Supplemental Information Appendix

A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria.

Nathan M. Belliveau[1], Stephanie L. Barnes[1], William T. Ireland[2], Daniel L. Jones[3], Mike J. Sweredoski[4], Annie Moradian[4], Sonja Hess[4,5], Justin B. Kinney[6], Rob Phillips[1,2,7,*]

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

[2]Department of Physics, California Institute of Technology, Pasadena, CA, 91125, United States;

[3]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden;

[4]Proteome Exploration Laboratory (PEL), Beckman Institute, California Institute of Technology, Pasadena, CA, 91125, United States;

[5]Current address: MedImmune, One Medimmune Way, Gaithersburg, MD, 20878, United States;

[6]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, United States;

[7]Department of Applied Physics, California Institute of Technology, Pasadena, CA, 91125, United States

* Corresponding author. Email address: phillips@pboc.caltech.edu

Contents

A Identification of unannotated promoters in *E. coli* with growth-dependent differential expression.

Here we briefly describe how the unannotated promoters of the main text (*purT*, *xylE*, and *dgoR*) were chosen. Fig. S1 summarizes the current state of regulatory knownledge in *E. coli* and those promoters considered in this work. Here, we parse the database RegulonDB that lists all known regulatory reatures in *E. coli*, with the striking finding that more than half the operons lack any annotated transcription factor binding sites (denoted by red lines). To identify candidate promoters with which to apply Sort-Seq, we made use of a variety of genome-wide datasets[1,2]. Specifically, in the case of the *purT* promoter, network inference approaches[2] led us to a number of unannotated genes that appeared to be sensitive to purine (others included *yieH* and *adeP*). Since the *purT* promoter lacked any experimental characterization and with ChIP-chip data suggesting PurR may be involved[3], it appeared to be a good starting point with which to apply our approach. The promoters of *xylE* and *dgoR*, were identified from a recent study by Schmidt *et al.*[1]. They used mass spectrometry and measured the copy number per cell of more than 2,300 proteins (about 55% of the *E. coli* proteome) across 22 growth conditions. These conditions included different carbon sources, temperature, pH, growth phase, media, and growth in chemostats. This provided us with a rich set of measurements with which to identify unannotated promoters where a particular growth condition influenced expression and may be under transcriptional regulation. The rest of this section describes how the data of Schmidt *et al.* was used to identify candidate promoters.

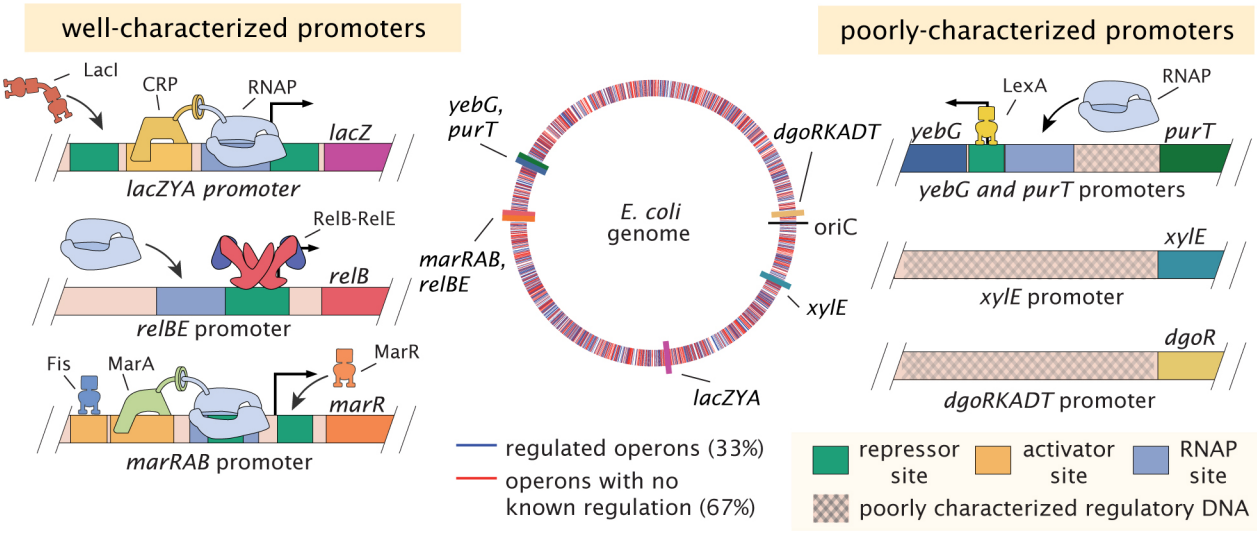

Figure S1. Summary of transcriptional regulatory knowledge in *E. coli*. left panel: Well-characterized promoters considered in this work. The schematics highlight the known regulatory architectures for the annotated promoters of *marRAB*, *relBE*, and *lacZYA*. The center plot identifies the genomic location of different operons in *E. coli*. Operons with annotated TF binding sites are shown in blue, while those lacking regulatory descriptions are shown in red[4]. The genomic location of the promoters considered in this work are labeled. Right panel: promoters associated with the operons of *yebG* and the poorly-characterized operons *purT*, *xylE*, and *dgoRKADT*. The promoters of *yebG* and *purT* are oriented in opposite directions. Repressor binding sites are shown in green, activator binding sites in yellow, and RNA polymerase (RNAP) binding sites in blue. The poorly characterized regulatory DNA is noted by a hashed pattern. The identification of regulated operons was performed using the annotated operons listed on RegulonDB[4], which are based on manually curated experimental and computational data. An operon was considered to be regulated if it had at least one transcription factor binding site associated with it.

In order to identify candidate genes using the mass spectrometry copy number data, we ranked each protein based on its copy number in a particular growth condition, divided by the average copy number across the 22 conditions. Regulated proteins should be among those that exhibit a large change in copy number in one or a few growth conditions. As a confirmation of this, among the proteins with known regulation, we came across the GalE protein which was found to have significantly higher expression when cells are grown in galactose (Fig. S2A). GalE is involved in galactose catabolism, and its expression is known to increase due to loss of repression of the *galE* promoter when cells were grown in galactose[5,6]. Among promoters that did not have any annotated regulation, we show the expression of DgoD for several different carbon sources (Fig. S2B). Cells grown in galactose showed much higher expression of the DgoD gene, with about 675 copies per cell, compared to at most 15 copies per cell across the other growth conditions. This is only one of many examples where a protein showed a large differential expression level across growth conditions (which include *xylE*), and suggests that many of these unannotated promoters may actually be under regulation.

Another way to view their data set is to calculate the coefficient of variation (the ratio of the standard deviation to the mean protein copy number) for each gene across the 22 growth conditions. In Fig. S2C, the coefficient of variation is plotted for each of the proteins measured in this study, separated by whether their promoter contains any known transcription factor binding sites (identified from RegulonDB[4]). For GalE, whose expression was perturbed by growth in galactose, we find a calculated coefficient of variation of 1.18. Using this as our reference for a regulated gene that was perturbed in the study, there appear to be many unannotated genes that show similar or larger coefficient of variations, further suggesting candidates that may be under regulation. Among these, DgoD for example has a coefficient of variation of 3.64. Among the other proteins we investigated, XylE also has a high coefficient of variation, equal to 2.73, and shows almost no expression unless cells are grown in the presence of xylose as the carbon source. While we only pursued the promoters associated with expression of DgoR, DgoD, DgoK, DgoA, and XylE, there are many other unannotated promoters that will be of interest in future work.

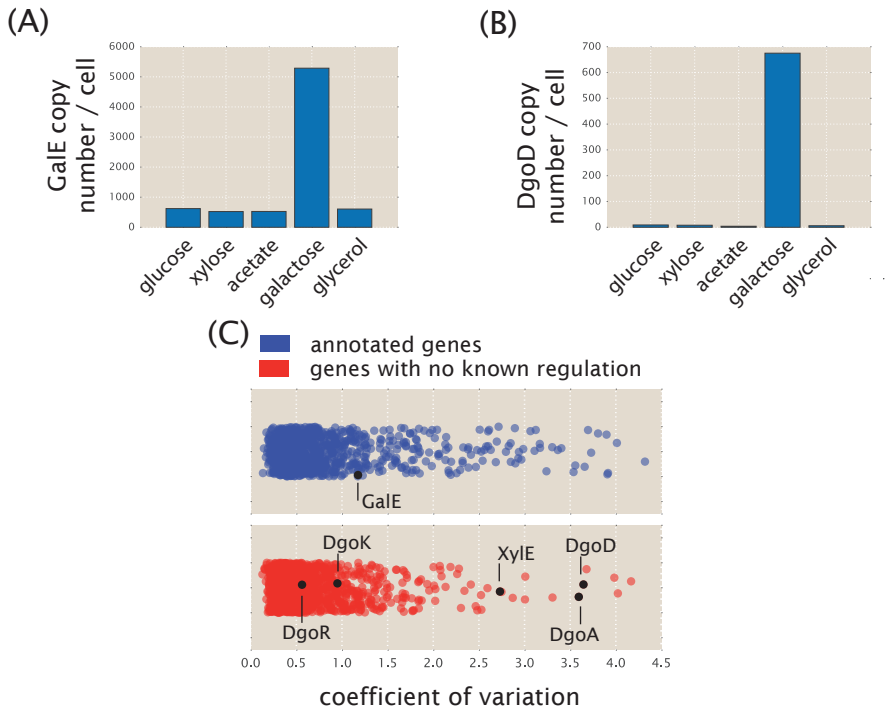

Figure S2. Identification of unannotated genes with potential regulation and distribution of known transcription factor binding sites in *E. coli.* (A) Here we show the protein copy numbers per cell for GalE across several carbon sources. Expression was sensitive to the presence of galactose which is consistent with its known regulation (with about 5000 copies per cell, versus about 500 for most other growth conditions). (B) DgoD was also found to be sensitive to the presence of galactose as the carbon source. The copy number was measured to be 675 copies per cell when cells were grown in galactose, and 15 copies per cell or less in all other conditions considered. For both (A) and (B), values are shown for growth in M9 minimal media, with glucose, xylose, acetate, galactose, and glycerol as carbon sources and obtained from[1]. (C) Coefficient of variation (standard deviation divided by mean copy number) across the 22 growth conditions for each protein measured in[1]. Proteins are identified as either having regulatory annotation (blue) or not (red) using the annotations in RegulonDB[4]. GalE is noted among the annotated genes and provides a reference as a gene that is known to be regulated and be perturbed in this study, as shown in (A).

B Characterization of library diversity and sorting sensitivity.

B.1 Sort-Seq of the *rel* promoter using different sorting conditions.

In the work of the main text, Sort-Seq was performed by sorting cell libraries into four bins based on their fluorescence, each containing about 15 percent of the population. The remaining population was not collected and was discarded to waste. Due to the variability in expression of a single clonal population (Fig. S3A), sorting into a larger number of narrower bins was not expected to provide additional resolution for the sequence-dependent fluorescence distribution. Given the success in identifying the known regulatory binding sites of the *lacZ*, *relB*, and *marR* promoters, and agreement between the inferred sequences logos and available sequence logos (see Supplemental Fig. S4), these conditions appeared to provide sufficient information to accurately analyze our libraries.

However, as an additional check that our results were not being influenced by the specific sorting

scheme, we also tested several other sorting conditions using our *relB* promoter library. Here cells were sorted into either four or eight bins, with a sorting gate containing between 10 and 22 percent of the population per bin. The associated expression shift plots and information footprints (defined in Supplemental Section H) are shown in Fig. S3B-D. In general we found little difference between each of these experiments. Energy matrices for the binding sites were similarly in agreement, with a Pearson correlation coefficient between matrix parameters generally greater than 0.9 across the different conditions tested.

B.2 Analysis of library diversity using data from the *mar* promoter.

Here we provide additional characterization of the mutagenized promoter libraries, using a library from the *marR* promoter as a representative example (70 bp region containing RNAP and MarR repressor sites). With the exception of the *lacZ* promoter, all library oligonucleotide pools were purchased from Integrated DNA Technologies (USA) with a target mutation rate of nine percent per nucleotide position. For the *lacZ* promoter library, we purchased an oligonucleotide pool using their Ultramer branded technology to allow for a longer mutagenized region that covered the known set of regulatory binding sites. While we intended to have a similar mutation rate, through sequencing we found a mutation rate closer to three percent per nucleotide position. While unexpected, it provided a test of two different mutation rates in our initial validation of the methodology using well-characterized promoters.

To get a better sense of how the mutation rate varies across the libraries, we plot a histogram of the number of mutations per base pair for the entire set of sequences found in the *marR* promoter library (Fig. S3E). We obtained an average mutation rate of 10.4% in this library, close to our target rate of 9%, though there is some variability in this mutation rate as might be expected given that the incorporation of mutations in the DNA synthesis procedure is a random process. Since we are using these sequence data sets to infer sequence-specific models of binding between DNA and transcription factors, it was also of interest to consider the mutational coverage found within the library. As shown in Fig. S3F, all single-point mutations and a large fraction of two-point mutations were present within the library. Due to the large number of possible three point mutants in a 60 bp region, only a small subset of the possible sequences will be found in the library.

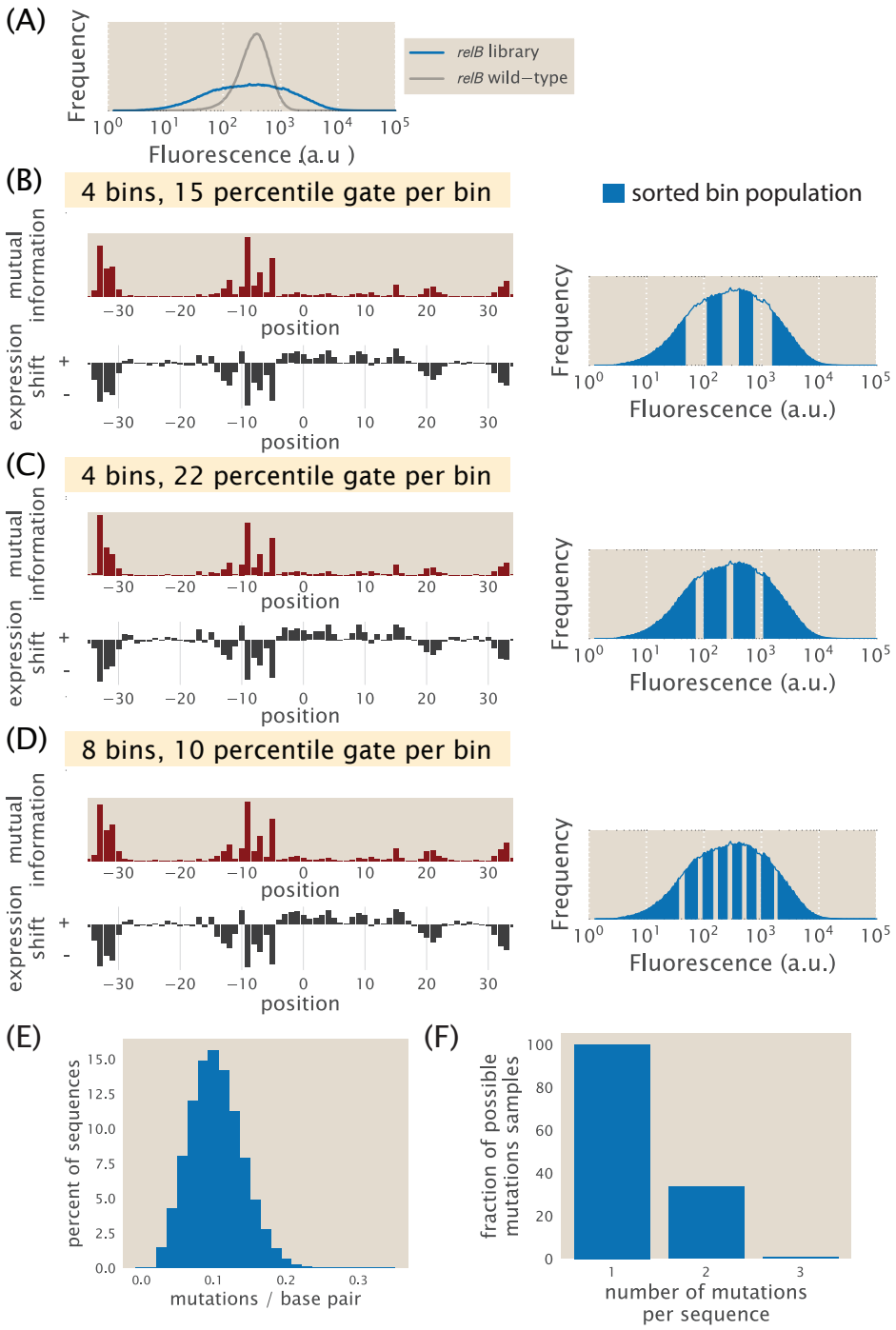

Figure S3. Analysis of the library mutation spectrum and effect of Sort-Seq sorting conditions. (A) Here we used our *relBE* promoter library to test whether the sorting procedure influenced our Sort-Seq data analysis. The fluorescence histogram of the wild-type promoter plasmid (single clonal population) and the mutated library for the *relB* promoter are shown. Expression shifts and information footprints are shown for cells sorted under three different scenarios in (B) -(D). In (B) cells were sorted using the approach of the main text where cells were sorted into 4 bins, each containing 15% of the population. *(continued on next page)*

Figure S3. *(continued from previous page)* In (C) cells were similarly sorted into 4 bins, but where each bin contained about 22% of the population. In (D) cells were sorted into 8 bins, each containing about 10% of the population. The histograms beside each information footprint identify the approximate gating windows used to sort each fluorescence bin population. Histograms were based on between 400,000-500,000 cell counts. The same cell culture was used for each of the three Sort-Seq experiments performed here, sorted during the same sorting session. Cells were grown in M9 minimal media with 0.5% glucose like in the main text. (E) Histogram showing the mutation rate across all sequences found in the 60 bp *marRAB* library containing the RNAP and MarR repressor binding sites. Analysis was based on sequences from all fluorescence sorted bins. (F) The fraction of all possible unique sequences with one, two, or three mutations is shown for the *marRAB* library of (E). The coverage quickly drops for possible three-point mutations due to the large sequence space at this mutation frequency.

