent TK310 from synthesizing or
 605 degrading cAMP, thus allowing *in vivo* cAMP concentrations to be quantitatively controlled by
 606 adding cAMP to the growth media. Into TK310 we introduced the $\Delta lacZ$ mutation, yielding
 607 strain DJ33; this mutation enables the use of β -galactosidase activity assays for measuring
 608 plasmid-based *lacZ* expression. In our initial experiments, we found that the growth rate of
 609 DJ33 in RDM' varied strongly with the amount of cAMP added to the media. Fortunately, we
 610 isolated a spontaneous knock-out mutation in *dksA* (thus yielding JK10), which caused the
 611 growth rate (~ 30 min doubling time) in RDM' to be independent of cAMP concentrations
 612 below ~ 500 μM .^a The TK310, DJ33, and JK10 genotypes were confirmed by whole genome
 613 sequencing using the PureLink Genomic DNA Mini Kit (ThermoFisher) for extracting genomic
 614 DNA from cultured cells and the Nextera XT DNA Library Preparation Kit (Illumina) for
 615 preparing whole-genome sequencing libraries.

616 Expression of the *lacZ* gene was driven from variants of a plasmid we call pJK48. These
 617 reporter constructs were cloned as follows. We started with the vector pJK14 from **Kinney**
 618 **et al. (2010)**. pJK14 contains a pSC101 origin of replication (~ 5 copies per cell; **Thompson**
 619 **et al. (2018)**), a kanamycin resistance gene, and a *ccdB* cloning cassette positioned immedi-
 620 ately upstream of a *gfpmut2* reporter gene and flanked by outward-facing BsmBI restriction
 621 sites. First, the *gfpmut2* gene in this vector was replaced with *lacZ*, yielding pJK47. Next, the
 622 ribosome binding site in the 5' UTR of *lacZ* was weakened, yielding pJK47.419; this weakening
 623 prevents *lacZ* expression from substantially slowing cell growth in RDM'. pJK47.419 was
 624 propagated in DB3.1 *E. coli* (Invitrogen), which is resistant to the CcdB toxin. The promoters
 625 we assayed were variants of what we call the "lac*" promoter. The lac* promoter is similar
 626 to the endogenous *lac* promoter of *E. coli* MG1655 except for (i) it contains a CRP binding
 627 site with a consensus right pentamer and (ii) it contains mutations that were introduced
 628 in an effort to remove previously reported cryptic promoters (**Reznikoff, 1992**). Promoter-
 629 containing insertion cassettes were created through overlap-extension PCR and flanked by
 630 outward-facing BsaI restriction sites. All primers were ordered from Integrated DNA Tech-
 631 nologies. Note that some of the primers used to create these inserts were synthesized using
 632 pre-mixed phosphoramidites at specified positions; this is how a 24% mutation rate in the
 633 -10 or -35 regions of the RNAP binding site was achieved. The resulting promoter sequences
 634 are illustrated in Appendix 1 Figure 1. To clone variants of pJK48, we separately digested the
 635 pJK47.419 vector with BsmBI (NEB) and the appropriate insert with BsaI (NEB). Digests were
 636 then cleaned up (Qiagen PCR purification kit) and ligated together in at 1:1 molar ratio for
 637 1 hour using T4 DNA ligase (Invitrogen). After 90 min dialysis, plasmids were transformed
 638 into electrocompetent JK10 cells. Individual clones were plated on LB supplemented with
 639 kanamycin (50 $\mu\text{g}/\text{ml}$). After initial cloning and plating, each colony was re-streaked, grown
 640 in LB+kan, and stored as a catalogued glycerol stock. The promoter region of each clone
 641 was sequenced in both directions. Only plasmids with validated promoter sequences were
 642 used for the measurements presented in this paper. The promoter sequences of all 448
 643 plasmids used in this study, as well as their measured t_+ and t_- values, are provided at
https://github.com/jbkinney/17_inducibility.

^aWe note that JK10 will not grow in minimal media in the absence of cAMP.

