

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

Nathan M. Belliveau^a, Stephanie L. Barnes^a, William T. Ireland^b, Daniel L. Jones^c, Mike J. Sweredoski^d, Annie Moradian^d, Sonja Hess^{d,e}, Justin B. Kinney^f, and Rob Phillips^{a,b,g,*}

^aDivision of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125; ^bDepartment of Physics, California Institute of Technology, Pasadena, CA, 91125; ^cDepartment of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; ^dProteome Exploration Laboratory (PEL), Beckman Institute, California Institute of Technology, Pasadena, CA, 91125; ^eCurrent address: MedImmune, One Medimmune Way, Gaithersburg, MD, 20878, United States; ^fSimons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724; ^gDepartment of Applied Physics, California Institute of Technology, Pasadena, CA, 91125

*Corresponding author. E-mail: phillips@pboc.caltech.edu

Gene regulation is one of the most ubiquitous processes in biology. But while the catalog of bacterial genomes continues to expand rapidly, we remain ignorant about how almost all of the genes in these genomes are regulated. At present, characterizing the molecular mechanisms by which individual regulatory sequences operate requires focused efforts using low-throughput methods. Here we show how a combination of massively parallel reporter assays, mass spectrometry, and information-theoretic modeling can be used to dissect bacterial promoters in a systematic and scalable way. We demonstrate this method on both well-studied and previously uncharacterized promoters in the enteric bacterium *Escherichia coli*. In all cases we recover nucleotide-resolution models of promoter mechanism. For some promoters, including previously unannotated ones, the approach allowed us to further extract quantitative biophysical models describing input-output relationships. This method opens up the possibility of exhaustively dissecting the mechanisms of promoter function in *E. coli* and a wide range of other bacteria.

1 The sequencing revolution has left in its wake an enormous
2 challenge: the rapidly expanding catalog of sequenced genomes
3 is far outpacing a sequence-level understanding of how the
4 genes in these genomes are regulated. This ignorance extends
5 from viruses to bacteria to archaea to eukaryotes. Even in
6 *E. coli*, the model organism in which transcriptional regula-
7 tion is best understood, we still have no indication if or how
8 more than half of the genes are regulated (Fig. S1; see also
9 RegulonDB (1) or EcoCyc (2)). In other model bacteria such
10 as *Bacillus subtilis*, *Caulobacter crescentus*, *Vibrio harveyii*,
11 or *Pseudomonas aeruginosa*, far fewer genes have established
12 regulatory mechanisms (3–5).

13 New approaches are needed for studying regulatory archi-
14 tecture in these and other bacteria. Although an arsenal of
15 genetic and biochemical methods have been developed for
16 dissecting promoter function at individual bacterial promoters
17 (reviewed in Minchin *et al.* (6)), these methods are not readily
18 parallelized. As a result, they will likely not lead to a com-
19 prehensive understanding of full regulatory genomes anytime
20 soon. RNA sequencing, chromatin immunoprecipitation, and
21 other high-throughput techniques are increasingly being used
22 to study gene regulation in *E. coli* (7–11), but these methods
23 are incapable of revealing either the nucleotide-resolution loca-
24 tion of all functional transcription factor binding sites, or the
25 way in which interactions between DNA-bound transcription
26 factors and RNA polymerase modulate transcription.

27 In recent years a variety of massively parallel reporter
28 assays have been developed for dissecting the functional archi-
29 tecture of transcriptional regulatory sequences in bacteria,
30 yeast, and metazoans. These technologies have been used to
31 infer biophysical models of well-studied loci, to characterize
32 synthetic promoters constructed from known binding sites,
33 and to search for new transcriptional regulatory sequences (12–
34 18). CRISPR assays have also shown promise for identifying

longer range enhancer-promoter interactions in mammalian
cells (19). However, no approach for using massively parallel
reporter technologies to decipher the functional mechanisms of
previously uncharacterized regulatory sequences has yet been
established.

Here we describe a systematic and scalable approach for
dissecting the functional architecture of previously uncharac-
terized bacterial promoters at nucleotide resolution using a
combination of genetic, functional, and biochemical measure-
ments. First, a massively parallel reporter assay (Sort-Seq
(12)) is performed on a promoter in multiple growth condi-
tions in order to identify functional transcription factor binding sites.
DNA affinity chromatography and mass spectrometry (20, 21)
are then used to identify the regulatory proteins that recognize
these sites. In this way one is able to identify both the func-
tional transcription factor binding sites and cognate transcrip-
tion factors in previously unstudied promoters. Subsequent
massively parallel assays are then performed in gene-deletion
strains to provide additional validation of the identified regu-
lators. The reporter data thus generated is also used to infer
sequence-dependent quantitative models of transcriptional regu-
lation. In what follows, we first illustrate the overarching
logic of our approach through application to four previously
annotated promoters: *lacZYA*, *relBE*, *marRAB*, and *yebG*.
We then apply this strategy to the previously uncharacterized
promoters of *purT*, *xylE*, and *dgoRKADT*, demonstrating the
ability to go from complete regulatory ignorance to explicit
quantitative models of a promoter’s input-output behavior.

Results

To dissect how a promoter is regulated, we begin by performing
Sort-Seq (12). As shown in Fig. 1A, Sort-Seq works by first
generating a library of cells, each of which contains a mutated

67 promoter that drives expression of GFP from a low copy
 68 plasmid (5-10 copies per cell (22)) and provides a read-out
 69 of transcriptional state. We use fluorescence-activated cell
 70 sorting (FACS) to sort cells into multiple bins gated by their
 71 fluorescence level and then sequence the mutated plasmids
 72 from each bin. We found it sufficient to sort the libraries
 73 into four bins and generated data sets of about 0.5-2 million
 74 sequences across the sorted bins (Fig. S3A-D). To identify
 75 putative binding sites, we calculate 'expression shift' plots that
 76 show the average change in fluorescence when each position of
 77 the regulatory DNA is mutated (Fig. 1B, top plot). Mutations
 78 to the DNA will in general disrupt binding of transcription
 79 factors (23), so regions with a positive shift are suggestive of
 80 binding by a repressor, while a negative shift suggests binding
 81 by an activator or RNA polymerase (RNAP).