C Generation of sequence logos.

Sequence logos provide a simple way to visualize the sequence specificity of a transcription factor to DNA, as well as the amount of information present at each position[7]. Here we describe how we generate them using either known genomic binding sites or the energy matrices from our Sort-Seq data. In each case we need to calculate a 4xL position weight matrix for a binding site of length L, which is used to estimate the position-dependent information content that will then be used to construct a sequence logo. In Section C.1 we consider position weight matrices from known genomic binding sites, while in Section C.2 we consider position weight matrices from our Sort-Seq data. Lastly, in Section C.3, we use our position weight matrices to construct sequence logos.

C.1 Generating position weight matrices from known genomic binding sites.

To construct a position weight matrix using these genomic binding sites, we must first align all the available binding site sequences and determine the nucleotide statistics at each position. Specifically, we count the number of each nucleotide, N_{ij}, at each position along the binding site. Here the subscript i refers to the position, while j refers to the nucleotide, A, C, G, or T. We can then calculate a position probability matrix (also 4xL) where each entry is found by dividing these counts by the total number of sequences in our alignment,

$$p_{ij} = \frac{N_{ij}}{N_g}. \tag{S1}$$

Note that in situations where the number of aligned sequences is small (e.g., less than five), pseudocounts[8] are often added to regularize the probabilities of the counts in the calculation of position probabilities,

$$p_{ij} = \frac{N_{i,j} + B_p}{N_g + 4 \cdot B_p}, \tag{S2}$$

where B_p is the value of the pseudocount. The argument for their use is that when selecting from a small number of binding site sequences, just by chance infrequent nucleotides will be absent, and assigning them a probability (p_{ij}, noted above) of zero may be too stringent of a penalty[8,9]. We let $B_p = 0.1$. In the limit of zero binding site sequences (i..e with no sequences observed), this will result in probabilities p_{ij} approximately equal to the background probability used in calculating the position weight matrix below (and a non-informative sequence logo).

Finally, the values of the position weight matrix are found by calculating the log probabilities relative to a background model[10],

$$PWM_{ij} = \log_2 \frac{p_{ij}}{b_j}. \tag{S3}$$

The background model reflects assumptions about the genomic background of the system under investigation. For instance, in many cases it may be reasonable to assume each base is equally likely to occur. Given that we know the base frequencies for *E. coli*, we choose a background model that reflects these frequencies (b_j: $A = 0.246$, $C = 0.254$, $G = 0.254$, and $T = 0.246$ for strain MG1655; BioNumbers ID 100528, http://bionumbers.hms.harvard.edu). From Equation S3, we can see that the value at the i, j^{th} position will be zero if the probability, p_{ij}, matches that of the background model, but non-zero otherwise. This reflects the fact that base frequencies matching the background model tell us nothing about the binding preferences of the transcription factor, while deviation from this background frequency indicates sequence specificity.

C.2 Generating position weight matrices from Sort-Seq data.

Next we construct a position weight matrix using our Sort-Seq data. Here we appeal to the result from Berg and von Hippel, that the logarithms of the base frequencies above should be proportional to their binding energy contributions[10,11]. Berg and von Hippel considered a statistical mechanical system containing L independent binding site positions, with the choice of nucleotide at each position corresponding to a change in the energy level by ε_{ij} relative to the lowest energy state at that position. ε_{ij} corresponds to the energy entry from our energy matrix, scaled to absolute units, $\varepsilon_{ij} = A \cdot \theta_{ij} + B$ (where θ_{ij} is the i, j^{th} entry as further discussed in Section H.3.3). An important assumption is that all nucleotide sequences that provide an equivalent binding energy will have equal probability of being present as a binding site. In this way, we can relate the binding energies considered here to the statistical distribution of binding sites in the previous section. The probability p_{ij} of choosing a nucleotide at position i will then be proportional to the probability that position i has energy ε_{ij}. Specifically, the probabilities will be given by their Boltzmann factors normalized by the sum of states for all nucleotides,

$$p_{ij} = \frac{b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}{\sum_{j=A}^{T} b_j \cdot e^{-\beta A \cdot \theta_{ij} \cdot s_{ij}}}, \tag{S4}$$

where $\beta = 1/k_B T$, with k_B is Boltzmann's constant and T the absolute temperature. As above, b_j refers to the background probabilities of each nucleotide. Note that the energy scaling factor B drops out of this equation since it is shared across each term.

One difficulty that arises when we use energy matrices that are not in absolute energy units is that we are left with an unknown scale factor A, preventing calculation of p_{ij}. We appeal to the expectation that mismatches usually involve an energy cost of 1-3 $k_B T$[12]. In other work within our group, we have found this to be a reasonable assumption for LacI. Therefore, we approximate it such that the average cost of a mutation $\langle A \times \theta_{i,j} \rangle = 2 k_B T$. We can then calculate a position weight matrix from Equation S3.

C.3 Construction of sequence logo

With our position weight matrices in hand we now construct sequence logos by calculating the average information content at each position along the binding site. With our four letter alphabet there is a maximum amount of information of 2 bits ($\log_2 4 = 2$ bits) at each position i. The information content will be zero at a position when the nucleotide frequencies match the genomic background, and will have a maximum of 2 bits only if a specific nucleotide is completely conserved. The total information content at position i is determined through calculation of the Shannon entropy, and is given by,

$$I_i = \sum_{j=A}^{T} p_{ij} \cdot log_2 \frac{p_{ij}}{b_i} = \sum_{j=A}^{T} p_{ij} \cdot \mathrm{PWM}_{ij}. \tag{S5}$$

Here, PWM_{ij} refers to the i, j^{th} entry in the position weight matrix[10,13]. The total information content contained in the position weight matrix is then the sum of information content across the length of the binding site.

To construct a sequence logo, the height of each letter at each position i is determined by,

$$\text{Seqlogo}_{ij} = p_{ij} \cdot I_i, \tag{S6}$$

which is in units of bits. This causes each nucleotide in the sequence logo to be displayed as the proportion of the nucleotide expected at that position scaled by the amount of information contained at that position[7]. To construct sequence logos we use custom Python code written by Justin Kinney and available on our GitHub repository for this work (https://www.github.com/RPGroup-PBoC/sortseq_belliveau; DOI: 10.5281/zenodo.1184169).

C.4 Comparison of Sort-Seq sequence logos.

For the various annotated binding sites identified in this work we used our Sort-Seq data to generate energy matrices. While these energy matrices provide a concrete way to understand the sequence-dependent DNA-protein interaction, it was also useful to generate sequence logos for visualization and to compare with sequence logos more conventionally generated using known genomic binding site sequences. In Fig. S4 we show this comparison for transcription factors with three or more known genomic binding sites, with agreement more apparent when genomic binding site logos are based on a larger number of known sequences.

We also report the Pearson correlation coefficient between the position weight matrices from the Sort-Seq inference and the genomic alignment. To compare the two position weight matrices we first apply identical 'gauge fixing' to both matrices being compared (discussed further in Section H.3.1). Each column of the matrices are set to have a mean energy of zero and their matrix norms (or inner products) are normalized to have value one. Under this constraint, the Pearson correlation coefficient is simply given by the summed product of matrix entries,

$$r = \frac{COV(\text{PWM'}_X, \text{PWM'}_Y)}{\sigma_X \cdot \sigma_Y} = \sum_{i=1}^{L} \sum_{j=A}^{T} \text{PWM'}_{X,i,j} \cdot \text{PWM'}_{Y,i,j}. \tag{S7}$$

Here, COV refers to the covariance between PWM'_X and PWM'_Y, where the superscript prime indicates that the matrices have been gauge fixed (mean energy in each column of zero and the matrix norm of 1). The subscript X, for example, would correspond to the Sort-Seq matrix, and Y, to the genomic matrix. σ_X and σ_Y refer to the standard deviation of the matrix entries for PWM'_X and PWM'_Y. We note that while Pearson correlation coefficient provide one useful metric to compare energy matrices, there are alternative metrics that are also commonly used (Kullback-Leibler divergence, Euclidean distance, and Pearson χ^2 test, among others; See Gupta et al. 2007[14] which is the publication for the TOMTOM motif comparison software and provides a good summary of these).

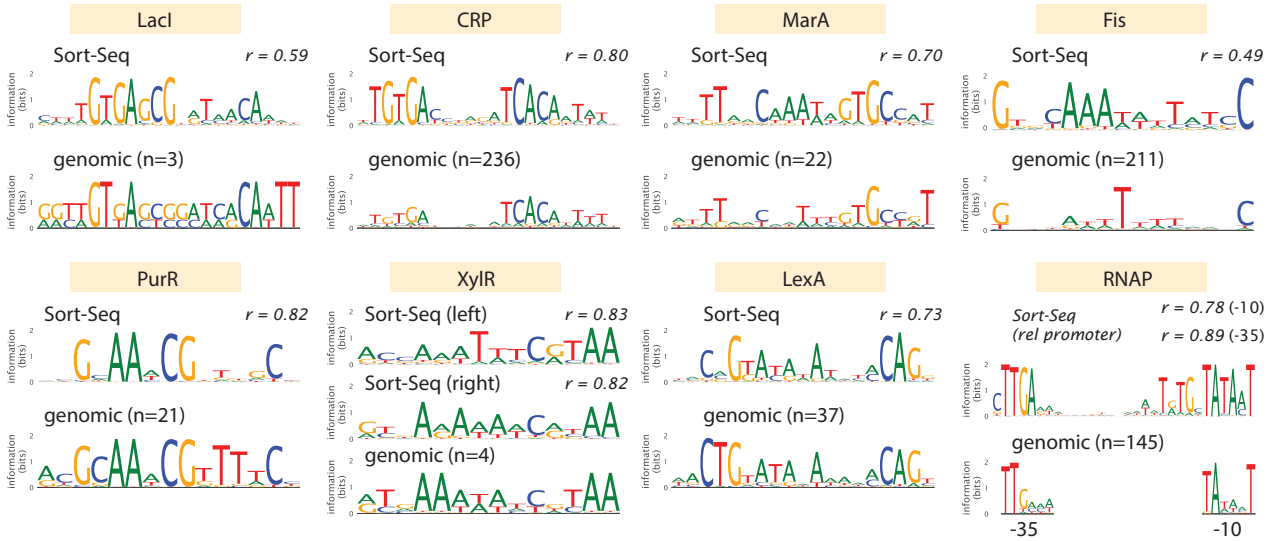

Figure S4. Comparison between Sort-Seq and genomic-based sequence logos. Comparisons are shown for LacI, CRP, MarA, Fis, PurR, XylR, LexA, and RNAP. Binding site sequences were obtained from RegulonDB, where n identifies the number of genomic binding sites that were used to construct the sequence logo. The Sort-Seq RNAP logo is based on data from the *rel* promoter. For the genomic RNAP logo, sequences were taken from computationally predicted RNAP binding sites on RegulonDB (top 3.3 % scored sequences using their reported metric) for the 6 bp regions of the -10 and -35 binding sites. Pearson correlation coefficients are calculated with Equation S7 using the position weight matrices from the Sort-Seq and genomic matrices. For LexA, the first four bp were not used in the calculation due to overlap with the -10 RNAP binding site of the *yebG* promoter.

D Statistical mechanical model of the DNA affinity chromatography approach.

In order to better understand the factors that govern the success of the DNA affinity chromatography method, we took a statistical-mechanical approach to help identify the key parameters that will influence the fold enrichment of transcription factors that we measure. We are interested in calculating the probability that the transcription factor of interest binds to the target DNA sequence used for purification. We will ignore possible binding by proteins to the magnetic beads, to which the DNA oligonucleotides are tethered.

To calculate the probability that the transcription factor of interest is bound, we will simplify our problem by assuming that all other proteins present in the lysate will bind the DNA with some average nonspecific binding energy. This must be included since these proteins will act as potential competitors for the tethered DNA. We must first enumerate the possible states of our DNA. For each DNA affinity purification, this will include the following three states: 1) no protein bound to the DNA, 2) the target transcription factor bound, and 3) a nonspecific protein is bound. These are shown in Supplemental Fig. S5D for each of the DNA oligonucleotides used for the two different purifications performed.

The non-normalized probability of each state occurring is simply given by $e^{-\beta(\varepsilon_i - \mu_i)}$. Here, ε_i is the protein-DNA binding energy and μ_i, the chemical potential, for species i[15]. $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. The chemical potential contains information about concentration, and it is possible to alternatively write the non-normalized probability in terms of these, which we will rewrite as $e^{-\beta(\varepsilon_i - \mu_i)} = C_i/C_o e^{-\beta \Delta\varepsilon_i}$. Here, C_i is the concentration of protein species i, and C_o, is the standard concentration, which is taken as 1 M. $\Delta\varepsilon_i$ is the binding energy for

species i, which will be taken as relative to the unbound state.

We can now write the statistical weight for each state, which is summarized in Fig. S5D. We allow the unbound state to act as our reference state with an energy equal to zero, corresponding to a statistical weight of 1. The probability of our target protein being bound to a certain DNA target, $P_{bound,DNA}$, will then be given by the statistical weight for the state where the target protein is bound, divided by the sum of statistical weights for each state. This is given by,

$$P_{bound,DNA} = \frac{\frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,DNA}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,DNA}}} \tag{S8}$$

where the subscript 'TF, DNA' identifies the target transcription factor and its binding to a particular DNA target. In regard to our two purifications shown in Supplemental Fig. S5D, $\Delta\varepsilon_{TF,s}$ refers to the binding energy of the transcription factor to its target binding site, while $\Delta\varepsilon_{TF,ns}$ refers to the nonspecific binding energy to non-target reference DNA. In addition, $\Delta\varepsilon_{ns}$ refers to the binding energy of other proteins present in the lysate, which may bind the DNA nonspecifically.