645 Appendix 2

646 Miller assays and the calibration of expression measurements



647
648
649 **Appendix 2 Figure 1.** Calibration of expression measurements with and without cAMP. (A)
650 Measurements of t_+^{raw} (in 250 μM cAMP) vs t_-^{raw} (in 0 μM cAMP) for promoters in which the CRP binding
651 site has been replaced by a non-functional “null” site. As expected, these data lie close to the $t_+^{\text{raw}} = t_-^{\text{raw}}$
652 diagonal (dotted line). (B) Upon closer inspection, however, we found that t_+^{raw} values consistently fall
653 slightly below corresponding t_-^{raw} values. Using least-squares fitting we found that, on average,
654 $t_+ / t_-^{\text{raw}} = 0.852^{+0.056}_{-0.053}$ where uncertainties indicate a 95% confidence interval (reflecting 1.96 times the
655 standard error of the mean in log space). To correct for this bias, we plot and fit models to $t_+ = t_-^{\text{raw}}$ and
656 $t_- = 0.855 \times t_+^{\text{raw}}$ throughout this paper.
657

658 We obtained t_+ and t_- measurements for each promoter as follows. First, the corresponding
659 *E. coli* clone was streaked out on LB+kan agar and grown overnight. A colony was then picked
660 and used to inoculate a 1.5 ml overnight LB+kan liquid culture. Either 8 μl , 6 μl , or 4 μl of
661 the overnight culture were then diluted into 200 μl RDM'+kan. 25 μl of each dilution was
662 then added to 175 μl RDM'+kan in a 96-well optical bottom plate and supplemented with
663 either 0 μM cAMP (for t_-^{raw}), 250 μM cAMP (for t_+^{raw}), or another cAMP concentration (for some
664 t_+^{raw} measurements in Figure 3). The plate was then covered with Breathe-Easier film (USA
665 Scientific) and cells were cultured for ~ 3 hr at 37°C, shaking at 900 RPM in a microplate
666 shaker. During this time, 5.5 ml of lysis buffer was freshly prepared using 1.5 ml RDM', 4.0
667 ml PopCulture reagent (Millipore), 114 μl of 35 mg/ml chloramphenicol (Sigma), and 44 μl of
668 40 U/ μl rLysozyme (Sigma).

669 Microplate film was removed and cell density (quantified by A_{600}) was measured using an
670 Epoch 2 Microplate Spectrophotometer (BioTek). Cells were then lysed by adding 25 μl
671 lysis buffer to each microplate well, incubating the microplate at room temperature for 10
672 minutes without shaking, then cooling the microplate at 4°C for a minimum of 15 minutes.
673 In each well of a 96-well optical bottom plate, 50 μl of lysate was then added to 50 μl of
674 pre-chilled Z-buffer containing 1 mg/ml ONPG (Sigma). Samples were sealed with optical
675 film and both A_{420} and A_{550} were periodically measured in the plate reader over an extended
676 period of time (every 1.5 min for 1 hour or every 15 min for 10 hours, depending on the level
677 of expression expected).

678 The raw expression levels were quantified from these absorbance data using the formula

$$t_{\pm}^{\text{raw}} = \frac{\Delta A_{420} - \Delta A_{550}}{V \cdot \Delta T \cdot A_{600}}, \quad (9)$$

where $V = 50$ is the volume of lysate in μl added to the ONPG reaction, ΔT is the change in time from the beginning of the measurement, and ΔA_X indicates a change in absorbance at $X \text{ nm}$ over this time interval. Only data from wells with $A_{600} \lesssim 0.5$ were analyzed. Note that the A_{550} term in Equation 9 is not multiplied by 1.75 as it is in **Miller (1972)**. This is because our A_{550} measurements are used to compensate for condensation on the microplate film, not cellular debris as in **Miller (1972)**; our lysis procedure produces no detectable cellular debris. In practice, Equation 9 was not evaluated using individual measurements, but was computed from the slope of a line fit to all of the non-saturated absorbance measurements. Raw A_{420} , A_{550} , and A_{600} values, as well as our analysis scripts, are available at https://github.com/jbkinney/17_inducibility. Median values from at least 3 independent Miller measurements (and often more) were used to define each measurement shown in Main Text figures.

Because we controlled the *in vivo* activity of CRP by supplementing media with or without cAMP, we tested whether CRP-independent promoters produce measurements that vary between these growth conditions. Specifically, we measured t_-^{raw} (in 0 μM cAMP) and t_+^{raw} (in 250 μM cAMP) for 39 promoters in which the CRP binding site was replaced with a “null” site (see Appendix 1, Figures 1B and 1C). These measurements are plotted in Figure 1 of this Appendix, and show a slight bias. To correct for this bias, we use an unadjusted $t_+ = t_+^{\text{raw}}$ together with an adjusted $t_- = 0.855 \times t_-^{\text{raw}}$ throughout the main text. Note that $t_+ = t_+^{\text{raw}}$ was used for all nonzero cAMP concentrations, including those in Main Text Figure 3B that differ from 250 μM . Some upward bias is therefore possible in these t_+ measurements, but we do not expect this to greatly affect our conclusions.