82 The identified binding sites are further interrogated by
 83 performing information-based modeling with the Sort-Seq data.
 84 Here we generate energy matrix models (12, 24) that
 85 describe the sequence-dependent energy of interaction of a
 86 transcription factor at each putative binding site. For each
 87 matrix, we use a convention that the wild-type sequence is
 88 set to have an energy of zero (see example energy matrix in
 89 Fig. 1B). Mutations that enhance binding are identified in blue,
 90 while mutations that weaken binding are identified in red. We
 91 also use these energy matrices to generate sequence logos (25)
 92 which provides a useful visualization of the sequence-specificity
 93 (see above matrix in Fig. 1B).

94 In order to identify the putative transcription factors, we
 95 next perform DNA affinity chromatography experiments using
 96 DNA oligonucleotides containing the binding sites identified
 97 by Sort-Seq. Here we apply a stable isotopic labeling of cell
 98 culture (SILAC (26)) approach, which enables us to perform
 99 a second reference affinity chromatography that is simultane-
 100 ously analyzed by mass spectrometry. We perform chromatog-
 101 raphy using magnetic beads with tethered oligonucleotides
 102 containing the putative binding site (Fig. 1C). Our reference
 103 purification is performed identically, except that the binding
 104 site has been mutated away. The abundance of each protein
 105 is determined by mass spectrometry and used to calculate
 106 protein enrichment ratios, with the target transcription factor
 107 expected to exhibit a ratio greater than one. The reference pu-
 108 rification ensures that non-specifically bound proteins will have
 109 a protein enrichment near one. This mass spectrometry data
 110 and the energy matrix models provide insight into the identity
 111 of each regulatory factor and potential regulatory mechanisms.
 112 In certain instances these insights then allow us to probe the
 113 Sort-Seq data further through additional information-based
 114 modeling using thermodynamic models of gene regulation. As
 115 further validation of binding by an identified regulator, we also
 116 perform Sort-Seq experiments in gene deletion strains, which
 117 should no longer show the associated positive or negative shift
 118 in expression at their binding site.

119 Sort-Seq recovers the regulatory features of well-char- 120 acterized promoters.

121 To first demonstrate Sort-Seq as a tool to discover regulatory
 122 binding sites *de novo* we began by looking at the promoters
 123 of *lacZYA* (*lac*), *relBE* (*rel*), and *marRAB* (*mar*). These pro-
 124 moters have been studied extensively (27-29) and provide a
 125 useful testbed of distinct regulatory motifs. To proceed we con-
 126 structed libraries for each promoter by mutating their known
 127 regulatory binding sites. (See Supplemental Information Sec-

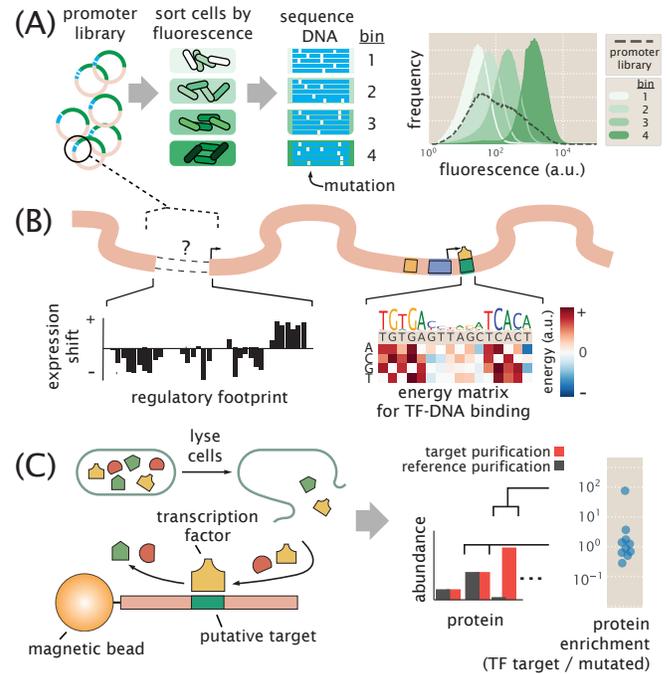


Fig. 1. Overview of approach to characterize transcriptional regulatory DNA, using Sort-Seq and mass spectrometry. (A) Schematic of Sort-Seq. A promoter plasmid library is placed upstream of GFP and is transformed into cells. The cells are sorted into four bins by FACS and after regrowth, plasmids are purified and sequenced. The entire intergenic region associated with a promoter is included on the plasmid and a separate downstream ribosomal binding site sequence is used for translation of the *GFP* gene. The fluorescence histograms show the fluorescence from a library of the *rel* promoter and the resulting sorted bins. (B) Regulatory binding sites are identified by calculating the average expression shift due to mutation at each position. In the schematic, positive expression shifts are suggestive of binding by repressors, while negative shifts would suggest binding by an activator or RNAP. Quantitative models can be inferred to describe the associated DNA-protein interactions. An example energy matrix that describes the binding energy between an as yet unknown transcription factor to the DNA is shown. By convention, the wild-type nucleotides have zero energy, with blue squares identifying mutations that enhance binding (negative energy), and where red squares reduce binding (positive energy). The wild-type sequence is written above the matrix. (C) DNA affinity chromatography and mass spectrometry is used to identify the putative transcription factor (TF) for an identified repressor site. DNA oligonucleotides containing the target binding site are tethered to magnetic beads and used to purify the target transcription factor from cell lysate. Protein abundance is determined by mass spectrometry and a protein enrichment is calculated as the ratio in abundance relative to a second reference experiment where the target sequence is mutated away.