We can now calculate the fraction of bound transcription factor, $P_{bound,DNA}$, using some reasonable values for *E. coli*[16,17]. Here we use $C_{TF} = 10^{-8}M$ (about 10 copies per cell), $C_o = 1M$, $\Delta\varepsilon_{TF,s}$ = - $15k_BT$, and $\Delta\varepsilon_{ns}$ = - $5k_BT$. $C_{ns} = 3 \cdot 10^{-3}M$, which is the approximate number of proteins in *E. coli*. The specific numbers will depend on the DNA target sequence used, the concentration of target protein, as well as the lysate preparation itself. Here we find $P_{bound} \approx 0.02$. In contrast, for the nonspecifically bound fraction we calculate about a ten fold higher fraction of nonspecific protein bound to the DNA. Even though the binding energy for a target transcription factor is significantly stronger than the competitor proteins that bind nonspecifically, we bind more nonspecific proteins due to their high concentration. This result in particular highlights our rationale for using an additional reference purification to distinguish the target transcription factor from non-specifically bound proteins[18]. We consider the consequences of this next.

In this second reference purification, the DNA no longer has the target binding site, and thus the value of $P_{bound,DNA}$ for the transcription factor should be significantly smaller. We can use Equation S8 to calculate expected ratio of transcription factor bound to target DNA versus reference DNA, given by,

$$\frac{P_{bound,target}}{P_{bound,reference}} = \frac{\frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,s}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,s}}} \cdot \frac{1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,ns}}}{\frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,ns}}} \tag{S9}$$

$$= \frac{e^{-\beta \Delta \varepsilon_{TF,s}}}{e^{-\beta \Delta \varepsilon_{TF,ns}}} \frac{1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,ns}}}{1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} + \frac{C_{TF}}{C_o} e^{-\beta \Delta \varepsilon_{TF,s}}}. \tag{S10}$$

Again, the subscript $\Delta\varepsilon_{TF,ns}$ refers to the binding energy of the transcription factor to the non-target (i.e. non-specific) reference DNA. Using the example values from our calculation of P_{bound} above, we find that $1 + \frac{C_{ns}}{C_o} e^{-\beta \Delta \varepsilon_{ns}} \gg e^{-\beta \Delta \varepsilon_{TF,s}} \gg e^{-\beta \Delta \varepsilon_{TF,ns}}$, with Equation S10 simplifying to

$$\frac{P_{bound,target}}{P_{bound,reference}} \approx \frac{e^{-\beta \Delta \varepsilon_{TF,s}}}{e^{-\beta \Delta \varepsilon_{TF,ns}}} = e^{-\beta(\Delta \varepsilon_{TF,s} - \Delta \varepsilon_{TF,ns})}. \tag{S11}$$

This result suggests that the enrichment ratio should mainly depend on the difference in binding energy between the DNA sequences used in the two purifications. Our results from purifying LacI with strains containing different LacI copy number per cell and with different DNA target sequences (see Section E.3 and Supplemental Fig. S5C) appear to agree with this result in general, where we see greater enrichment when using the strong Oid target LacI binding site sequence than the weaker O3 binding site sequence. This appears to influence the enrichment ratio more significantly than protein concentration, although further work will be needed to fully characterize this relationship.

E DNA affinity chromatography and mass spectrometry experimentation and analysis.

In this section we provide additional details on the use of DNA affinity chromatography and mass spectrometry to identify the transcription factors that bind to our putative binding sites. In particular, we provide additional data to demonstrate protein labeling and characterize the dynamic range expected from our enrichment measurements (see Methods Section for more details about the approach). We also provide data from an affinity chromatography experiment where the same DNA oligonucleotide sequence was used for both target and control purifications. The ideal result from such an experiment is that each protein detected is found in equal abundance between the two purifications performed, yielding an enrichment ratio equal to one. However, there is some inherent variability in such a measurement and this provides some characterization of that uncertainty. Lastly, we provide additional data showing that we can purify and identify transcription factors at concentrations ranging from about 10 to 1,000 copies per cell.

E.1 Characterization of SILAC labeling and measurement of protein enrichment ratios.

To ensure *E. coli* cells incorporated the heavy isotope of lysine ($^{13}C^{15}_{6}N_2$-L-lysine, heavy lysine), we first generated an auxotropic strain which was unable to synthesize its own lysine through deletion of the *lysA* gene[19]. LysA is an enzyme that catalyzes the last step in lysine biosynthesis. Furthermore, to ensure proteins would be sufficiently labeled when growing cultures for lysate preparation, we inoculated our cultures with a large dilution of 1:5,000. This large dilution is important since the inoculate represents an unlabeled fraction of the cell population. We checked the effective labeling efficiency by combining lysates from cells grown with heavy and light (natural) lysine over a range of ratios between 0.1/1 to 1,000/1 (heavy / light). The measured ratio in abundance for each of the proteins detected among the two lysates are plotted in Fig. S5A. In calculating these values, we found that the median average from our 1/1 combination was measured to be 0.71 (heavy / light). This suggested there may have been some inaccuracy in the Bradford assay that was used to measure protein concentration prior to mixing our lysates. We therefore renormalized the ratios according to this measured ratio. The data suggests a labeling efficiency of at least 99% (red dashed line, in comparison to perfect labeling shown by the gray dashed line). One important aspect highlighted by this data is that the highest enrichment ratio we should expect to measure in our DNA affinity experiments is several hundred fold.

E.2 Characterization of protein enrichment variability from identical DNA targets.

For each DNA affinity chromatography experiment, we are trying to identify a DNA-binding protein that shows up in higher abundance when we use the target binding site sequence identified by Sort-Seq (i.e. a transcription factor binding site), relative to a purification where that target sequence has been mutated away. To ensure that our measured enrichment ratios were not an artifact of noise in the measurement, it was important to also check the measurement variability when both lysate purifications used an identical DNA sequence. In this way, we could characterize the inherent variability in such a measurement. To proceed, we performed experiments using the control DNA sequence that was used in our purification of the *purT* promoter target (Fig. 5C, though any DNA oligonucleotide could have been used). We performed this in triplicate and consider the average enrichment ratios for each protein measured across the three experiments. In the left panel of Fig. S5B we show the average enrichment values that were measured for each of the detected proteins. Since many of the data points fall on top of one another, we also provide a histogram of the associated data (Fig. S5B, right plot). Here we have taken the logarithm of the enrichment ratios so that the bins are equally spaced. The shaded region in both plots identifies the range between the 2.5^{th} and 97.5^{th} percentiles, highlighting that the majority of proteins were found between an enrichment ratio of 0.2 and 3.3 (or log enrichment ratio of between -1.5

and 1.2). The ideal enrichment expected would be a value of 1.0 or log ratio of 0. In the main text, the enrichment values for transcription factors found using targets associated with the *lacZ*, *relB*, *purT*, *xylE*, and *dgoR* promoters fall well outside of the range of variability found here.

E.3 Identification of LacI by mass spectrometry using strains with a variable LacI copy number.

Finally, one experiment that we performed, in addition to purifying LacI with different strength binding site targets (i.e., Fig. 4A), was to consider the copy number per cell of the LacI target, as copy number should influence the fraction of bound LacI to the DNA target (see details in Section D). Here we used strains whose protein concentration has been measured during growth in M9 minimal media with 0.5% glucose and whose average LacI number had previously been measured to range from the native expression of 11 ± 2 tetramers per cell, to a maximum concentration of 870 ± 170 tetramers per cell. In Supplemental Fig. S5C we show the enrichment ratios measured for LacI from individual experiments (n = 1-2 per strain). Here we were able to purify LacI using either the weak O3 or strong Oid binding site sequence for each of the different strains, though we also see that the O3 target sequence provides an enrichment that is much closer to the tail of the control experiment in Fig. S5B. Additionally, while the copy number of LacI appears to affect the enrichment ratio in some experiments, it does not have a consistently significant effect.

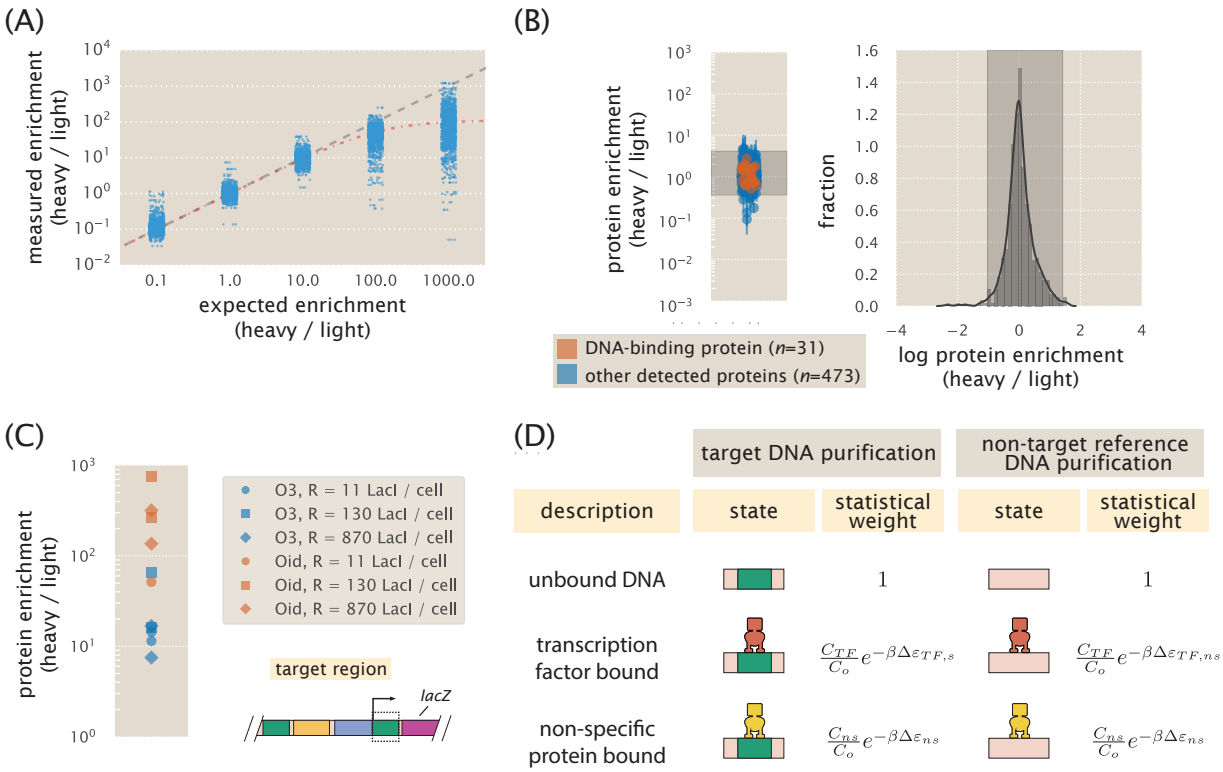

Figure S5. Identification of transcription factors using DNA-affinity chromatography and mass spectrometry. (A) Characterization of stable isotopic lysine labeling and mass spectrometry measurement sensitivity. Lysates from cell cultures grown in either heavy ($^{13}C^{15}_{6}N_2$-L-lysine) or normal L-lysine were combined at ratios between 0.1:1 to 1000:1 heavy:light and the measured ratios in abundance are plotted for each protein. Note that for the 1:1 ratio we found a median ratio of 0.71. We therefore renormalized the ratio values using this as a correction factor. Data points represent the average values from $n = 3$ replicates. The gray line represents the expected measurement under perfect labeling, while the red line represents a 99.1 % labeling efficiency (assuming that some fraction of heavy lysate is unlabeled). (B) DNA-affinity purification using the same DNA oligonucleotide to purify protein for both heavy and light cell lysates ($n = 3$). The scatter plot shows the average enrichment values for each protein detected. Proteins with DNA binding motifs[20] are shown in red ($n = 41$), while other detected proteins are in blue ($n = 581$). Error bars represent the standard deviation, calculated from log protein enrichment values. The histogram shows the distribution of the measured ratios for all detected proteins, with 95% of the measurements contained between a log enrichment of -1.5 and 1.2, as indicated by the shaded region. Lysates were prepared from cells grown in M9 minimal media with 0.5% glucose. (C) DNA-affinity purification of LacI using three different *E. coli* strains with repressor copy numbers per cell of 11 ± 2, 130 ± 20, and 870 ± 170 (tetramers per cell)[21]. Operator strength was varied by purifying LacI with either the weak O3 or strong Oid operators. LacI was detected as the most significantly enriched protein among all proteins detected. Each data point represents the enrichment from a single purification experiment ($n = 1$-2 for each strain). (D) States and weights are shown for an oligonucleotide in which a target transcription factor and other cellular proteins compete for a DNA binding site. Within the cell lysate, the target protein is present at a concentration C_{TF}, while all other proteins, which may bind the DNA nonspecifically are present at a concentration C_{ns}. C_o is the standard concentration. The difference in energy between a repressor bound to the target DNA binding site and an unbound DNA is $\Delta\varepsilon_{TF,s}$ when the binding site is present. Otherwise, the binding energy is given by $\Delta\varepsilon_{TF,ns}$. Other proteins that bind nonspecifically, irrespective of the DNA sequence, have a binding energy of $\Delta\varepsilon_{ns}$.

F Selection of the mutagenesis window for promoter dissection by Sort-Seq.

In designing our mutagenized promoter libraries, we found it useful to consider what was known regarding both the genes of interest and general patterns of transcriptional regulation in *E. coli* and bacteria more broadly. Two useful resources were RegulonDB[4] and EcoCyc[20], which summarize much of what is known about transcriptional regulation in *E. coli*. RegulonDB, in particular, aims to compile all available data regarding gene regulation in *E. coli* into a single database and is the most complete record available for *E. coli*[22].

While Sort-Seq enables us to identify all proteins involved at a promoter, one potential limitation is that a transcription factor binding site will only be identified if it was contained within our mutagenized region. Using the known transcription factor binding sites in *E. coli* as a guide in our design, we made an educated guess regarding where we should search for transcription factor binding sites. Fig. S6 shows a histogram of all of the transcription factor binding site positions from RegulonDB. By staggering a set of 60bp windows to cover a 150 bp region, we found we would expect to capture 73 percent of the known transcription factor binding sites. We chose 60 bp-70 bp windows for most libraries since they could be readily synthesized by Integrated DNA Technologies (USA) and were more economical than longer oligonucleotides. We also included about 15 bp of overlap between staggered regions to provide some replicates of the mutated base pairs on the different libraries.

It is also useful to note that our approach does not require that this specific strategy be used to create mutagenized promoter constructs. The methodology only requires compatibility between the length of the mutagenized region probed and the sequencing platform used. Microarray synthesized oligonucleotides provide another approach for targeted oligonucleotide design[23], and error-prone PCR can enable longer mutagenized windows within a single library[24,25]. In addition, advances in sequencing, either through longer reads or alternative sequencing platforms such as PacBio (Pacific Bioscience, USA) and MinION (Oxford Nanopore Technologies, UK) are making it possible to sequence longer mutagenized regions, and CRISPR technologies could make it possible to identify longer range interactions such as DNA looping in bacteria (e.g., a one megabase region was considered in Fulco *et al.*[26]).

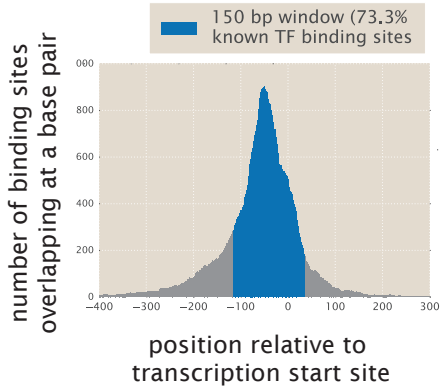

Figure S6. Distribution of known transcription factor binding sites in *E. coli*. The histogram shows the genome-wide distribution of transcription factor binding sites relative to their respective transcription start sites. Binding sites were compiled from RegulonDB and used to calculate the number of overlapping binding sites at each position using the length and position of each binding site sequence. The location of the 150 bp mutation window used in this study is shown in blue, expected to capture upwards of 70% of known transcription factor binding site position.

G Additional data from Sort-Seq experiments of the main text.

Here we provide additional data and analysis on the promoters of *rel*, *mar*, *yebG*, *purT*, *xylE*, and *dgoR* to provide additional support for the results and conclusions made in the main text.