704 **Appendix 3**

706
707
708
709
710
711
712
713
Parameter inference

714 Allelic manifold parameters were fit to measured t_+ and t_- values as follows. First, outlier
715 data points were called by eye and excluded from the parameter fitting procedure. We
716 denote the remaining measurements using $t_+^{i,\text{data}}$ and $t_-^{i,\text{data}}$, where $i = 1, 2, \dots, n$ indexes the n
717 non-outlier data points. Corresponding model predictions $t_+^i(\theta)$ and $t_-^i(\theta)$, where θ denotes
718 model parameters, were then fit to these data using nonlinear least squares optimization.
719 Specifically, we inferred parameters $\theta^* = \operatorname{argmin}_\theta \mathcal{L}(\theta)$ where the loss function is given by
720

$$\mathcal{L}(\theta) = \sum_{i=1}^n \left(\left[\log \frac{t_+^i(\theta)}{t_+^{i,\text{data}}} \right]^2 + \left[\log \frac{t_-^i(\theta)}{t_-^{i,\text{data}}} \right]^2 \right). \quad (10)$$

721 These optimal parameter values θ^* were used to generate the best-estimate allelic mani-
722 folds, which are plotted in black in Main Text figures. Uncertainties in θ were estimated by
723 performing the same inference procedure on bootstrap-resampled data. For each variable
724 $X \in \{F, P, \alpha', \beta', t_{\text{sat}}, t_{\text{bg}}\}$, we report
725

$$X = (X_{50})^{+(X_{84}-X_{50})}_{-(X_{50}-X_{16})} \quad (11)$$

726 where X_{50} , X_{84} , and X_{16} respectively denote the median, 84th percentile, and 16th percentile
727 of X values obtained from bootstrap resampling. In the case of $X \in \{F, P, \alpha\}$, we also report
728

$$\Delta G_X = -k_B T \log X_{50} \pm k_B T \left(\frac{\log X_{84} - \log X_{16}}{2} \right), \quad (12)$$

729 where $1 \text{ kcal/mol} = 1.62 k_B T$, corresponding to 37°C . We now describe each specific inference
730 procedure in more detail.

731 **Inference for Main Text Figure 2B**

732 We inferred $\theta = \{t_{\text{sat}}, t_{\text{bg}}, F, P_1, P_2, \dots, P_n\}$, with model predictions given by
733

$$t_+^i(\theta) = t_{\text{sat}} \frac{P_i}{1 + F + P_i} + t_{\text{bg}}, \quad t_-^i(\theta) = t_{\text{sat}} \frac{P_i}{1 + P_i} + t_{\text{bg}}. \quad (13)$$

734 Parameters were fit to the $n = 39$ non-outlier measurements made for promoters with $+0.5$
735 bp or $+4.5$ bp architecture. We found that $F = 23.9_{-2.5}^{+3.1}$ and $t_{\text{bg}} = 2.30 \times 10^{-3}$ a.u., while t_{sat}
736 values remained highly uncertain.

737 **Inference for Main Text Figure 3B**

738 We performed a separate inference procedure for each of the seven cAMP concentra-
739 tions $C \in \{250, 125, 50, 25, 10, 5, 2.5\}$, indicated in μM units. Specifically, we inferred $\theta_C = \{F_C, P_1, P_2, \dots, P_{n_C}\}$ where n_C is the number of promoters for which t_+ was measured using
740 cAMP concentration C . Model predictions were given by
741

$$t_+^i(\theta_C) = t_{\text{sat}} \frac{P_i}{1 + F_C + P_i} + t_{\text{bg}}, \quad t_-^i(\theta_C) = t_{\text{sat}} \frac{P_i}{1 + P_i} + t_{\text{bg}}, \quad (14)$$