128 tion B and Fig. S3E,F for additional characterization). We
 129 begin by considering the *lac* promoter, which contains three *lac*
 130 repressor (LacI) binding sites, two of which we consider here,
 131 and a cyclic AMP receptor (CRP) binding site. It exhibits the
 132 classic catabolic switch-like behavior that results in diauxie
 133 when *E. coli* is grown in the presence of glucose and lactose
 134 sugars (27). Here we performed Sort-Seq with cells grown in
 135 M9 minimal media with 0.5% glucose. The expression shifts
 136 at each nucleotide position are shown in Fig. 2A, with anno-
 137 tated binding sites noted above the plot. The expression shifts
 138 reflect the expected regulatory role of each binding site, show-
 139 ing positive shifts for LacI and negative shifts for CRP and
 140 RNAP. The difference in magnitude at the two LacI binding

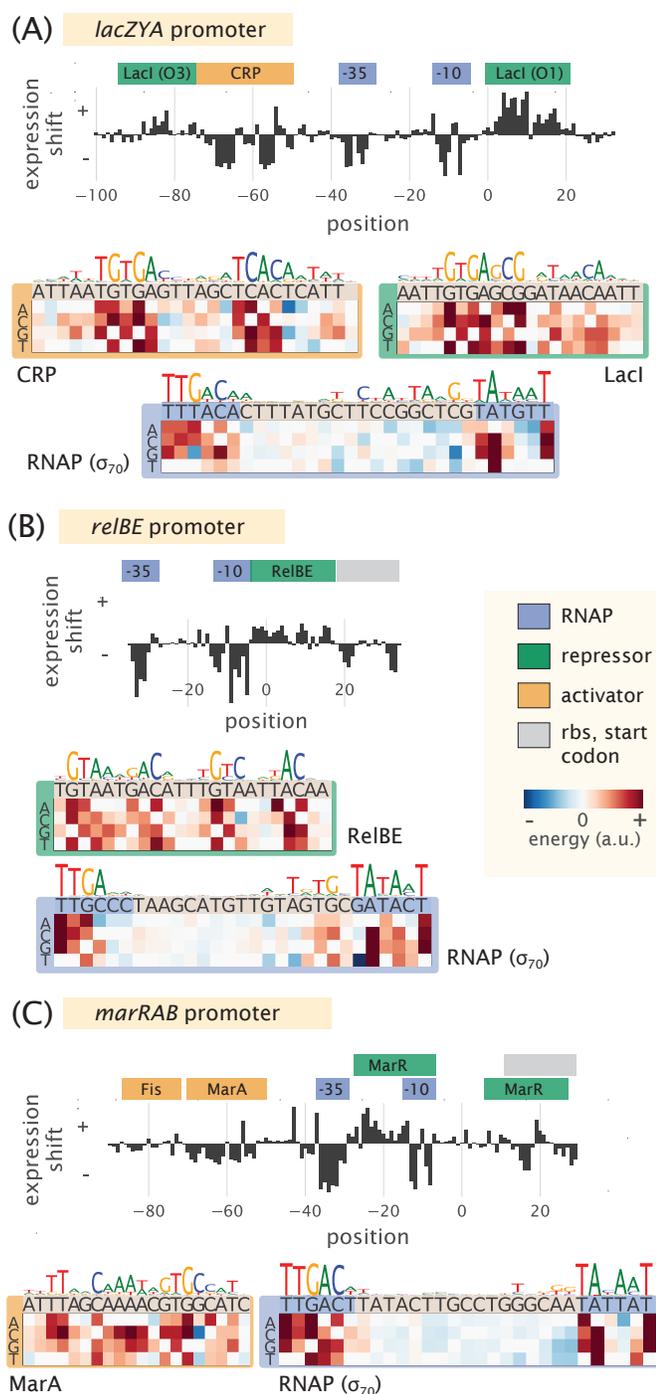


Fig. 2. Characterization of the regulatory landscape of the *lac*, *rel*, and *mar* promoters. (A) Sort-Seq of the *lac* promoter. Cells were grown in M9 minimal media with 0.5% glucose at 37°C. Expression shifts are shown, with annotated binding sites for CRP (activator), RNAP (-10 and -35 subsites), and LacI (repressor) noted. Energy matrices and sequence logos are shown for each binding site. (B) Sort-Seq of the *rel* promoter. Cells were also grown in M9 minimal media with 0.5% glucose at 37°C. The expression shifts identify the binding sites of RNAP and RelBE (repressor), and energy matrices and sequence logos are shown for these. (C) Sort-Seq of the *mar* promoter. Here cells were grown in lysogeny broth (LB) at 30°C. The expression shifts identify the known binding sites of Fis and MarA (activators), RNAP, and MarR (repressor). Energy matrices and sequence logos are shown for MarA and RNAP. Annotated binding sites are based on those in RegulonDB.

sites likely reflect the different binding energies between these two binding site sequences, with LacI O3 having an *in vivo* dissociation constant that is almost three orders of magnitude weaker than the LacI O1 binding site (27, 30).