G.1 The *rel* and *mar* promoters

In our analysis of the *rel* and *mar* promoters in the main text, it was noted that the sequence specificity of the repressors RelBE and MarR lacked any prior characterization. In order to validate that the observed features of the expression shift plots were due to binding by these regulatory proteins, we performed additional Sort-Seq experiments in deletion strains for these regulators. The expression shift plots were shown in the main text (Fig. 3). Here we provide a more quantitative analysis to show that the energy matrices for binding by RelBE and MarR poorly describe the sequence data when *relBE* and *marR* are deleted, respectively.

Since the transcription factors have been deleted, we expect the energy matrix predictions of each sequence's binding energy to provide no clear trend across the sorted bins (i.e., zero or little mutual information). To first give a sense for how mutual information is calculated, in Fig. S7A and Fig. S7B we show the estimated joint distributions when we apply the RelBE energy matrix (from Fig. 2B of the main text) to either a replicate Sort-Seq experiment or to the $\Delta relBE$ deletion data. When applying the RelBE energy matrix to the wild-type data, we find a clear trend, with strongest binding energies (lowest rank order) more likely found at the lowest fluorescence bin, and weakest binding energies more likely in the highest fluorescence bin.

Next we focus in on our data from the deletion strains of $relBE$ and $marR$ (Figure S7C and S7D, respectively). In each case, we find that our energy matrices poorly describe the data and are not substantially better than a randomly generated matrix. In Figure S7B it might have been noted that there were still some positions with non-zero expression shift (i.e., still appear informative). In order to show that this remaining information cannot be accounted for from our energy matrices, we also estimated the maximum information present in the Δ strain data sets (by directly fitting a matrix to the Δ strain data). Importantly, we find that these features cannot be explained by our wild-type strain RelBE and MarR energy matrices, and must be due to other features in the data.

G.2 The *yebG* promoter

The *yebG* promoter is among a variety of genes known to increase expression when cells are under DNA damage stress (SOS response)[27]. It shares the intergenic region with the *purT* promoter. In the main text we considered the *yebG* promoter in cells grown in standard M9 minimal media with 0.5% glucose (Fig. 5A, *Left*). While the expression shifts appeared to align with annotated binding sites for LexA (positive shift), and the RNAP binding site (negative shift), we did not show evidence for the identity of each binding protein in the main text. Here we present results from our inference of energy matrices using our Sort-Seq data, which confirm the identity of the binding proteins. We also explore regulation of *yebG* by perturbing the regulatory state through induction of the SOS response[27,28].

We begin by considering the Sort-Seq data from cells grown in M9 minimal media with 0.5% glucose. In Fig. S8A we show the inferred energy matrices associated with the annotated site for LexA. This was in excellent agreement with the known sequence specificity of LexA (see Fig. S4 for a direct comparison with the genomic sequence logos). We note, however, that the RNAP binding site was shifted by 9 bp from the annotated binding site[28], with an overlap between the -10 RNAP site and 4 bp of the LexA binding site.

We were also interested in confirming that the *yebG* promoter responds DNA stress and is induced as part of the SOS response. By repeating Sort-Seq in cells grown in non-lethal concentrations of mitomycin C (1 μg/ml)[28] we observed a dramatic increase in expression relative to growth without mitomycin C. Fluorescence histograms showing expression from our plasmid reporter in

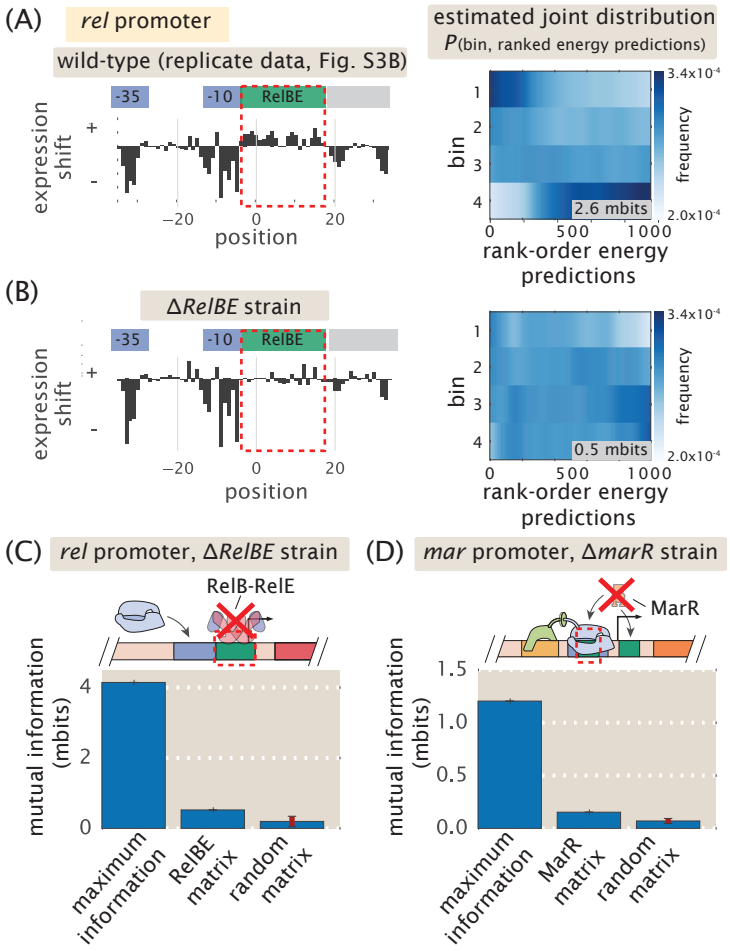

Figure S7. Predictive information of transcription factor energy matrices when applied to Sort-Seq data. In (A) and (B) we use our RelBE energy matrix to predict binding energies across all sequence data for a replicate experiment with wild-type *E. coli* and a $\Delta relBE$ strain, respectively. The 2-d histograms show the estimated joint probability distributions between bin and rank-ordered energies (generated by binning sequences into 1000 bins). The calculated information (in mbits) shown in the joint distribution plot represents the mutual information from these estimated joint distributions. In (C) and (D) we focus on our transcription factor deletion strains ($relBE$ in (C) and $marR$ in (D)), and similarly calculate mutual information between bin and energy matrix predictions (again, using their rank-ordered predictions). The 'maximum information' represents the estimated maximum information that might be obtained by fitting an energy matrix to the Δ strain data. The 'random matrix' represents the average mutual information calculated from 20 randomly generated energy matrices (error bar represents standard deviation) applied to the sequence data. To provide consistent comparisons, all matrices were 'gauge fixed' such that the mean energy in each column of zero and the matrix norm of 1. Note that for MarR we show analysis for the left MarR binding site. In the right binding site, there is additional information corresponding to the ribosomal binding site. The joint probability distribution and associated mutual information are calculated following the procedure described in Section H.3.

non-mutagenized promoter constructs are shown in Fig. S8B. From the expression shift plots and information footprint (which are defined in Section H and used in Kinney *et al.*[29]) in Fig. S8D we find that this is due to loss of repression at the LexA binding site. This is consistent with the expectation that LexA undergoes proteolysis as part of the SOS response[27].

G.3 The *purT* promoter

When cells were grown in the presence of adenine, we identified a putative repressor site between the -10 and -35 regions of the RNAP binding site of the *purT* promoter. In our initial attempt to identify the associated transcription factor we performed a DNA affinity purification using conditions that matched the growth conditions where repression was observed. However, as shown in Fig. S8C, the most significantly enriched protein (GlpR) only showed an enrichment of about 2.9, which was near the shaded region associated with most other non-specific proteins detected. Only upon repeating our purification in the presence of hypoxanthine (10 μg/ml) (Fig. 5C) did we find enrichment of PurR (approximately 350 fold relative to our reference purification).

G.4 The *xylE* promoter

In the main text it was noted that we could not perform Sort-Seq on the *xylE* promoter unless cells were grown in xylose. In Fig. S8E, we show the associated fluorescence histograms from libraries grown in either glucose or xylose. Interestingly, each mutated window was essentially identical to autofluorescence when cells were grown in glucose. In contrast, growth in xylose showed differential expression for each of the mutated regions. While the promoter was expected to be sensitive to the presence of xylose (causing an increase in expression[1]), this was still a non-obvious result without prior knowledge of whether repressors or activators were involved.

In our analysis we also noted in the main text that the identified set of activator binding sites conformed well with the two other promoters regulated by XylR and CRP, namely *xylFG* and *xylAB*. Here we scanned our inferred energy weight matrix across the intergenic regions of *xylFG* and *xylAB*, in order gain further confidence that the identified feature matched the known binding specificity of these transcription factors. These are shown in Fig. S8F. At each position in these plots, we use the energy matrix to calculate the binding energy of the putative transcription factors. For each we identify a strong peak that does indeed align well with the annotated binding sites of XylR and CRP. While our predicted binding energies are not in absolute k_BT units, they are much more negative than the promoter background and predict a similar binding energy (in gauge fixed, arbitrary energy units) to the binding site region of the *xylE* promoter.

G.5 The *dgoR* promoter

The last promoter we considered was associated with the expression of the *dgoRKADT* operon. Due to the complexity observed, we were unable to show all data in the main text that supported our identification of the regulatory architecture. In particular, here we show the sensitivity to the different carbon sources considered and additional analysis of the identified regulatory binding sites for DgoR, RNAP, and CRP.

G.5.1 The *dgoR* promoter is induced when cells are grown in galactose and D-galactonate.

Prior to performing Sort-Seq on this promoter, we confirmed prior observations that expression was sensitive to the presence of galactose and D-galactonate[1,30]. Using a wild-type promoter plasmid for the *dgoR* promoter, cells were grown in M9 minimal media with either 0.5% glucose, 0.23% D-galactose, or 0.23% D-galactonate. Fluorescence histograms are shown in Fig. S9A, where we observed higher expression in galactose over glucose, and even higher expression when cells were grown in D-galactonate.

G.5.2 An RNAP binding site is apparent in the downstream region of the *dgoR* promoter when cells were grown in glucose.

In Fig. 7A we showed plots comparing the expression shifts upon mutation when cells were grown in either glucose or D-galactonate. In Fig. S9B we reproduce the expression shift plots along with an energy matrix for the region from approximately -70 to -30, which helped us to identify the RNAP binding site

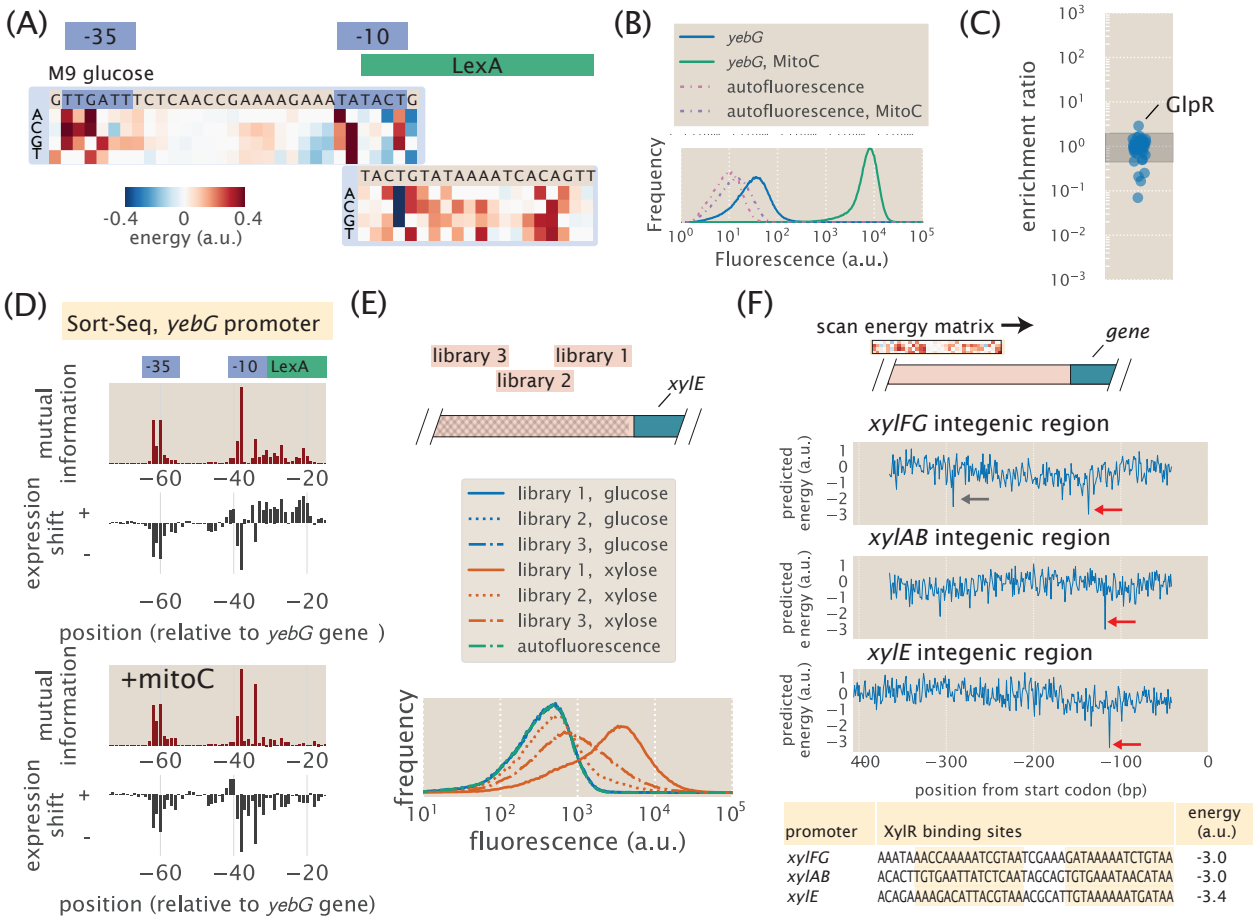

Figure S8. Extended analysis of the *yebG*, *purT*, and *xylE* promoters. (A) Energy matrices were inferred for the binding sites of LexA and RNAP. Data are from cells grown in M9 minimal media with 0.5% glucose. (B) Fluorescence histograms for a wild-type *yebG* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, and with or without mitomycin C (1 μg/ml). Mitomycin C induces the SOS response[28] and dramatically increases expression from the *yebG* promoter. Autofluorescence histograms refer to cells that did not contain the GFP promoter plasmid. (C) DNA affinity chromatography performed using the identified repressor site on the *purT* promoter. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose and binding was performed in the presence of adenine (100 μg/ml) to match the growth conditions where repression was observed. (D) Information footprints and expression shift plots are shown for the *yebG* promoter in the presence or absence of mitomycin C (1 μg/ml). Cells were grown in M9 minimal media 0.5% glucose. (E) Fluorescence histograms are shown for the three *xylE* libraries (different mutated regions), with cells grown in M9 minimal media with either 0.5% glucose or 0.5% xylose. While xylose led to differential expression for the different libraries, cells grown in glucose were identical to autofluorescence. (F) The energy matrix associated with two tandem putative binding sites for xylR and CRP (Fig. 6C) was scanned across the intergenic regions of *xylAB*, *xylFG*, and *xylE*. The predicted energy is plotted for each position, and a strong binding site was identified in each promoter (red arrow). For *xylAB*, and *xylFG*, this matched the known binding sites for XylR and CRP on these promoters and their sequences and binding energy predictions are noted below the plots. The promoters of *xylAB* and *xylFG* share the same intergenic regions, but are in opposite coding directions. The reverse complement of the binding site identified in the *xylAB* promoter also showed a strong binding energy prediction (gray arrow in *xylFG* scan).

in this region. While the -10 TATAAT motif is quite apparent, the -35 site is less clear. Interestingly, while the -35 region shows a most energetically favorable sequence of TTTACA (close to the consensus of TTGACA), the wild-type sequence is CCCCCC and suggests this is a weak RNAP binding site.

G.5.3 Deletion of the *dgoR* gene recovers the induced phenotype.

Comparing the expression shift values at each position in cells grown in either glucose or D-galactonate, we find that they are poorly correlated (Fig. S9C, left plot). However, upon identifying DgoR as a putative regulator in the upstream region of the promoter, we then performed Sort-Seq in a $\Delta dgoR$ strain. This was shown in Fig. 7D with cells grown in glucose. Interestingly, the expression shifts were much more similar to the wild-type cells grown in D-galactonate, suggesting that deletion of *dgoR* has switched regulation to the induced state (Fig. S9C, right plot).