742 where $t_{\text{sat}} = 15.1$ a.u. is the median saturated transcription rate from Main Text Figure 5C,
743 and $t_{\text{bg}} = 2.30 \times 10^{-3}$ a.u. is the median background transcription rate from Main Text Fig. 2B.
744 Note that many of the t_-^i measurements were used in the inference procedures for multiple
745 values of C , whereas each t_+^i measurement was used in only one such inference procedure.
746

750 **Inference for Main Text Figure 5B**
 751 Using data from both the -10 and -35 allelic series for the -61.5 bp promoter architecture, we
 752 inferred $\theta = \{t_{\text{sat}}, t_{\text{bg}}, \alpha', P_1, \dots, P_n\}$. Model predictions were given by
 753

$$754 \quad t_+^i(\theta) = t_{\text{sat}} \frac{\alpha' P_i}{1 + \alpha' P_i} + t_{\text{bg}}, \quad t_-^i(\theta) = t_{\text{sat}} \frac{P_i}{1 + P_i} + t_{\text{bg}}. \quad (15)$$

755 For each inferred α' , a value for α was computed using $\alpha = \alpha'(1 + F^{-1}) - F^{-1}$, where $F = 23.9$
 756 is the median CRP binding factor inferred for Main Text Figure 2B.
 757

758 **Inference for Main Text Figure 5C**
 759 In a single fitting procedure, we inferred $\theta = \{t_{\text{sat}}, t_{\text{bg}}^{-82.5}, \dots, t_{\text{bg}}^{-60.5}, \alpha'_{-82.5}, \dots, \alpha'_{-60.5}, P_1, \dots, P_n\}$
 760 using
 761

$$762 \quad t_+^i(\theta) = t_{\text{sat}} \frac{\alpha'_{D_i} P_i}{1 + \alpha'_{D_i} P_i} + t_{\text{bg}}^{D_i}, \quad t_-^i(\theta) = t_{\text{sat}} \frac{P_i}{1 + P_i} + t_{\text{bg}}^{D_i}. \quad (16)$$

763 where each $D_i \in \{-82.5, -81.5, -76.5, -72.5, -71.5, -66.5, -65.5, -64.5, -63.5, -62.5, -61.5, -60.5\}$
 764 represents the position of the CRP binding site (in bp relative to the TSS) for promoter i . Note
 765 that a single value for t_{sat} was inferred for all promoter architectures, while both t_{bg}^D and α'_D
 766 varied with CRP position D . The corresponding values of α plotted in Main Text Figure 5D
 767 and listed in Main Text Table 1 were computed using $\alpha_D = \alpha'_D(1 + F^{-1}) - F^{-1}$ where $F = 23.9$
 768 is the median CRP binding factor inferred for Main Text Figure 2B. Among other results, we
 769 find that $t_{\text{sat}} = 15.1_{-0.5}^{+0.6}$ a.u..
 770

771 **Inference for Main Text Figure 8C**
 772 For each spacing D , we separately inferred $\theta_D = \{\alpha'_D, \beta'_D, t_{\text{bg}}^D\}$ using
 773