Next we consider the *rel* promoter that transcribes the toxin-antitoxin pair RelE and RelB. It is one of about 36 toxin-antitoxin systems found on the chromosome, with important roles in cellular physiology including cellular persistence (31). When the toxin, RelE, is in excess of its cognate binding partner, the antitoxin RelB, the toxin causes cellular paralysis through cleavage of mRNA (32). Interestingly, the antitoxin protein also contains a DNA binding domain and is a repressor of its own promoter (33). We similarly performed Sort-Seq, with cells grown in M9 minimal media. The expression shifts are shown in Fig. 2B and were consistent with binding by RNAP and RelBE. In particular, a positive shift was observed at the binding site for RelBE, and the RNAP binding site mainly showed a negative shift in expression.

The third promoter, *mar*, is associated with multiple antibiotic resistance since its operon codes for the transcription factor MarA, which activates a variety of genes including the major multi-drug resistance efflux pump, ArcAB-tolC, and increases antibiotic tolerance (29). The *mar* promoter is itself activated by MarA, SoxS, and Rob (via the so-called marbox binding site), and further enhanced by Fis, which binds upstream of this marbox (34). Under standard laboratory growth it is under repression by MarR (29). We found that the promoter's fluorescence was quite dim in M9 minimal media and instead grew libraries in lysogeny broth (LB) at 30°C (35). Again, the different features in the expression shift plot (Fig. 2C) appeared to be consistent with the noted binding sites. One exception was that the downstream MarR binding site was not especially apparent. Both positive and negative expression shifts were observed along its binding site, which may be due to overlap with other features present including the native ribosomal binding site. There have also been reported binding sites for CRP, Cra, CpxR/CpxA, and AcrR (1). However the studies associated with these annotations either required overexpression of the associated transcription factor, were computationally predicted, or demonstrated through *in vitro* assays and not necessarily expected under the growth condition considered here.

While each promoter qualitatively showed the expected regulatory behavior in each expression shift plot, it was important to show that we could also recover the quantitative features of binding by each transcription factor. Here we inferred energy matrices and associated sequence logos for the binding sites of RNAP, LacI, CRP, RelBE, MarA, and Fis. These are shown in Fig. 2A-C and Fig. S4, and indeed, agreed well with sequence logos generated from known genomic binding sites for these transcription factors (Pearson correlation coefficient $r=0.5-0.9$; see Supplemental Information Section C). For the repressors RelBE and MarR, there was no data available that characterized their sequence specificity with which to compare against. Here, instead, we validated our data by performing Sort-Seq in strains where the *relBE* or *marR* genes were deleted. In each case this resulted in a loss of the expression shift associated with binding by these repressors (Fig. 3), suggesting that the observed features are due to binding by these transcription factors.

141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200

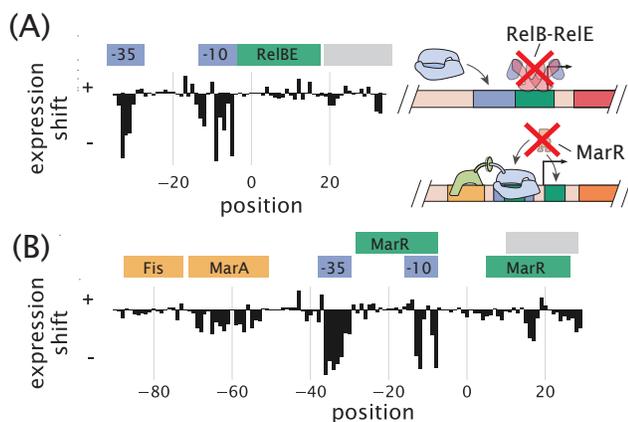


Fig. 3. Expression shifts reflect binding by regulatory proteins. (A) Expression shifts for the *rel* promoter, but in a Δrel genetic background. Cells were grown in conditions identical to Fig. 2B but do not show a positive expression shift across the entire RelBE binding site. (B) Expression shifts for the *mar* promoter, but in a $\Delta marR$ genetic background. The positive expression shift observed where MarR is expected to bind is no longer observed. Binding site annotations are identified in blue for RNAP sites, green for repressor sites, yellow for activator sites, and gray for ribosomal binding site and start codons. These annotations refer to the binding sites noted on RegulonDB that were observed in the Sort-Seq data.

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

202 It was next important to show that DNA affinity chromatography could be used to identify transcription factors in *E. coli*.
 203 In particular, a challenge arises in identifying transcription factors in most organisms due to their very low abundance.
 204 In *E. coli* the cumulative distribution in protein copy number shows that more than half have a copy number less than 100
 205 per cell, with 90% having copy number less than 1,000 per cell. This is several orders of magnitude below that of many
 206 other cellular proteins (36).

207 We began by applying the approach to known binding sites for LacI and RelBE. For LacI, which is present in *E. coli*
 208 in about 10 copies per cell, we used the strongest binding site sequence, Oid (*in vivo* $K_d \approx 0.05$ nM), and the weakest
 209 natural operator sequence, O3 (*in vivo* $K_d \approx 110$ nM) (27, 30, 37). In Fig. 4A we plot the protein enrichments from each
 210 transcription factor identified by mass spectrometry. LacI was found with both DNA targets, with fold enrichment greater
 211 than 10 in each case, and significantly higher than most of the proteins detected (indicated by the shaded region, which
 212 represents the 95% probability density region of all proteins detected, including non-DNA binding proteins). Purification
 213 of LacI with about 10 copies per cell using the weak O3 binding site sequence are near the limit of what would be necessary
 214 for most *E. coli* promoters.