While it is unclear what causes the noisy profiles in the expression shift plots, one hypothesis was that the different RNAP binding sites were producing at least two distinct mRNA transcriptions, whose 5' untranslated might influence transcript stability and GFP expression. In particular, the upstream RNAP binding site will generate a much longer 5' untranslated region and mutations that influence mRNA structure and stability might show up as an effect on expression within the region we considered by Sort-Seq. Using the Salis lab ribosomal binding site calculator[31] and RNA structure predictions with NUPACK[32], we predicted the secondary structure of the two expected mRNAs transcripts (Fig. S9D). We find that the longer transcript (expected when cells are grown with D-galactonate), does indeed predict a strong secondary structure that alter translation from this transcript.

G.5.4 Simulations of upstream promoter region identify multiple overlapping RNAP binding sites.

Next we consider additional analysis to support the presence of overlapping RNAP sites that was noted in Fig. 7C. Since Sort-Seq does not differentiate between multiple transcription start sites, the sorted data will represent a mixture of all transcripts generated from the promoter. Using our RNAP energy matrix from the *relBE* promoter (with an additional 1 bp spacer included to increase the distance between -10 and -35 to 18 bp), we were able to identify multiple overlapping sequences that each predicted a similar binding energy by RNAP. The sequence logo in Fig. 7C of the main text (top logo) therefore likely represents the convolution of these multiple binding sites and would explain why we do not see the conventional -35 RNAP motif in the sequence logo.

To convince ourselves that this was a reasonable hypothesis, we performed several Sort-Seq simulations of the *dgoR* promoter to estimate what we may have expected if 1-3 of these identified RNAP binding sites were functional. These simulations use energy matrices and a thermodynamic model of regulation to predict gene expression as a function of regulatory sequence in an attempt to mimic a real Sort-Seq experiment. The code used is available on our GitHub repository (https://www.github.com/RPGroup-PBoC/sortseq_belliveau; DOI: 10.5281/zenodo.1184169) and we briefly describe the approach here. We began by first generating a library of five million mutated *dgoR* promoter sequences (10% mutation rate). We then assumed that transcription from each RNAP is proportional to $P/N_{NS} \cdot e^{-\beta E}$ where P is the RNAP copy number per cell, $N_{NS} = 4.6 \times 10^6$ refers to the number of non-specific binding sites on the genome, and $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature. We introduced noise into our simulation by assuming that the RNAP copy number P was normally distributed across our library with a mean value of $3,000$ and standard deviation of 750 copies per cell[1, 33]. As defined in Supplemental Section H.3.1, E represents the binding energy as determined from the energy matrix.

Using these calculations to predict expression from each mutated sequence, the sequences were then computationally sorted in the same manner as that performed experimentally. We did this assuming the presence of one, two, or three active RNAP binding sites based on those identified. As shown in Fig. S9F, the presence of three RNAP binding sites produces a result that conforms much better with experimental results than the presence of only one RNAP binding site. Note that binding sites were successively included into the model based on their predicted binding energies (wild-type RNAP 1: -1.99

a.u., wild-type RNAP 2: -1.74 a.u., wild-type RNAP 3: -1.60 a.u.; versus an average of -0.14 a.u. and standard deviation of 0.56 a.u. when the energy matrix is scanned across the promoter).

G.5.5 The presence of the class II CRP activator binding site is enhanced using strain JK10, grown with cAMP.

Lastly, we show additional evidence to support the claim of a putative binding site for CRP. Since CRP binds to DNA by co-activation through binding with cAMP, we used the strain JK10 (based on TK310[29]; MG1655 $\Delta cyaA\Delta cpdA$), where we could control binding of CRP to DNA by direct supplement of cAMP to the growth media. Here we grew cells in EZrich MOPS media (Teknova, CA, USA) with D-galactonate as the carbon source and supplemented with 500 μM cAMP. While the sequence logos in Fig. 7E showed a good match with the left site of the CRP binding site, our hypothesis here was that addition of a high concentration of cAMP might enhance the CRP motif in our data. This appeared to be the case, and the right side of the binding site (which overlaps the -35 RNAP binding site) shows a stronger preference for the sequence CAC than present with the wild-type *E. coli* strain (important for binding by CRP in both the *lac* and *xylE* promoters).

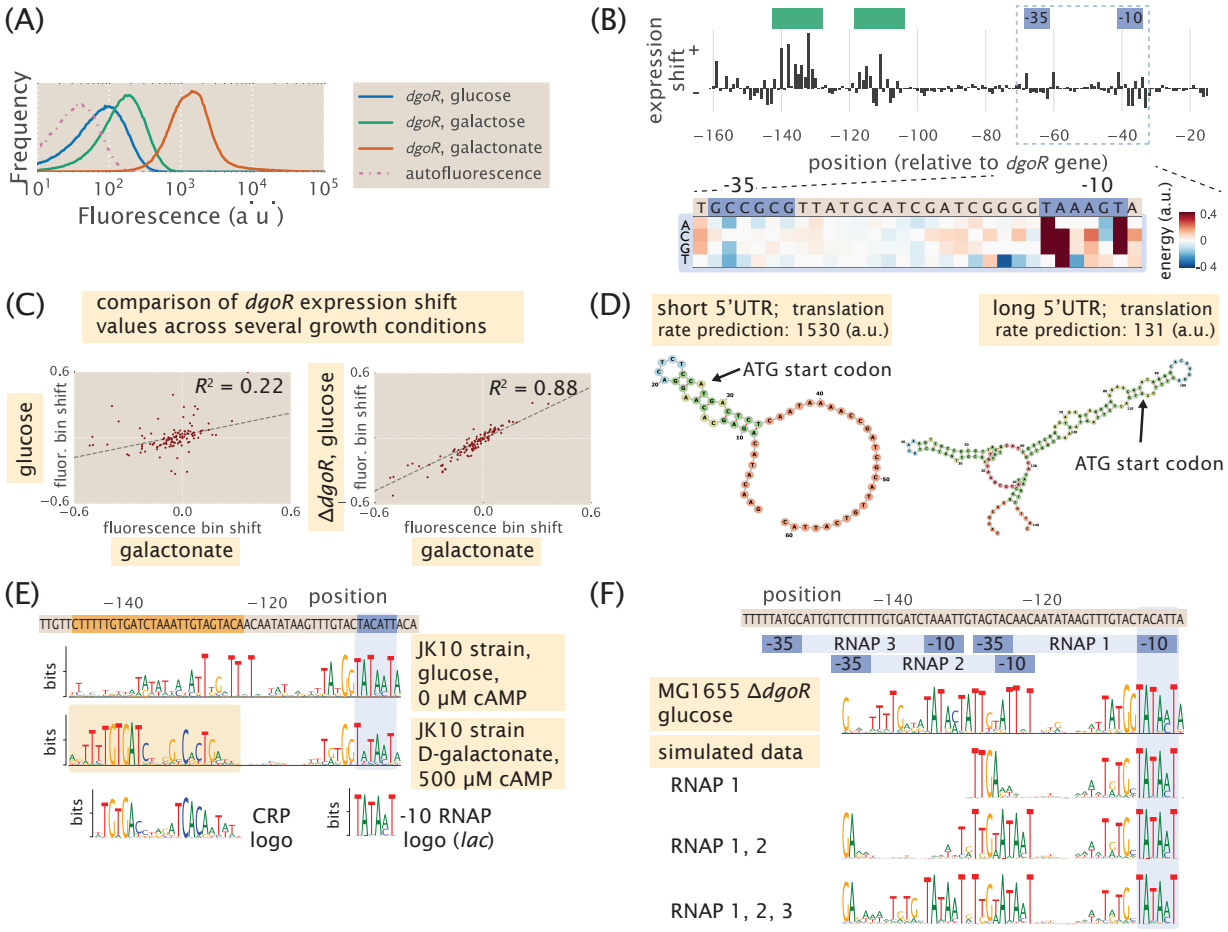

Figure S9. Extended analysis of the *dgoR* promoter. (A) Flow cytometry histograms of cells containing a wild-type *dgoR* promoter plasmid are shown for cells grown in M9 minimal media with 0.5% glucose, 0.23% galactose, or 0.23% D-galactonate. (B) Identification of an RNAP binding site that appears active when cells are grown in M9 minimal media with 0.5% glucose. The inferred energy matrix exhibits a clear -10 RNAP binding site (consensus sequence is TATAAT) and a poor -35 binding site (CCCCCC). (C) Expression shift values are plotted against each other (glucose vs. D-galactonate, and Δ*dgoR* glucose vs. D-galactonate) for positions -120 bp to -14 bp relative to the *dgoR* coding gene. Note that these are the same values used to generate the plot in Fig. 7A, just plotted against each other for each position. Δ*dgoR* cells appear to have the same regulatory phenotype as cells grown in D-galactonate, with a line of best fit showing much higher correlation between these data sets. (D) Predicted RNA transcript structure based on the two distinct RNAP binding sites. Growth in D-galactonate leads to the long 5' untranslated region and is found to produce strong secondary structure which predicts significantly lower translation rates of the *dgoR* gene than with the short transcript. The ATG start codon is identified. (E) Sequence logos were generated for the most upstream 60bp region containing the putative RNAP and CRP binding sites. Data is from Sort-Seq in strain JK10 (derivative of TK310[29]) and binding of CRP was induced through addition of 500 μM cAMP. Cells were grown in EZrich MOPS media (Teknova, CA, USA) with D-Galactonate as the carbon source. In comparison to the sequence logos shown in Fig. 7C (growth in D-galactonate), the right side of the CRP binding site is now in better agreement with the logo from the *lac* promoter. (F) Sequence logos are shown for simulated data for the upstream region of the *dgoR* promoter assuming one, two, or three RNAP binding sites. The top sequence logo shows the experimental result for Sort-Seq performed in a Δ*dgoR* genetic background, with cells grown in glucose.

H Extended Sort-Seq data analysis details.

H.1 Calculation of expression shifts

One of the first ways we analyze the sequence data from our Sort-Seq experiment is to look at the consequence of mutations at each position on the overall fluorescence. Specifically, at each position we calculate the average fluorescence bin of mutated nucleotides and compare this to the average bin for all the sequences in the data set (i.e. expression shift). Since we find that most mutations are deleterious to the binding of transcription factors or RNAP, we can use the change in fluorescence to identify regions associated with binding by repressors or activators and RNAP. This provides an alternative to the information footprints calculated in Kinney et al., 2010. While the information footprints can also be useful, the sign of the expression shifts is useful to determine the type of regulatory protein.

First we calculate the average bin for all the sequences in the data set. We let N_f be the total number of sequences in each bin, where f refers to the bin number ($f = 1$, 2, 3, and 4, for four bins). The average fluorescence bin is then given by the arithmetic average across all bins,

$$\langle f \rangle = \sum_{f=1}^{4} f \cdot p(f) = \sum_{f=1}^{4} f \cdot \frac{N_f}{\sum_{f=1}^{4} N_f}, \tag{S12}$$

where $p(f)$ is the fraction of sequences in bin f. Note that the denominator is just the total number of sequences, $N = \sum_{f=1}^{4} N_f$, and that this average will be independent of position.

Next we need to determine the average fluorescence bin of a mutated nucleotide at each position i. Since the number of mutated nucleotides may differ at each position, we define the number of mutated nucleotides in each bin and position as $M_{f,i}$. The subscript 'f, i' is used to identify which bin f and position i are being considered. The average fluorescence bin of a mutated nucleotide can then similarly be found,

$$\langle f_{mut,i} \rangle = \sum_{f=1}^{4} f \cdot p_{mut,i}(f) = \sum_{f=1}^{4} f \cdot \frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}}, \tag{S13}$$

where in this case, $p_{mut,i}(f)$ refers to the fraction of mutated nucleotides in bin f, and at position i.

Finally, we can now calculate the average fluorescence bin shift upon mutation, which is given by the differences in Equation S13 and Equation S12,

$$\langle \Delta f_{mut,i} \rangle = \langle f_{mut,i} \rangle - \langle f \rangle == \sum_{f=1}^{4} f \cdot \left(\frac{M_{f,i}}{\sum_{f=1}^{4} M_{f,i}} - \frac{N_f}{\sum_{f=1}^{4} N_f}\right). \tag{S14}$$

Note that when we plot the fluorescence bin shift for a region where we have multiple data points (i.e. from different mutated, but overlapping regions of the DNA), we plot the average calculated value of $\langle \Delta f_{mut,i} \rangle$ from the different experiments.

We also note that it is possible to re-weight each bin by its mean fluorescence, $\widetilde{f}$ (i.e. instead of $f = 1, 2, 3, 4$, use the average fluorescence shift in arbitrary fluorescence units). Here we replace f with $\widetilde{f}$ in Equation S14. For example, under situations where different sort conditions were used across experiments, this re-normalization should allow better comparison of values across experiments. The fluorescence values for $\widetilde{f}$ can be determined by regrowing the sorted cells and measuring the mean fluorescence of each sorted cell population.

H.2 Calculation of information footprints

Another way that we analyze the data from our Sort-Seq experiments is to calculate an information footprint[29]. This allows us to identify whether there are any positions along the mutagenesis window that are informative in relating sequence S and fluorescence bin f. Said differently, an informative region would be one that if given some knowledge about the sequence, we should be able to predict

which fluorescence bin the promoter sequence might be found in. The mathematical way of implementing this intuition is to use the quantity known as the mutual information.

We can calculate the mutual information between sequence and fluorescence bin, $I(b_j, f)$, at each position i along the mutagenesis window by calculating the fraction of each nucleotide b_j (= A, C, G, T) found within each bin f. This allows us to estimate the joint probability distribution $p_i(b_j, f)$ at each position i. For example, $p_{10}(A, 2)$ would denote the probability that we observe an A in the second fluorescence bin at position i=10 along our promoter. The mutual information at each position is then defined by,

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log \left(\frac{p_i(b_j, f)}{p_i(b_j) p_i(f)} \right) \tag{S15}$$

where we have summed over all nucleotides and the N_f fluorescent bins that the sequences were found in. There is also a finite sample correction that can be applied,[34], since Equation S15 tends to overestimate the true mutual information. This is given by

$$I_i(b_j, f) = \sum_{b_j=A}^{T} \sum_{f=1}^{N_f} p_i(b_j, f) \log \left(\frac{p_i(b_j, f)}{p_i(b_j) p_i(f)} \right) - \frac{(n_{b_j} - 1) \cdot (n_f - 1) \cdot log_2 e}{2 \cdot N} + \mathcal{O}(N^{-2}), \tag{S16}$$

where $n_{b_j} = 4$ is the number of nucleotides, and n_f is the number of bins that cells have been sorted into.

H.3 Inference of energy matrix models with Sort-Seq data.

In order to predict the influence of DNA sequence on binding by regulatory proteins, we use the Sort-Seq data to generate quantitative models of the sequence-dependent binding energy. Through a relationship between likelihood and mutual information, Kinney *et al.*[29,35] showed that in the large data limit it is possible to infer biophysical parameters such as the binding energies that relate the interaction between proteins and DNA sequence. In this section we describe in detail the approach used to infer energy matrices from our Sort-Seq data using Markov Chain Monte Carlo (MCMC). A full discussion of MCMC is beyond the scope of this work, but we point the interested reader to further details regarding inference using mutual information in work from Kinney *et al.*[29,33,36]. We also stress that while we make extensive use of linear energy matrix models, the inference procedure is in no way limited to such models and can be extended to allow, for example, epistatic effects through the addition of other parameters. The simple linear models, however, provide us with a useful starting point to gain insight and describe the protein-DNA interaction.

We begin with a summary of the procedure used to infer an energy matrix model using MCMC, and use the RNAP binding site of the *relB* promoter as an example. The inference was performed using the MPAthic software[37]. A general schematic of the procedure is shown in Supplemental Fig. S10. More specific details are then discussed in the following subsections. First we must initialize a 4xL set of energy parameters, $\Theta = \{\theta_{i,j}\}$, for a binding site of length L and four base pairs (see Supplemental Fig. S10, part 1). We begin by randomly selecting parameter values for our energy matrix with which to initialize the MCMC. Here we select values from a normal distribution centered at zero with variance equal to 1, although this choice does not appear to be too critical and rather, just provides us with a starting point for our MCMC chain. Using this energy matrix we then estimate the mutual information between the binned sequences and the associated set of energy model predictions. As shown in Supplemental Fig. S10, part 2, initially the energy matrix will be of little value in describing the observed sequence data since it was randomly chosen. This is shown by the almost uniform joint probability distribution and low mutual information in Supplemental Fig. S10A, and Fig. S10B.