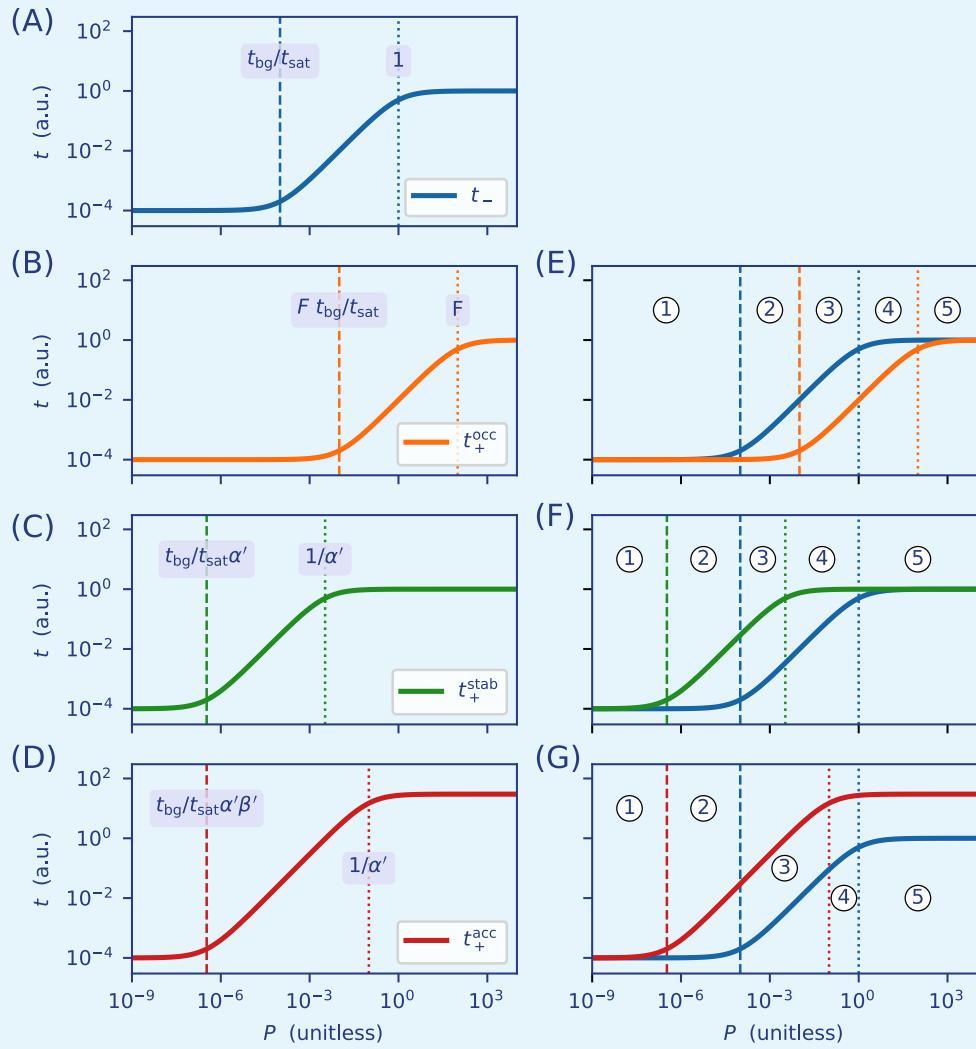
$$774 \quad t_+^i(\theta_D) = \beta'_D t_{\text{sat}} \frac{\alpha'_D P_i}{1 + \alpha'_D P_i} + t_{\text{bg}}^D, \quad t_-^i(\theta_D) = t_{\text{sat}} \frac{P_i}{1 + P_i} + t_{\text{bg}}^D. \quad (17)$$

775 where $t_{\text{sat}} = 15.1$ a.u. is the median saturated transcription rate inferred for Main Text Figure
 776 5C. We then computed $\alpha_D = \alpha'_D(1 + F^{-1}) - F^{-1}$ and $\beta_D = \beta'_D(1 + \alpha_D^{-1} F^{-1}) - \alpha_D^{-1} F^{-1}$, using the
 777 median CRP binding factor $F = 23.9$ inferred for Main Text Figure 2B.
 778

779 Appendix 4

780

Derivation of allelic manifold regimes



781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

Appendix 4 Figure 1. Derivation of the regimes of allelic manifolds. Panels A-D show simulated induction curves for transcription t as a function of the RNAP binding factor P . Dashed lines indicate boundaries between the minimal and linear regimes of each curve, while dotted lines indicate boundaries between linear and maximal regimes. A formula for the value of P at each regime boundary is also shown. All simulations used $t_{sat} = 1$ a.u., $t_{bg} = 10^{-4}$ a.u., $F = 100$, and P ranging from 10^{-9} to 10^4 . (A) Induction curve for unregulated transcription; see Equation 18. (B) Induction curve for transcription repressed by occlusion; see Equation 19. (C) Induction curve for transcription activated by stabilization ($\alpha = 300$); see Equation 20. (D) Induction curve for transcription activated by acceleration ($\alpha = 10$, $\beta = 30$); see Equation 21. Panels E-G show how overlaps between the six regimes of two induction curves (three for t_- and three for t_+) result in five distinct regimes for the corresponding allelic manifold. (E) Regimes of the allelic manifold for occlusion, which is shown in Fig. 1C. (F) Regimes of the allelic manifold for stabilization, which is shown in Fig. 4C. (G) Regimes of the allelic manifold for acceleration, which is shown in Fig. 7C.