215 To ensure this success was not specific to LacI, we also applied chromatography to the RelBE binding site. RelBE
 216 provides an interesting case since the strength of binding by RelB to DNA is dependent on whether RelE is bound in complex
 217 to RelB (with at least a 100 fold weaker dissociation constant reported in the absence of RelE (38, 39)). As shown
 218 in Fig. 4B, we found over 100 fold enrichment of both proteins by mass spectrometry. To provide some additional intuition
 219 into these results we also considered the predictions from a

237 statistical mechanical model of DNA binding affinity (See Supplemental Information Section D). As a consequence of
 238 performing a second reference purification, we find that fold enrichment should mostly reflect the difference in binding energy
 239 between the DNA sequences used in the two purifications, and be much less dependent on whether the protein was in low or
 240 high abundance within the cell. This appeared to be the case when considering other *E. coli* strains with LacI copy numbers
 241 between about 10 and 1,000 copies per cell (Fig. S5C). Further characterization of the measurement sensitivity and dynamic
 242 range of this approach is noted in Supplemental Information Section E.

243 Sort-Seq discovers regulatory architectures in unannotated regulatory regions.

244 Given that more than half of the promoters in *E. coli* have no annotated transcription factor binding sites in RegulonDB, we
 245 narrowed our focus by using several high-throughput studies to identify candidate genes to apply our approach (40, 41).
 246 The work by Schmidt *et al.* (41) in particular measured the protein copy number of about half the *E. coli* genes across
 247 22 distinct growth conditions. Using this data, we identified genes that had substantial differential gene expression patterns
 248 across growth conditions, thus hinting at the presence of regulation and even how that regulation is elicited by environmental
 249 conditions (see further details in Supplemental Information Section A and Fig. S2A-C). On the basis of this survey, we chose
 250 to investigate the promoters of *purT*, *xylE*, and *dgoRKADT*. To apply Sort-Seq in a more exploratory manner, we considered
 251 three 60 bp mutagenized windows spanning the intergenic region of each gene. While it is certainly possible
 252 to investigate other genes, we chose to focus on these three genes for the purposes of this study.

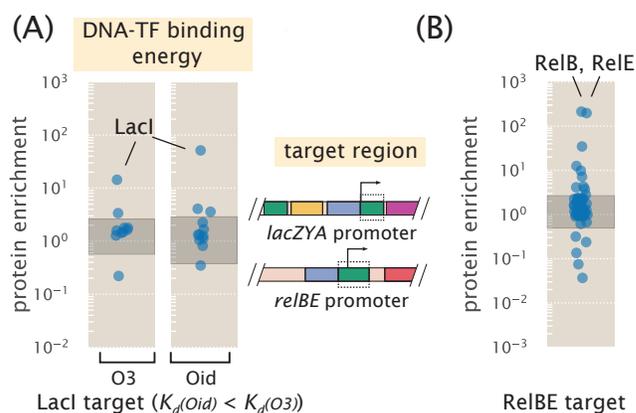


Fig. 4. DNA affinity purification and identification of LacI and RelBE by mass spectrometry using known target binding sites.

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

267 sible that regulatory features will lie outside of this window,
 268 a search of known regulatory binding sites suggest that this
 269 should be sufficient to capture just over 70% of regulatory
 270 features in *E. coli* and provide a useful starting point (Fig. S6).

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

273 The first of our candidate promoters is associated with expression
 274 of *purT*, one of two genes found in *E. coli* that catalyze
 275 the third step in *de novo* purine biosynthesis (42, 43). Due to a
 276 relatively short intergenic region, about 120 bp in length that
 277 is shared with a neighboring gene *yebG*, we also performed
 278 Sort-Seq on the *yebG* promoter (oriented in the opposite direction
 279 (44); see schematic in Fig. 5A). To begin our exploration
 280 of the *purT* and *yebG* promoters, we performed Sort-Seq with
 281 cells grown in M9 minimal media with 0.5% glucose. The
 282 associated expression shift plots are shown in Fig. 5A. While
 283 we performed Sort-Seq on a larger region than shown for
 284 each promoter, we only plot the regions where regulation was
 285 apparent.

286 For the *yebG* promoter, the features were largely consistent
 287 with prior work, containing a binding sites for LexA and RNAP.
 288 However, we found that the RNAP binding site is shifted 9
 289 bp downstream from what was identified previously through a
 290 computational search (44), demonstrating the ability of our
 291 approach to identify and correct errors in the published record.
 292 We were also able to confirm that the *yebG* promoter was
 293 induced in response to DNA damage by repeating Sort-Seq
 294 in the presence of mitomycin C (a potent DNA cross-linker
 295 known to elicit the SOS response and proteolysis of LexA (45);
 296 see Fig. S7A, B, and D).

297 Given the role of *purT* in the synthesis of purines, and the
 298 tight control over purine concentrations within the cell (42),
 299 we performed Sort-Seq of the *purT* promoter in the presence
 300 or absence of the purine, adenine, in the growth media. In
 301 growth without adenine (Fig. 5A, right plot), we observed two
 302 negative regions in the expression shift plot. Through inference
 303 of an energy matrix, these two features were identified as the
 304 -10 and -35 regions of an RNAP binding site. While these two
 305 features were still present upon addition of adenine, as shown
 306 in Fig. 5B, this growth condition also revealed a putative
 307 repressor site between the -35 and -10 RNAP binding sites,
 308 indicated by a positive shift in expression (green annotation).