We now begin the MCMC by perturbing the energy matrix parameters using the Metropolis-Hastings algorithm with the PyMC package in Python[38] (within the MPAthic software[37]). After each step of the chain, we re-calculate the mutual information between the data and new model

predictions, which allows us to calculate how well this new set of energy matrix parameters describe the data. Dependent on whether the new energy matrix parameters lead to an improvement in mutual information, these new parameters are either retained or discarded and the process is repeated (again, according to the Metropolis-Hastings algorithm[38]). As discussed in Supplemental Section H.3.1, we also renormalize the matrix entries to constrain certain gauge freedoms after each iteration.

After a sufficient number of steps, and assuming that a model exists that can describe the Sort-Seq data, we will arrive at a model whose joint probability distribution between model predictions and binned sequences show a clear correlation. This is shown by the joint probability distribution in Supplemental Fig. S10C, as well as the plateau in the mutual information trace in Supplemental Fig. S10A, since changes to the energy matrix parameters are unable to increase the mutual information any further. In this first portion of MCMC we have performed many samplings to reach a high probability region where the energy matrix will be more representative of the distribution we are sampling from. This first step is usually referred to as the 'burn-in' period[38] and allows us to begin sampling from the distribution, $p(\Theta|data)$ (defined in Supplemental H.3.2), that describes the distribution of energy matrix model parameters.

Finally, now that we are sampling from the desired distribution, we can estimate energy matrix parameters just by sampling this distribution many times. This brings us to part 3 of Supplemental Fig. S10. While the mutual information no longer shows a substantial change, the parameters of the energy matrix are continuing to be perturbed following the Metropolis-Hastings algorithm, and according to the distribution $p(\Theta|data)$. We can now estimate each entry in the energy matrix by taking the arithmetic mean of the matrix parameters across all the sampling steps. This is shown by a set of contour plots and marginalized distributions for the binding energy parameters from column five of the RNAP energy matrix (Fig. S10D). To ensure that multiple energy minima were not present in this energy landscape, we repeated the inference procedure 20 times and used the average across all appropriate MCMC chains to estimate the energy matrix parameters. The calculated mutual information will be indifferent the particular sign of the energy matrix and adjust the energy matrices such that the wild-type sequence has a negative predicted binding energy and check that energy predictions from the energy matrices from each MCMC are correlated (keeping energy matrices that provide a Pearson correlation coefficient of 0.85 or greater across model predictions). Note that for inference of parameters using thermodynamic models, separate from these energy weight matrices, we did find the presence of multiple minima and apply a parallel tempering MCMC procedure to properly sample these distributions (see Supplemental H.3.4 for more detail).

Using the schematic in Supplemental Fig. S10 as our guide, the sub-sections that follow expand on the details introduced here to perform this inference procedure. In particular, we begin by describing the linear energy matrix model (Supplemental Section H.3.1). We then outline the Bayesian approach taken to formally write the posterior distribution, $p(\Theta|data)$, that provides us with a relationship between the energy matrix parameters and observed sequence data (Supplemental Section H.3.2). When sampling this distribution we need to estimate mutual information at each iteration of the MCMC sampling procedure, and describe how to calculate it in Supplemental Section H.3.3.

H.3.1 Linear energy matrix models are used to describe DNA-protein interaction.

We begin by outlining the linear energy matrix model shown in Fig. S10A that describes the binding interaction between the DNA and a DNA-binding protein. We treat each base pair position j along a binding site as contributing a certain amount to the binding energy, where the total binding energy is then the sum of the contributions from all base pairs. Mathematically the energy matrix model is described by a 4xL matrix, Θ, consisting of energy parameters $\{\theta_{ij}\}$. Here each column j of matrix parameters will represent the energies for each nucleotide $i = A, C, G$, or T ($= 1, 2, 3$, or 4) associated with position j of the binding site. For example, $\theta_{2,3}$ represents the energy parameter for nucleotide C at position 3. To make our computation of binding energies more convenient, we also represent our DNA sequence as another matrix, S, having identical dimensions, 4xL. This matrix consists of parameters $\{s_{ij}\}$, where the ij^{th} entry again corresponds to the the nucleotide identity i and sequence position j. Each parameter will have a value of 1 if it corresponds to the sequence's nucleotide identity at position j,

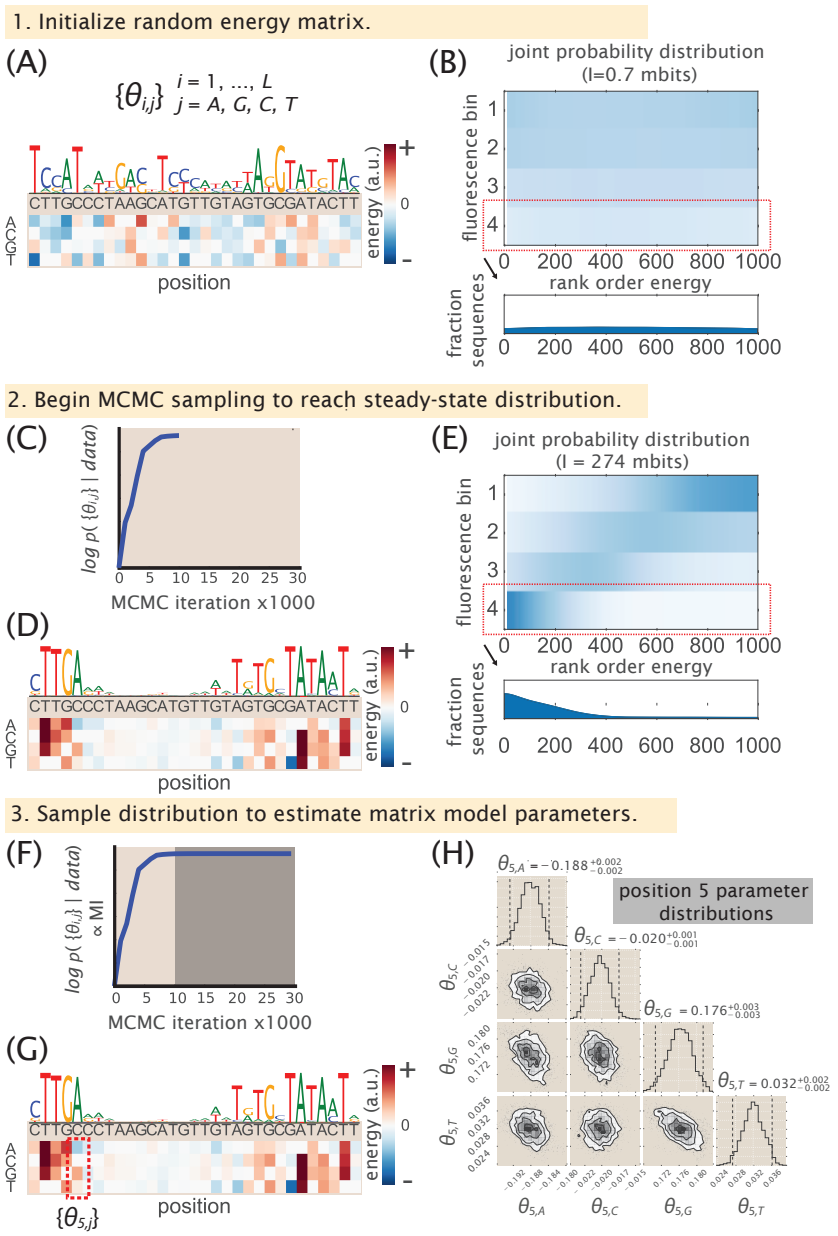

Figure S10. Schematic of the inference procedure used to determine energy matrices from Sort-Seq data using Markov Chain Monte Carlo. 1. To begin the inference of a set of 4xL model parameters, $\{\theta_{ij}\}$, are chosen from a normal distribution. (A) Example set of parameters used to initialize the MCMC sampling. Matrix entries are first normalized such that energy predictions have mean of zero and standard deviation of one. For plotting energy matrices, each column has been shifted such that the wild-type sequence has zero energy. The associated sequence logo is shown above the energy matrix. (B) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (A), using all sequences in the *rel* promoter data set. The bottom plot shows, the histogram of rank ordered predictions of only bin four, corresponding to the red boxed region, which is nearly uniform due to the randomly chosen matrix entries used to predict energies from each sequence. Since the matrix parameters were randomly chosen, the nearly uniform distribution results in low mutual information (0.7 mbits, where 1 mbit $= 10^{-3}$ bits) between fluorescence bin and rank order energy predictions. *(Caption continued on next page)*

Figure S10. *(continued from previous page)* 2. MCMC sampling of the energy matrix model is performed using the Sort-Seq data associated with the *rel* RNAP binding site. (C) The log posterior, Eq. (S20), is plotted for the first 1000 iterations and corresponds to the 'burn-in' period. The log posterior is proportional to the mutual information between fluorescent bin and rank order energy predictions (see Appendix H.3.3). During each sampling iteration, the parameters will be retained or discarded with some probability given by the the Metropolis-Hasting algorithm[38]. (D) The energy matrix and sequence logos are shown using the set of parameters at the 1000^{th} iteration. (E) Estimated joint probability distribution between fluorescence bin and rank order energy predictions using the energy matrix in (D). The energy matrix provides energy predictions for each sequence that clearly distributes across the sorted bins and results in much higher mutual information (274 mbits). 3. Finally, matrix parameters are estimated by continuing to sample the posterior distribution many more times and determined from a weighted average of these samples. (F) The log posterior is plotted for the entire set of MCMC iterations. The sampled model parameters during the shaded region are used to estimation each matrix entry. (G) The mean energy matrix entries from these samples are plotted. (H) Contour plots and marginalized distributions summarize the sampled values for each of the four parameters at position five of the RNAP energy matrix. Note that entries in (G) have been shifted such that the wild-type nucleotide has zero energy.

and a value of 0 otherwise. For example, for a sequence with a C at position $j = 4$, the entry $s_{2,4} = 1$ and $s_{i=1,3,4,j=4} = 0$. The binding energy, E, of any sequence, S, will then be given by

$$E = \sum_{i=A}^{T} \sum_{j=1}^{L} \theta_{ij} \cdot s_{ij}. \tag{S17}$$

One aspect we have not considered thus far is the scale of the energy parameter. When considering binding between between DNA and a DNA-binding protein, a statistical mechanical approach would suggest that the probability of such an event occurring will be given by the Boltzmann factor, $e^{-\varepsilon_s/(k_B T)}$[16]. Here ε_s is the binding energy that describes this interaction in absolute energy units (e.g. units of $k_B T$; 1 kcal/mol = 1.62 $k_B T$ at 37°C), k_B is the Boltzmann constant, and T is temperature. In relation to the binding energy, E, described by our Equation S17 above, $\varepsilon_s = A \cdot E + B$, where the constant A scales the energy matrix into absolute energy units, while B provides an additive shift that depends on the choice of reference energy. Here, the matrix entries that are used to calculate E are 'gauge fixed' such that the mean energy in each column is set to zero and the matrix norm (or inner product) has a value of 1. Note however that when plotting each energy matrix we find it useful to shift the energy in each column such that the wild-type sequence has zero energy.

When fitting the data to a model of the form $e^{-\varepsilon_s/(k_B T)}$, the fitting procedure is unable to determine the scale factors A and B noted above. For example, in most instances we report energy values in arbitrary units. This is consequence of the fitting procedure, where in the absence of a specific thermodynamic model, there remain some scale parameters that cannot be determined[29]. This parameter insensitivity has been termed 'diffeomorphic modes' and is discussed at length in other work[36]. One especially interesting aspect of this is that when considering biophysical models of regulation, diffeomorphic modes often disappear and make it possible to infer parameters that were not accessible by fitting simpler models. For the cases of repression by PurR at the *purT* promoter, or activation by CRP at the *dgoR* promoter, this allowed us to estimate binding energy in absolute energy. We discuss this further in Supplemental Section H.3.4.

H.3.2 Probability distribution relating energy matrix model parameters to the Sort-Seq data.

Given our FACS-sorted sequence data, we want to find the set of energy matrix parameters that best describe the distribution of sequences across our fluorescence bins (i.e. parameters that provide binding energy predictions that describe the data as shown in Supplemental Fig. S10C). To perform this

inference we take a Bayesian approach in our analysis, and as mentioned earlier, rely on MCMC to sample from the complex distribution relating our energy matrix parameters to the sequence data. While a full discussion of Bayesian analysis is outside the scope of this section, the book, Data Analysis by Sivia and Skilling[39], and online material available from the Caltech course, *BE/Bi 103: Data analysis in the biological sciences*, taught by Justin Bois (http://bois.caltech.edu/teaching.html), are excellent resources.

Formally, we want to find the set of energy matrix parameters that maximize the probability distribution of our energy predictions (through our energy matrix model) given our Sort-Seq sequence data, $p(E|\{S,f\})$, where $\{S,f\}$ refers to our array of N sequences S and the bin f where they were found (referred to as the 'data' in the initial summary of the inference procedure). x_S is the binding energy as defined in Equation S17. From Bayes' theorem, we can re-write this distribution as,

$$p(E|\{S,f\}) = \frac{p(\{S,f\}|E)p(E)}{p(\{S,f\})} \propto p(\{S,f\}|E)p(E), \tag{S18}$$

where the term $p(\{S,f\}|E)$ is called the likelihood, and $p(E)$ is known as the prior and encompasses our prior knowledge on the energy matrix parameters. The denominator $p(\{S,f\})$ is known as the marginalized likelihood and acts as a normalization factor, but is unimportant for our inference.

To proceed we follow the approach of Kinney *et al.*[29,35]. We assume a uniform prior over the energy matrix model parameters. In addition, we also assume our sequence measurements are independent. The second assumption allows us to write $p(\{S,f\}|E)$ as the product of probabilities across all sequences contained within our data set, $p(\{S,f\}|E) = \prod_{s=1}^{N} p((S_i,f_i)|E)$. This is also referred to as the error model since by relating the binned sequence data to binding energy, it must also encompass the additional noise sources from our experiment that actually led to our array of sequence data. Noise sources that might influence this include the sensitivity of the FACS GFP measurements, and the rate of mis-sorting events. Expression variability due to stochastic gene expression, differences in cell size, and plasmid copy number fluctuations are also likely to contribute. However, since these are not known exactly, Kinney *et al.* computed the likelihood by averaging over an ensemble of all possible error models. Using a uniform prior over the possible error models they found,

$$p(\{S,f\}|E) = \left\langle \prod_{s=1}^{N} p((S_i,f_i)|E) \right\rangle_{\text{all possible } p(S_i,f_i|E)} = C \cdot 2^{N\cdot(I(f,E)+\Delta)}, \tag{S19}$$

where N is the total number of sequences considered, $I(f,E)$ is the mutual information between the observed fluorescence bins and binding energies predicted by the energy matrix for all the sequences, and C is a constant of integration that will be unimportant to us. Here, Δ is a small correction that goes to zero as N goes to infinity[35]. Inserting Equation S19 into Equation S18, we can write,

$$p(E|\{S,f\}) \propto 2^{N\cdot I(f,E)}. \tag{S20}$$

Here we have assumed that N is sufficiently large so that the prior (which does not scale with N), as well as the Δ term in Equation S19 can be ignored. To reiterate in reference to our MCMC procedure (shown in Supplemental Fig. S10), this is the probability distribution that we are sampling from to find the set of energy matrix parameters that describe our sorted sequence data set. The mutual information values shown in the plots of Fig. S10C, F (mutual information traces in part 2 and 3) are reflected by our choice of energy matrix parameters. MCMC enables us to sample from the distribution and essentially find the set of matrix parameters that maximize this mutual information. In the next section we continue by describing how we estimate mutual information.

H.3.3 Estimating mutual information using the energy model predictions.