Each transcription rate modeled in this work is a sigmoidal function of the unitless RNAP-

797
 798 DNA binding factor P . As such, a log-log plot of transcription t as a function of P reveals
 799 a sigmoidal curve having three distinct regimes. The "minimal" regime of this induction
 800 curve comprises values of P that are sufficiently small for t to be well-approximated by
 801 its smallest value (t_{bg} in all cases). The "maximal" regime occurs when P is so large that t
 802 is well-approximated by its largest value (either t_{sat} or $\beta' t_{\text{sat}}$). Between these maximal and
 803 minimal regimes lies a "linear" regime in which t is approximately proportional to P .

For unregulated transcription, which in this paper is denoted t_- , these three regimes are given by,

$$t_- = t_{\text{sat}} \frac{P}{1 + P} + t_{\text{bg}} \approx \begin{cases} t_{\text{bg}} & \text{for } P \ll \frac{t_{\text{bg}}}{t_{\text{sat}}} \\ t_{\text{sat}} P & \text{for } \frac{t_{\text{bg}}}{t_{\text{sat}}} \ll P \ll 1 \\ t_{\text{sat}} & \text{for } 1 \ll P \end{cases} \quad (18)$$

See Figure 1A. For transcription that is repressed by occlusion (with $F \gg 1$), which we denote here by t_+^{occ} , these three regimes are shifted (relative to t_-) to larger values of P by a factor of approximately F . As a result,

$$t_+^{\text{occ}} = t_{\text{sat}} \frac{P}{1 + F + P} + t_{\text{bg}} \approx \begin{cases} t_{\text{bg}} & \text{for } P \ll F \frac{t_{\text{bg}}}{t_{\text{sat}}} \\ t_{\text{sat}} \frac{P}{1+F} & \text{for } F \frac{t_{\text{bg}}}{t_{\text{sat}}} \ll P \ll F \\ t_{\text{sat}} & \text{for } F \ll P \end{cases} \quad (19)$$

See Figure 1B. By contrast, for transcription that is activated by stabilization, denoted here by t_+^{stab} , these three regimes shift (relative to t_-) to lower values of P by a factor of $1/\alpha'$, giving

$$t_+^{\text{stab}} = t_{\text{sat}} \frac{\alpha' P}{1 + \alpha' P} + t_{\text{bg}} \approx \begin{cases} t_{\text{bg}} & \text{for } P \ll \frac{t_{\text{bg}}}{t_{\text{sat}} \alpha'} \\ t_{\text{sat}} \alpha' P & \text{for } \frac{t_{\text{bg}}}{t_{\text{sat}} \alpha'} \ll P \ll \frac{1}{\alpha'} \\ t_{\text{sat}} & \text{for } \frac{1}{\alpha'} \ll P \end{cases} \quad (20)$$

See Figure 1C. For transcription that is activated partially by acceleration and partially by stabilization, here denoted by t_+^{acc} , two parameters govern the shape of the induction curve. As a result, the boundary between the minimal and linear regimes are shifted (relative to t_-) to lower values of P by a factor of $1/\alpha' \beta'$, while the boundary between the linear regime and the maximal regime is shifted down by a factor of only $1/\alpha'$. As a result,

$$t_+^{\text{acc}} = \beta' t_{\text{sat}} \frac{\alpha' P}{1 + \alpha' P} + t_{\text{bg}} \approx \begin{cases} t_{\text{bg}} & \text{for } P \ll \frac{t_{\text{bg}}}{t_{\text{sat}} \alpha' \beta'} \\ t_{\text{sat}} \alpha' \beta' P & \text{for } \frac{t_{\text{bg}}}{t_{\text{sat}} \alpha' \beta'} \ll P \ll \frac{1}{\alpha'} \\ t_{\text{sat}} \beta' & \text{for } \frac{1}{\alpha'} \ll P \end{cases} \quad (21)$$

See Figure 1D.

Each allelic manifold described in the main text has five distinct regimes. These arise from overlaps between the three regimes of t_- and the three regimes of t_+ . Specifically, the five regimes of the allelic manifold for repression by occlusion, which are listed in Main Text Figure 1D, arise from the overlaps between the three regimes for t_- and the three regimes for t_+^{occ} . These overlaps are indicated in Figure 1E. Similarly, the five regimes of the allelic manifold for activation by stabilization (Main Text Figure 4D) arise from the overlaps between

the regimes of t_- and t_+^{stab} , illustrated in Figure 1F, while the regimes of the manifold for activation by acceleration (Main Text Figure 7D) arise from overlaps between the regimes of t_- and t_+^{acc} , illustrated in Figure 1G.