309 Following our strategy to find not only the regulatory sequences,
 310 but also their associated transcription factors, we next applied DNA
 311 affinity chromatography using this putative binding site sequence.
 312 In our initial attempt however, we were unable to identify any
 313 substantially enriched transcription factor (Fig. S7C). With
 314 repression observed only when cells were grown in the presence
 315 of adenine, we reasoned that the transcription factor may require
 316 a related ligand in order to bind the DNA, possibly through an
 317 allosteric mechanism. Importantly, we were able to infer an
 318 energy matrix to the putative repressor site whose sequence-specificity
 319 matched that of the well-characterized repressor, PurR ($r=0.82$; see
 320 Fig. S4). We also noted ChIP-chip data of PurR that suggests it
 321 might bind within this intergenic region (43). We therefore
 322 repeated the purification in the presence of hypoxanthine, which
 323 is a purine derivative that also binds PurR (46). As shown in
 324 Fig. 5C, we now observed a substantial enrichment of PurR
 325 with this putative binding site sequence. As further validation,
 326 we performed Sort-Seq once more in the adenine-rich growth
 327

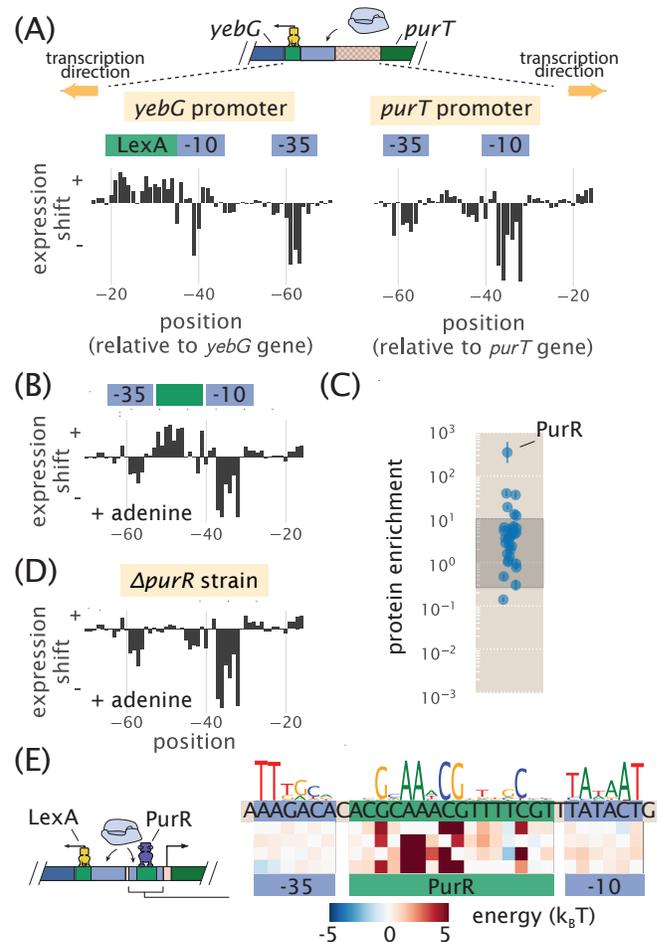


Fig. 5. Sort-Seq distinguishes directional regulatory features and uncovers the regulatory architecture of the *purT* promoter. (A) A schematic is shown for the approximately 120 bp region between the *yebG* and *purT* genes, which code in opposite directions. Expression shifts are shown for 60 bp regions where regulation was observed for each promoter, with positions noted relative to the start codon of each native coding gene. Cells were grown in M9 minimal media with 0.5% glucose. The -10 and -35 RNAP binding sites of the *purT* promoter were determined through inference of an energy matrix and are identified in blue. (B) Expression shifts for the *purT* promoter, but in M9 minimal media with 0.5% glucose supplemented with adenine (100 μg/ml). A putative repressor site is annotated in green. (C) DNA affinity chromatography was performed using the identified repressor site and protein enrichment values for transcription factors are plotted. Cell lysate was produced from cells grown in M9 minimal media with 0.5 % glucose. Binding was performed in the presence of hypoxanthine (10 μg/ml). Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all protein detected. (D) Identical to (B) but performed with cells containing a $\Delta purR$ genetic background. (E) Summary of regulatory binding sites and transcription factors that bind within the intergenic region between the genes of *yebG* and *purT*. Energy weight matrices and sequence logos are shown for the PurR repressor and RNAP binding sites. Data was fit to a thermodynamic of simple repression, yielding energies in units of $k_B T$.

condition, but in a $\Delta purR$ strain. In the absence of PurR, the putative repressor binding site disappeared (Fig. 5D), which is consistent with PurR binding at this location.

In Fig. 5E we summarize the regulatory features between

328
 329
 330
 331

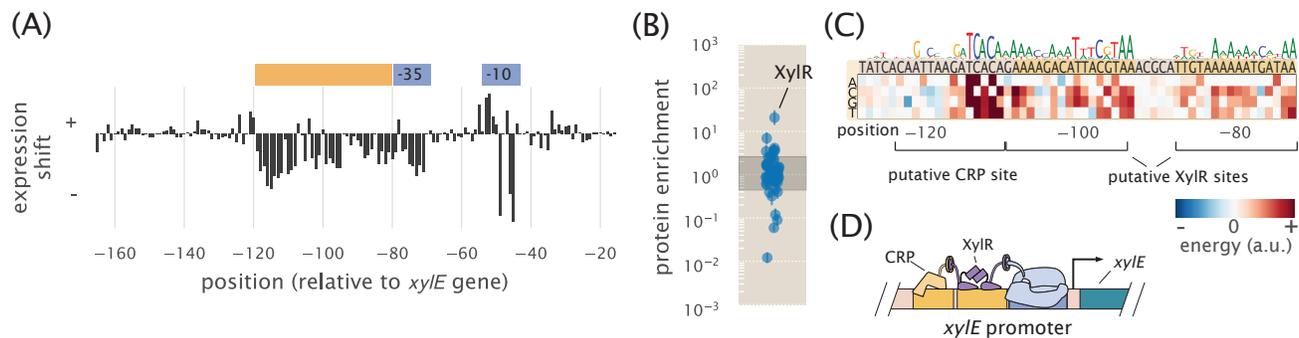


Fig. 6. Sort-Seq identifies a set of activator binding sites that drive expression of RNAP at the *xylE* promoter. (A) Expression shifts are shown for the *xylE* promoter, with Sort-Seq performed on cells grown in M9 minimal media with 0.5% xylose. The -10 and -35 regions of an RNAP binding site (blue) and a putative activator region (orange) are annotated. (B) DNA affinity chromatography was performed using the putative activator region and protein enrichment values for transcription factors are plotted. Cell lysate was generated from cells grown in M9 minimal media with 0.5% xylose and binding was performed in the presence of xylose supplemented at the same concentration as during growth. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates. The gray shaded region represents 95% probability density region of all proteins detected. (C) An energy matrix was inferred for the region upstream of the RNAP binding site. The associated sequence logo is shown above the matrix. Two binding sites for XylR were identified (see also Fig. S4 and Fig. S7F) along with a CRP binding site. (D) Summary of regulatory features identified at *xylE* promoter, with the identification of an RNAP binding site and tandem binding sites for XylR and CRP.