In the last section we found that the energy matrix parameters should be related to the data through Equation S20. By performing many samples from this distribution using MCMC, it is possible to estimate the most probable energy matrix parameters, $\theta_{i,j}$, that make up our energy matrix. Here we

consider how to estimate the mutual information term in Equation S20 needed for our calculation. While a non-trivial problem in general, the following approach appears to work well in practice. In this case the fluorescence bins, f, are discrete variables while our binding energies, E, are continuous, with the mutual information given by,

$$I(f,E) = \int_{E=-\infty}^{E=\infty} dE \sum_{f} p(f,E) \log_2 \frac{p(f,E)}{p(E) \cdot p(f)}. \tag{S21}$$

In our sequence data set, we can easily estimate $p(f)$ by counting the number of sequences in each fluorescence bin. However, we do not have direct access to the probability distribution $p(E)$ *a priori*.

To proceed, we further bin our N sequences into 1000 bins, by rank ordering them by their associated binding energy predictions (using the energy matrix of the current MCMC step). This provides us with an estimate of the probability distribution in binding energy across our sequences. Specifically, this is shown for fluorescence bin 4 in Supplemental Fig. S10B and E. While this is not a direct estimate of $p(E)$, we invoke the fact that the mutual information will be invariant under monotonic transformations ($I(f,E) = I(f,z_s)$)[29]. Hence, instead of calculating $I(f,E)$, we instead calculate $I(f,z_s)$, where z_s is instead the ranked ordering of the N sequences.

In order to calculate the mutual information we now construct a 2-d histogram (joint distribution) by binning the rank ordered energy predictions into $z_s = 1$ to 1000 bins across each of the different fluorescence bins. We define this by the frequency matrix $F(f,z_s)$, and from our finite data set, use kernel density estimation with a kernel width equal to 4% to estimate the joint distribution. This is what is plotted in Supplemental Fig. S10B, and E, where the mutual information is then calculated as,

$$I(f,z_s)_{smooth} = \sum_{z_s=1}^{1000} \sum_{f} F(f,z_s) \log_2 \frac{F(f,z_s)}{F(z_s) \cdot F(f)}. \tag{S22}$$

H.3.4 Inference of thermodynamic model parameters using parallel tempering Markov chain Monte Carlo (MCMC).

So far, we have applied MCMC using an error-model-averaged likelihood to infer the parameters of an energy matrix. One limit initially observed by Kinney *et al.*[29] was an inability of the fitting procedure to constrain certain parameters (due to free diffeomorphic modes, noted earlier). Interestingly however, it was found that certain diffeomorphic modes often disappear when fitting the Sort-Seq data to non-linear models. For a thorough discussion of diffeomorphic modes refer to the work of Kinney *et al.*[40]. We applied this strategy in several of our data sets from the *purT*, *dgoR*, and *xylE*, where specific thermodynamic models appeared appropriate. Here we briefly outline the models used and the main results from our MCMC analysis.

We begin with the *purT* promoter. Here we identified an RNAP binding site that is repressed by PurR, which binds between the -10 and -35 RNAP sites. Given the presence of only these two binding sites, we modeled the promoter as having a simple repression architecture[16]. Some additional complexity arises due to the presence of other PurR binding sites on the genome, and the allosteric dependence of a purine metabolite for co-repression. Following the approach of Weinert *et al.*[15], this can be quantitatively described by,

$$P_{bound} = \frac{\lambda_p e^{-\beta\varepsilon_p}}{1 + \lambda_p e^{-\beta\varepsilon_p} + \lambda_r e^{-\beta\varepsilon_r}}. \tag{S23}$$

Here λ_p and λ_r represent the fugacity, which describes the relative availability of RNAP and PurR, respectively, to bind their binding sites. These parameters depend on the concentration of each protein (through their chemical potentials), and for PurR, will also depend on its allosteric state. ε_p and ε_r represent the binding energies of RNAP and PurR to their binding sites, respectively.

As noted in Supplemental Section H.3.1, we can also describe each binding energy through the gauge-fixed energy matrix prediction (see Section H.3.1), which is multiplied by a scale factor and

additive shift (e.g. $\varepsilon_r = A_r \cdot x_r + B_r$, where A_r is the scale factor, x_r is the energy matrix prediction, and B_r is the additive shift). To being fitting to the model described by Equation S23, we first inferred the energy matrices for RNAP and PurR following the MCMC procedure noted above. We then performed a second MCMC to fit the remaining thermodynamic parameters. In this second MCMC we sampled using error-model-averaged likelihood against the posterior $p(P_{bound}|\{S, f\})$. This allowed us to infer the following parameters: $A_r = -11.55^{+0.2}_{-0.5} k_B T$, $\lambda_r e^{-\beta B_r} = e^{0.64^{+0.1}_{-0.3}}$, and $A_p = 2.4^{+0.4}_{-0.1} k_B T$, where A_p is the RNAP scale factor. Here the error bars represent the median of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions. Note that in this second MCMC, we performed parallel tempering MCMC (using the PTSampler in package emcee,[41]) to better sample the posterior distributions of our thermodynamic parameters (see supplemental of Kinney *et al*, 2010).

Next we consider the *dgoR* promoter. While we found the promoter to be quite complex, here we use data from the JK10 strain (see Section G.5) where activation by CRP appeared to dominate transcription. Here we apply the model used by Kinney *et al.*[29], which consists of a binding site for RNAP and CRP, but also includes an interaction energy between these two proteins. Again using fugacity terms to describe the availability of each protein, this will be given by,

$$P_{bound} = \frac{\lambda_p e^{-\beta\varepsilon_p} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p + \varepsilon_a + \varepsilon_i)}}{1 + \lambda_p e^{-\beta\varepsilon_p} + \lambda_a e^{-\beta\varepsilon_a} + \lambda_a \cdot \lambda_p e^{-\beta(\varepsilon_p + \varepsilon_a + \varepsilon_i)}}. \tag{S24}$$

In this architecture we have the fugacity λ_a for the activator CRP and its binding energy to the binding site, ε_a. In addition, there is an additional energy term ε_i that describes the interaction between RNAP and CRP. Again, we can write $\varepsilon_p = A_p \cdot x_p + B_p$. We can also write the CRP binding energy as $\varepsilon_a = A_a \cdot x_a + B_a$, where similarly, A_a is the scale factor, x_a is the gauge-fixed energy prediction, and B_a is an additive shift. Using parallel tempering MCMC to sample $p(P_{bound}|\{S, f\})$, we obtained the following values: $\varepsilon_i = -7.3^{+1.9}_{-1.4} k_B T$, $A_a = -13.6^{+2.6}_{-2.2} k_B T$, $\lambda_a e^{-\beta B_a} = e^{-1.89^{+0.4}_{-0.6}}$, and $A_p = -12.7^{+3.4}_{-2.8} k_B T$. As with the *purT* case above, the error bars represent the median of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions.

Lastly we consider the *xylE* promoter. This promoter contains two XylR sites which are likely bound as a dimer[42]. There is also a CRP site directly upstream of the xylR sites. The binding signature of CRP is only observed for the right half of the binding site, implying the left half of the protein does not make as significant DNA contact. Since CRP still has a powerful impact on gene expression, it suggests that there is a cooperative interaction between xylR and the weak CRP site. The short distance between the xylR sites and the RNAP also suggests that there is a direct interaction between the xylR sites and the RNAP. In addition, there is also a spacing between the RNAP polymerase and the CRP site of 35 bp (approximately three helical turns of the DNA). For this spacer length in the *lac* promoter there is a expected to be a significant interaction energy even in the absence of XylR[43,44]. A thermodynamic model of RNAP polymerase binding probability for this architecture will be

$$P_{bound} = \frac{f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}{g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i})}, \tag{S25}$$

where

$$\begin{aligned} f(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= \lambda_p e^{-\beta\varepsilon_p} + \lambda_p\lambda_x e^{-\beta(\varepsilon_p + \varepsilon_x + \varepsilon_{x_i})} \\ &\quad + \lambda_p\lambda_c e^{-\beta(\varepsilon_p + \varepsilon_c + \varepsilon_{c_i})} + \lambda_p\lambda_c\lambda_x e^{-\beta(\varepsilon_p + \varepsilon_x + \varepsilon_c + \varepsilon_{c_i} + \varepsilon_{x_i} + \varepsilon_{cx_i})} \end{aligned} \tag{S26}$$

$$\begin{aligned} g(\lambda_p, \lambda_x, \lambda_c, \varepsilon_p, \varepsilon_x, \varepsilon_c, \varepsilon_{x_i}, \varepsilon_{c_i}, \varepsilon_{cx_i}) &= 1 + \lambda_x e^{-\beta\varepsilon_x} + \lambda_c e^{-\beta\varepsilon_c} + \lambda_x\lambda_c e^{-\beta(\varepsilon_x + \varepsilon_c + \varepsilon_{cx_i})} \\ &\quad + \lambda_p e^{-\beta\varepsilon_p} + \lambda_p\lambda_x e^{-\beta(\varepsilon_p + \varepsilon_x + \varepsilon_{x_i})} \\ &\quad + \lambda_p\lambda_c e^{-\beta(\varepsilon_p + \varepsilon_c + \varepsilon_{c_i})} \\ &\quad + \lambda_p\lambda_c\lambda_x e^{-\beta(\varepsilon_p + \varepsilon_x + \varepsilon_c + \varepsilon_{c_i} + \varepsilon_{x_i} + \varepsilon_{cx_i})} \end{aligned} \tag{S27}$$

Here, the λ_x and ε_x terms mark the fugacity and binding energy of XylR respectively. The λ_c and ε_c represent the fugacity and binding energy of CRP, and λ_p and ε_p do the same for RNAP. The terms ε_{x_i}, ε_{c_i}, and ε_{cx_i} are interaction terms between XylR and RNAP, CRP and RNAP, and CRP and XylR, respectively.

Due to the position of the library windows (with a 60 bp window containing the two XylR binding sites, but only partial binding sites for CRP and RNAP), we were unable to fit this model to the data. The fitting procedure requires sequences with mutations throughout the multiple binding sites and further experimentation will be needed to fit and characterize the proposed model further.

I Extended experimental details

In this section we provide additional details to describe the specifics of the work flow. In general, an experiment is begun by constructing the mutated promoter libraries for Sort-Seq. Next transform libraries into cells and use FACS to sort by fluorescence. Using putative regulatory sequences identified by Sort-seq, we perform DNA affinity chromatography and mass spectrometry, which is necessary to identify the transcription factors that bind to these putative binding sites.

I.1 *E. coli* strain construction

Here we describe the approach used to generate these deletion strains. Briefly, an overnight culture of MG1655 containing the plasmid pSIM6 was diluted 1:100 in 50 ml LB media and grown to an OD600 of ≈ 0.4 at 30°C. The culture was immediately placed in a water bath shaker at 43°C for 15 minutes and then cooled in an ice bath for 10 minutes. Cells were then spun down for 10 minutes (4,000 g, 4°C) and resuspended on ice in 50 ml of chilled water. This was repeated three times before resuspending in 200 μL of chilled water to generate competent cells. Homologous primer extension sequences for the appropriate gene were obtained from Baba *et al.*[45] and used to generate linear DNA containing a kanamycin resistance gene insert by PCR, which contained homology for the region on the chromosome to be deleted[46]. Electroporation of the competent cells was performed using 1 μL purified PCR product (about 100 ng DNA), mixed with 50 μL cells. Cells were immediately resuspended in 750 μL SOC media and placed on a shaker at 30°C for outgrowth, for 90-120 minutes. Cells were then plated on an LB-agar plate containing kanamycin (30 μg/ml) and grown overnight at 30°C. The deletions were confirmed by both colony PCR and sequencing. After confirmation, the deletion was transferred to a clean MG1655 strain through P1 transduction and selection on kanamycin. In the case of the lysine auxotrophic strain, we also confirmed deletion of *lysA* by checking that the cells were unable to grow in M9 minimal media unless lysine was supplemented (40 μg/ml).

To generate strains with different LacI tetramer copy numbers per cell (associated with data in Fig. S5C), the LacI constructs from Garcia *et al.*[21] were P1 transduced into the $\Delta lacIZYA$ strain (integrated at the *ybcN* locus).

I.2 Sort-Seq library construction

Mutagenized single-stranded oligonucleotide pools were purchased from Integrated DNA Technologies (Coralville, IA), with a target mutation rate of 9%. Note that in the case of the *lacZ* promoter, the library is identical to that used in the experiments of Razo-Mejia *et al.*[47], and had a mutation rate of approximately 3%.

Note that to assemble PCR amplified library inserts with the plasmid backbone, we used Gibson assembly[48] (New England Biolabs, MA, USA). Otherwise, we follow the approach of Kinney *et al.* and amplify the backbone using a template plasmid containing the toxic gene *ccdB* (located where the library was to be inserted). This helped ensure that no template plasmid was propagated into the final plasmid library (see methods in reference[29] for more detail).

For each library construction, 40 ng of insert and 50 ng of backbone were combined in a 20 μL Gibson assembly reaction. To achieve high transformation efficiency, reaction buffer components from

the Gibson Assembly reaction were removed by drop dialysis and cells were transformed by electroporation of freshly prepared cells. Following an initial outgrowth in 1 mL of SOC media, cells were diluted into 50 mL of LB media and grown overnight under kanamycin selection. Transformation typically yielded $10^6 - 10^7$ colonies as assessed by plating 100 μL of cells diluted 1:10^4 onto an LB plate containing kanamycin.

I.3 Sort-Seq experiments

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

A Beckman Coulter MoFlo XDP cell sorter was used for all Sort-Seq experiments. Prior to sorting, we would obtain fluorescence histograms using between 200,000 and 500,000 cell events per culture. These histograms were used to set the four binning gates, which each covered $\sim 15\%$ of the histogram. During sorting of each library, 500,000 cells were collected into each of the four bins. Finally, sorted cells were re-grown overnight in 10 ml of LB media, under kanamycin selection.

I.4 Sort-Seq sequencing

The plasmid from cells in each bin were miniprepped following overnight growth (Qiagen, Germany). PCR was used to amplify the mutated region from each plasmid for Illumina sequencing, adding Illumina adapter sequences and custom barcode sequences. Sequencing was performed by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) or NGX Bio (NextSeq sequencer; San Fransisco, CA). Single-end 100bp or paired-end 150bp flow cells were used, with a target read count of about 500,000 sequences per library bin. Joining of paired-end reads was performed with the FLASH tool[49]. For quality filtering, we collected sequences whose barcodes had a PHRED score greater than 20 at each position. Some libraries also contained non-mutagenized regions, and sequences that did not contain the expected sequence were excluded from our analysis. The total number of useful reads available to produce expression shift plots, energy weight matrices, and sequence logos from each Sort-Seq experiment generally ranged between 300,000 to 2,000,000 reads. Energy matrices were inferred using Bayesian parameter estimation with an error-model-averaged likelihood as previously described[29,36], using the MPAthic software[37]. A more detailed description of the data analysis procedures is available in Section H.

I.5 DNA affinity chromatography and mass spectrometry

Here we provide additional details on SILAC incorporation, preparation of DNA-tethered magnetic beads, and the LC-MS/MS method.

I.5.1 Lysate preparation and SILAC incorporation

SILAC labeling[50–52] was implemented by growing cells in either the stable isotopic form of lysine ($^{13}C_6H_{14}{}^{15}N_2O_2$), referred to as the heavy label, or natural lysine, referred to as the light label. By differentially labeling cell lysates we were able to simultaneously quantify the abundance of protein between two DNA affinity purification samples (i.e. one using a target binding site sequence and another as a reference control). This allows us to identify whether any protein shows a preference for the target binding site sequence. Cell lysates were prepared using MG1655 $\Delta lysA$ cells. For each heavy and light labelled cells, 500 ml M9 minimal media was inoculated 1:5,000 with an overnight LB culture of $\Delta lysA$ cells, and grown to an OD600 of ≈ 0.6 (supplemented with the appropriate lysine; 40 μg/ml).

Cultures were pelleted, and lysed using a Cell Disruptor (CF Range, Constant Systems Ltd., UK) and concentrated to ∼150 mg/ml using Amicon Ultra-15 centrifugation units (3kDa MWCO, Millipore).