332 the coding genes of *purT* and *yebG*, including the new features
 333 identified by Sort-Seq. With the appearance of a simple repres-
 334 sion architecture (47) for the *purT* promoter, we extended our
 335 analysis by developing a thermodynamic model to describe
 336 repression by PurR. This enabled us to infer the binding en-
 337 ergies of RNAP and PurR in absolute $k_B T$ energies (48), and
 338 we show the resulting model in Fig. 5E (see additional details
 339 in Supplemental Information Section Information H.3.4).

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

342 The next unannotated promoter we considered was associated
 343 with expression of *xylE*, a xylose/proton symporter involved in
 344 uptake of xylose. From our analysis of the Schmidt *et al.* (41)
 345 data, we found that *xylE* was sensitive to xylose and proceeded
 346 by performing Sort-Seq in cells grown in this carbon source.
 347 Interestingly, the promoter exhibited essentially no expression
 348 in other media (Fig. S7E). We were able to locate the RNAP
 349 binding site between -80 bp and -40 bp relative to the *xylE* gene
 350 (Fig. 6A, annotated in blue). In addition, the entire region
 351 upstream of the RNAP appeared to be involved in activating
 352 gene expression (annotated in orange in Fig. 6A), suggesting
 353 the possibility of multiple transcription factor binding sites.

354 We applied DNA affinity chromatography using a DNA
 355 target containing this entire upstream region. Due to the
 356 stringent requirement for xylose to be present for any mea-
 357 surable expression, xylose was supplemented in the lysate
 358 during binding with the target DNA. In Fig. 6B we plot the
 359 enrichment ratios from this purification and find XylR to be
 360 most significantly enriched. From an energy matrix inferred
 361 for the entire region upstream of the RNAP site, we were able
 362 to identify two correlated 15 bp regions (dark yellow shaded
 363 regions in Fig. 6C). Mutations of the XylR protein have been
 364 found to diminish transport of xylose (49), which in light of
 365 our result, may be due in part to a loss of activation and ex-
 366 pression of this xylose/proton symporter. These binding sites
 367 were also similar to those found on two other promoters known
 368 to be regulated by XylR (*xylA* and *xylF* promoters), whose
 369 promoters also exhibit tandem XylR binding sites and strong

binding energy predictions with our energy matrix (Fig. S7F).

370
 371 Within the upstream activator region in Fig. 6A there still
 372 appeared to be a binding site unaccounted for with these tan-
 373 dem XylR binding sites. From the energy matrix, we were
 374 further able to identify a binding site for CRP, which is noted
 375 upstream of the XylR binding sites in Fig. 6C. While we did
 376 not observe a significant enrichment of CRP in our protein pu-
 377 rification, the most energetically favorable sequence predicted
 378 by our model, TGC GACC NAGATCACA, closely matches the
 379 CRP consensus sequence of TGTGANNNNNTCACA. In
 380 contrast to the *lac* promoter, binding by CRP here appears
 381 to depend more on the right half of the binding site sequence.
 382 CRP is known to activate promoters by multiple mechanisms
 383 (50), and CRP binding sites have been found adjacent to the
 384 activators XylR and AraC (49, 51), in line with our result.
 385 While further work will be needed to characterize the spe-
 386 cific regulatory mechanism here, it appears that activation of
 387 RNAP is mediated by both CRP and XylR and we summarize
 388 this result in Fig. 6D (and considered further in Supplemental
 389 Information Section H.3.4).

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

392 As a final illustration of the approach developed here, we con-
 393 sidered the unannotated promoter of *dgoRKADT*. The operon
 394 codes for D-galactonate-catabolizing enzymes; D-galactonate
 395 is a sugar acid that has been found as a product of galac-
 396 tose metabolism (52). We began by measuring expression
 397 from a non-mutagenized *dgoRKADT* promoter reporter to
 398 glucose, galactose, and D-galactonate. Cells grown in galac-
 399 tose exhibited higher expression than in glucose, as found by
 400 Schmidt *et al.* (41), and even higher expression when cells
 401 were grown in D-galactonate (Fig. S8A). This likely reflects
 402 the physiological role provided by the genes of this promoter,
 403 which appear necessary for metabolism of D-galactonate. We
 404 therefore proceeded by performing Sort-Seq with cells grown
 405 in either glucose or D-galactonate, since these appeared to
 406 represent distinct regulatory states, with expression low in
 407 glucose and high in D-galactonate. Expression shift plots from