To generate each lysate an overnight starter culture of cells was grown in LB media supplemented with kanamycin (30 μg/ml). An aliquot was washed twice in M9 minimal media and resuspended to an OD600 of ≈1.0. For both heavy and light labeling, 500 ml M9 minimal media was then inoculated at 1:5,000 and grown to an OD600 of ≈0.6 (supplemented with the appropriate lysine; 40 μg/ml). Cultures were pelleted using an ultracentrifuge (8,000 g, 40 minutes) at 4°C and resuspended in chilled 20 ml lysis buffer containing 1% (w/v) n-dodecyl-beta-maltoside. The pellets could also be stored at -80°C for later use. Cells were then lysed with a Cell Disruptor (CF Range, Constant Systems Ltd., UK) and following removal of debris by centrifugation, concentrated to ∼150 mg/ml using Amicon Ultra-15 centrifugation units (3kDa MWCO, Millipore). This provided about 600 μl of lysate, suitable for about six 80 μl DNA affinity purifications. Total protein concentration was assayed using the Bradford reagent (Sigma-Aldrich, St. Louis, MO). Following adjustment of protein concentration, sheared salmon sperm competitor DNA was added to the lysates (1 μg/ml; Life Technologies, Carlsbad, CA) and incubated for 10 minutes at 4°C. Finally, following centrifugation at 14,000 g to remove insoluble matter, the cell lysates were incubated for 1 hour with washed magnetic beads that contained no tethered DNA (0.5 mg beads per 100 μl lysate). Lysates were then either placed on ice or stored at 4°C prior to use.

Before performing affinity chromatography experiments, we also confirmed heavy lysine was being incorporated. Here, MG1655 $\Delta lysA$::kan cells from an overnight M9 minimal media culture were diluted 1:200 and 1:1,000, and grown in 1 ml M9 minimal media supplemented with 40 μg/ml heavy lysine. Following approximately 7 and 10 cell divisions, cells were resuspended in lysis buffer (50 mM HEPES pH 7.5, 70 mM potassium acetate, 5 mM magnesium acetate, 0.2% (w/v) n-dodecyl-beta-D-maltoside, Roche protease inhibitor cOmplete tablet) and lysed by performing 10 freeze-thaw cycles with dry ice. Cellular debris was removed by centrifugation at 14000 g at 4°C on a tabletop centrifuge. Finally cellular lysates were prepared for mass spectrometry by in-solution digestion with endoproteinase Lys-C (Promega, Madison, WI). Digestion was performed as described elsewhere[53] and labeling of the heavy isotope was confirmed by mass spectrometry measurement. In addition, we also characterized the SILAC enrichment ratio measurement by directly combining measurements from heavy and light lysates over a range from 0.1:1 to 1,000:1 heavy:light (see Section E).

I.5.2 Preparation of DNA-tethered magnetic beads

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dynabeads MyOne T1, ThermoFisher, Waltham, MA) containing tethered DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Note that single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand.

To begin preparation of tethered beads, DNA was suspended in annealing buffer (20 mM Tris-HCl, 10 mM MgCl2, 100 mM KCl) to 50 μM. Complementary strands were annealed by mixing 30 μL of the sense strand and 40 μL of the complement strand. Excess complement strand ensured all biotinylated-DNA would be in a double stranded form. Annealing was then performed using a thermocycler: 90°C for 5 minutes, gradient from 90°C to 65°C @ 0.1C /sec, incubated for 10 minutes at 65°C and allowed to return to room temperature on the thermocycler. Prior to attaching DNA, 150 μL beads were washed twice with 600 μL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and then twice with DW buffer (20 mM Tris-HC pH 8.0, 2 M NaCl, 0.5 mM EDTA[18]). Approximately 640 pmol of DNA were then diluted to 600 μL in DW Buffer and incubated with the washed beads overnight at 4°C and on a rotatory wheel. Bound DNA was measured by determining the DNA concentration before and after incubation with beads using a NanoDrop (ThermoFisher, Waltham, MA). Finally, beads were washed once with 600 μL TE buffer and three washes of 600 μL DW buffer, and resuspended in 150 μL DW buffer.

DNA affinity chromatography

DNA affinity chromatography was performed by incubating cell lysate with magnetic beads (Dynabeads MyOne T1, ThermoFisher, Waltham, MA) containing tethered DNA. The DNA was tethered through a linkage between streptavidin on the beads and biotin on the DNA. Single-stranded DNA was purchased from Integrated DNA Technologies with the biotin modification on the 5' end of the oligonucleotide sense strand. Prior to DNA affinity purification the DNA tethered beads were incubated with blocking buffer (20 mM Hepes, pH 7.9, 0.05 mg/ml BSA, 0.05 mg/ml glycogen, 0.3 M KCl, 2.5 mM DTT, 5 mg/ml polyvinylpyrrolidone, 0.02% (w/v) n-dodeyl-β-D-maltoside; about 1.3 ml/mg beads[18]) for one hour at 4°C for passivation. Excess blocking buffer was removed by washing the beads twice with 600 μL lysis buffer.

Cell lysates were incubated on a rotating wheel with the DNA tethered beads overnight at 4°C. Beads were recovered with a magnet and washed three times using an equivalent volume of lysis buffer. The beads were then washed once more, but with NEB Buffer 3.1 (New England Biolabs, MA, USA). Both purifications (with the target DNA and reference control) were combined by resuspending in 50 μL NEB Buffer 3.1, and 10 μl of the restriction enzyme PstI (100,000 units/ml, New England Biolabs) was added and incubated for 1.5 hours at 25°C. PstI cleaves the sequence CTGCAG, which was included between the biotin label and binding site sequence, allowing the DNA to be released from the magnetic beads. The beads were then removed and the samples prepared for mass spectrometry by in-gel digestion with endoproteinase Lys-C.

Note that in general, proteins were purified from a heavy lysate using DNA containing the target binding site sequence, while devoting the light lysate to a control DNA sequence. However, for our LacI and RelBE experiments, we also performed the alternative scenario, using the target sequence with the light lysate, and did not observe notable differences.

In-gel digestion of purified protein samples

Protein samples were diluted with 4x SDS-PAGE sample buffer and incubated for five minutes at 95°C and loaded on a SDS-PAGE gel (Any kD Mini-PROTEAN TGX Precast Protein Gels, 10-well , 50 µl; BioRad, CA, USA). Electrophoresis was performed for 45-55 minutes (200V) to provide 1-D size separation, and stained using the Colloidal Blue Staining Kit (ThermoFisher Scientific, MA, USA) for visualization. Destaining was performed with 100 mM ammonium bicarbonate, and the gel was cut into four sections, each of which was cut into roughly 1 mm pieces for in-gel digestion. The gel pieces were reduced, alkylated, and digested by endoproteinase Lys-C overnight at 37°C. This enzymatically cleaves proteins after lysine residues and is necessary for determining whether detected peptides are from the light or heavy lysine labeled purification. Digested peptides were then extracted from the gel and lyophilized. The peptide samples were further purified using StageTips to remove residual salts[54] and re-suspended in 0.2% formic acid.

I.5.3 LC-MS/MS method details

Liquid chromatography tandem-mass spectrometry (LC-MS/MS) experiments were carried out as previously described[55].

The LacI target purification experiments were performed on a nanoflow LC system, EASY-nLC II coupled to a hybrid linear ion trap Orbitrap Classic mass spectrometer equipped with a Nanospray Flex Ion Source (Thermo Fisher Scientific). The in-gel digested peptides were directly loaded at a flow rate of 500 nl/min onto a 16-cm analytical HPLC column (75 μm ID) packed in-house with ReproSil-Pur C18AQ 3 μm resin (120 Å pore size, Dr. Maisch, Ammerbuch, Germany). The column was enclosed in a column heater operating at 45°C. After 30 min of loading time, the peptides were separated in a solvent gradient at a flow rate of 350 nl/min. The gradient was as follows: 0–30% B (80 min), and 100% B (10 min). The solvent A consisted of 97.8% H2O, 2% ACN, and 0.2% formic acid and solvent B consisted of 19.8% H2O, 80% ACN, and 0.2% formic acid. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1600) in the Orbitrap (resolution

100,000) and subsequent 15 CID MS/MS scans (Top 15 method) in the linear ion trap. Collision induced dissociation (CID) was performed at normalized collision energy of 35% and 30 msec of activation time.

All other measurements were performed on a hybrid ion trap-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific), which provided greater detection sensitivity and other fragmentation techniques as described. The Orbitrap was operated in data-dependent acquisition mode to automatically alternate between a full scan (m/z=400–1,800) in the Orbitrap (resolution 120,000) and subsequent 5 MS/MS scans also acquired in Orbitrap with 15,000 resolution. The MS/MS spectra were acquired for the top 5 ions alternating between higher collision dissociation (HCD) and electron transfer dissociation (ETD) fragmentations that are well suited for higher charge peptides. Higher collision dissociation was performed at a normalized collision energy of 30% and electron transfer dissociation reaction time was set to 100 msec. The analytical column for this instrument was a PicoFrit column (New Objective, Woburn, MA) packed in house with ReproSil-Pur C18AQ 1.9 μm resin (120Å pore size, Dr. Maisch, Ammerbuch, Germany) and the column was heated to 60°C. The peptides were separated either with a 90 or 60 min gradient (0-30% B in 90 min or 0-30% B in 60 min) at a flow rate of 220 nL/min.

I.5.4 Mass spectrometry data processing

Thermo RAW files were processed using MaxQuant (v. 1.5.3.30)[56,57]. Spectra were searched against the UniProt *E. coli* K-12 database (4318 sequences) as well as a contaminant database (256 sequences). Additional details are provided in the supplemental methods. Precursor ion mass tolerance was 4.5 ppm after recalibration by MaxQuant. Fragment ion mass tolerance was 20 ppm for high-resolution HCD and ETD spectra, and 0.5 Da for low-resolution CID spectra. Variable modifications included oxidation of methionine and protein N-terminal acetylation. Carboxyamidomethylation of cysteine was specified as a fixed modification. LysC was specified as the digestion enzyme and up to two missed cleavages were allowed. A decoy database was generated by MaxQuant and used to set a score threshold so that the false discovery rate was less than 1% at both the peptide and protein level. For all experiments match between runs and re- quantify were enabled. One evidence ratio per replicate per protein was required for quantitation.

To calculate the overall protein ratio, the non-normalized protein replicate ratios were log transformed and then shifted so that the median protein log ratio within each replicate was zero (i.e., the median protein ratio was 1:1). The overall experimental log ratio was then calculated from the average of the replicate ratios. Proteins were considered if they were known to be transcription factors, or predicted to bind DNA (using gene ontology term GO:0003677, for DNA-binding in BioCyc).

References

[1] Schmidt, A. *et al.* The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol* **34**, 104–111 (2016).

[2] Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796–804 (2012).

[3] Cho, B.-K. *et al.* The PurR regulon in *Escherichia coli* K-12 MG1655. *Nucleic Acids Res* **39**, 6456–6464 (2011).

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

[5] Irani, M. H., Orosz, L. & Adhya, S. A control element within a structural gene: The *gal* operon of *Escherichia coli*. *Cell* **32**, 783–788 (1983).

[6] Semsey, S., Krishna, S., Sneppen, K. & Adhya, S. Signal integration in the galactose network of *Escherichia coli*. *Mol Microbiol* **65**, 465–476 (2007).

[7] Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097–6100 (1990).

[8] Nishida, K., Frith, M. C. & Nakai, K. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res* **37**, 939–944 (2009).

[9] Xia, X. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. *Scientifica* **2012**, 1–15 (2012).

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

[11] Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. *J Mol Biol* **193**, 723–743 (1987).

[12] Lässig, M. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinform* **8**, S7–21 (2007).

[13] Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**, 415–431 (1986).

[14] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* **8**, R24 (2007).

[15] Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R. & Kegel, W. K. Scaling of Gene Expression with Transcription-Factor Fugacity. *Phys Rev Lett* **113**, 258101 (2014).

[16] Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**, 116–124 (2005).

[17] Moran, U., Phillips, R. & Milo, R. SnapShot: Key Numbers in Biology. *Cell* **141**, 1262–1262.e1 (2010).

[18] Mittler, G., Butter, F. & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res* **19**, 284–293 (2009).

[19] Ong, S.-E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc* **1**, 2650–2660 (2007).

[20] Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* **41**, D605–D612 (2013).

[21] Garcia, H. G. & Phillips, R. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci USA* **108**, 12173–12178 (2011).

[22] Rydenfelt, M., Garcia, H. G., Cox, R. S., III & Phillips, R. The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS One* **9**, e114347 (2014).

[23] Bonde, M. T. *et al.* Direct Mutagenesis of Thousands of Genomic Targets Using Microarray-Derived Oligonucleotides. *ACS Synth Biol* **4**, 17–22 (2015).

[24] Rohlhill, J., Sandoval, N. R. & Papoutsakis, E. T. Sort-Seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* Growth on methanol. *ACS Synth Biol* **6**, 1584–1595 (2017).

[25] Zhang, H., Susanto, T. T., Wan, Y. & Chen, S. L. Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* **113**, 4182–4187 (2016).

[26] Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).

[27] Wade, J. T., Reppas, N. B., Church, G. M. & Struhl, K. Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes Dev* **19**, 2619–2630 (2005).

[28] Lomba, M. R., Vasconcelos, A. T., Pacheco, A. B. F. & Almeida, D. F. Identification of *yebG* as a DNA damage-inducible *Escherichia coli* gene. *FEMS Microbiol Ecol* **156**, 119–122 (1997).

[29] Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* **107**, 9158–9163 (2010).

[30] Cooper, R. A. The utilisation of D-galactonate and D-2-oxo-3-deoxygalactonate by *Escherichia coli* K-12. Biochemical and genetical studies. *Arch Microbiol* **1**, 199–206 (1978).

[31] Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**, 946–950 (2009).

[32] Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* **32**, 170–173 (2011).

[33] Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346**, 1533–1536 (2014).

[34] Treves, A. & Panzeri, S. The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Comput* **7**, 399–407 (1995).

[35] Kinney, J. B., Tkačik, G. & Callan, C. G. Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* **104**, 501–506 (2007).

[36] Atwal, G. S. & Kinney, J. B. Learning Quantitative Sequence-Function Relationships from Massively Parallel Experiments. *J Stat Phys* **162**, 1203–1243 (2016).

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

[38] Patil, A., Huard, D. & Fonnesbeck, C. J. PyMC: Bayesian Stochastic Modelling in Python. *J Stat Softw* **35**, 1–811 (2010).

[39] Sivia, D. & Skilling, J. *Data Analysis: A Bayesian Tutorial* ((Oxford, Oxford University Press), 2006).

[40] Kinney, J. B. & Atwal, G. S. Parametric Inference in the Large Data Limit Using Maximally Informative Models. *Neural Comput* **26**, 637–653 (2014).

[41] Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: The mcmc hammer. *Publ Astron Soc Pac* **125**, 306–312 (2013).

[42] Song, S. & Park, C. Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *J Bacteriol* **179**, 7025–7032 (1997).

[43] Ushida, C. & Aiba, H. Helical phase dependent action of CRP: effect of the distance between the CRP site and the -35 region on promoter activity. *Nucleic Acids Res* **18**, 6325–6330 (1990).

[44] Gaston, K., Bell, A., Kolb, A., Buc, H. & Busby, S. Stringent spacing requirements for transcription activation by CRP. *Cell* **62**, 733–743 (1990).

[45] Baba, T. *et al.* Construction of *escherichia coli* k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).

[46] Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* **97**, 6640–6645 (2000).

[47] Razo-Mejia, M. *et al.* Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. *Phys Biol* **11**, 026005 (2014).

[48] Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).

[49] Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

[50] Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376–386 (2002).

[51] Kerner, M. J. *et al.* Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in *Escherichia coli*. *Cell* **122**, 209–220 (2005).

[52] Soufi, B. & Macek, B. Stable Isotope Labeling by Amino Acids Applied to Bacterial Cell Culture. In *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)*, 9–22 (Humana Press, New York, NY, New York, NY, 2014).

[53] Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359–362 (2009).

[54] Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2**, 1896–1906 (2007).

[55] Kalli, A. & Hess, S. Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer. *Proteomics* **12**, 21–31 (2011).

[56] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–1372 (2008).

[57] Cox, J. *et al.* A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **4**, 698–705 (2009).