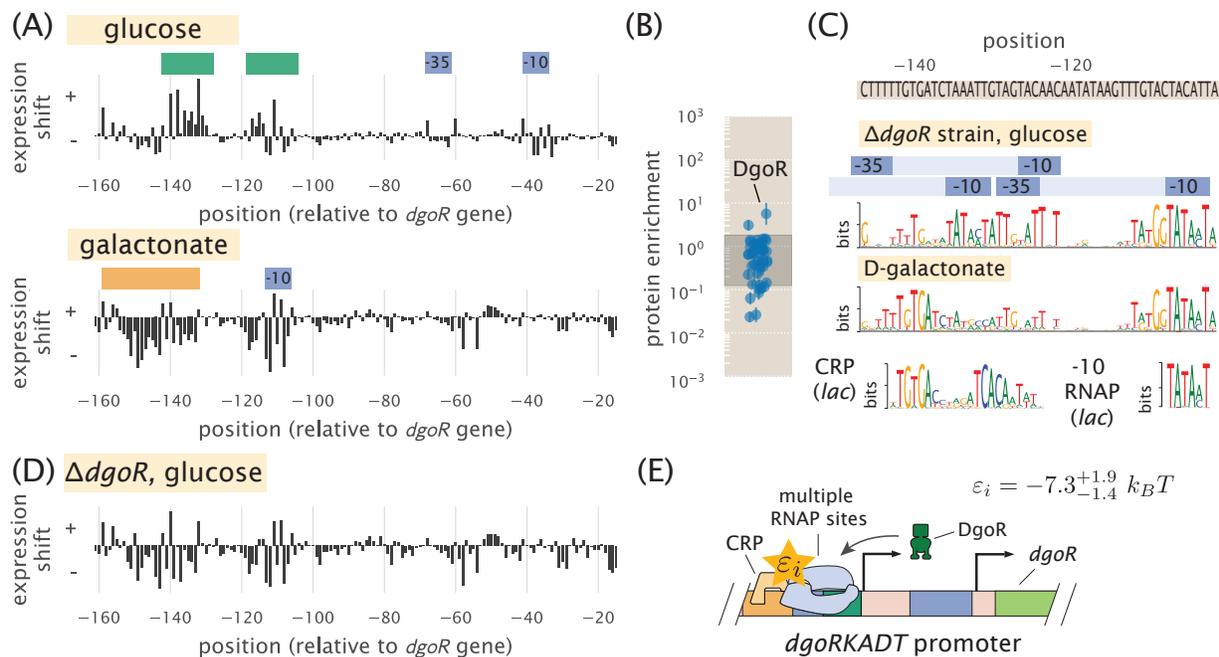


Fig. 7. The *dgoRKADT* promoter is induced in the presence of D-galactonate due to loss of repression by DgoR and activation by CRP. (A) Expression shifts due to mutating the *dgoRKADT* promoter are shown for cells grown in M9 minimal media with either 0.5% glucose (top) or 0.23% D-galactonate (bottom). Regions identified as RNAP binding sites (-10 and -35) are shown in blue and putative activator and repressor binding sites are shown in orange and green, respectively. (B) DNA affinity purification was performed targeting the region between -145 to -110 of the *dgoRKADT* promoter. The transcription factor DgoR was found most enriched among the transcription factors plotted. Error bars represent the standard error of the mean, calculated using log protein enrichment values from three replicates, and the gray shaded region represents 95% probability density region of all proteins detected. (C) Sequence logos were inferred for the most upstream 60 bp region associated with the upstream RNAP binding site annotated in (A). Multiple RNAP binding sites were identified using Sort-Seq data performed in a $\Delta dgoR$ strain, grown in M9 minimal media with 0.5% glucose. (further detailed in Fig. S8). Below this, a sequence logo was also inferred using data from Sort-Seq performed on wild-type cells, grown in D-galactonate, identifying a CRP binding site (class II activation (50)). (D) Expression shifts are shown for the *dgoRKADT* promoter when performed in a $\Delta dgoR$ genetic background, grown in 0.5% glucose. This resembles growth in D-galactonate, suggesting D-galactonate may act as an inducer for DgoR. (E) Summary of regulatory features identified at *dgoRKADT* promoter, with the identification of multiple RNAP binding sites, and binding sites for DgoR and CRP. The interaction energy between CRP and RNAP, ϵ_i , was inferred to be $-7.3^{+1.9}_{-1.4} k_B T$, where the superscripts and subscripts represent the upper and lower bounds of the 95th percentile of the parameter value distribution.

each growth conditions are shown in Fig. 7A.

We begin by considering the results from growth in glucose (Fig. 7A, top plot). Here we identified an RNAP binding site between -30 bp and -70 bp, relative to the native start codon for *dgoR* (Fig. 7B). Another distinct feature was a positive expression shift in the region between -140 bp and -110 bp, suggesting the presence of a repressor binding site. Applying DNA affinity chromatography using this target region we observed an enrichment of DgoR (Fig. 7B), suggesting that the promoter is indeed under repression, and regulated by the first coding gene of its transcript. As further validation of binding by DgoR, the positive shift in expression was no longer observed when Sort-Seq was repeated in a $\Delta dgoR$ strain (Fig. 7D and Fig. S8C). We also were able to identify additional RNAP binding sites that were not apparent due to binding by DgoR. While only one RNAP -10 motif is clearly visible in the sequence logo shown Fig. 7C (top sequence logo; TATAAT consensus sequence), we used simulations to demonstrate that the entire sequence logo shown can be explained by the convolution of three overlapping RNAP binding sites (See Supplemental Information Section D and Fig. S8F).

Next we consider the D-galactonate growth condition (Fig. 7A, bottom plot). Like in the expression shift plot for

the $\Delta dgoR$ strain grown in glucose, we no longer observe the positive expression shift between -140 bp and -110 bp. This suggests that DgoR may be induced by D-galactonate or a related metabolite. However, in comparison with the expression shifts in the $\Delta dgoR$ strain grown in glucose, there were some notable differences in the region between -160 bp and -140 bp. Here we find evidence for another CRP binding site. The sequence logo identifies the sequence TGTGA (Fig. 7C, bottom logo), which matches the left side of the CRP consensus sequence. In contrast to the *lac* and *xylE* promoters however, the right half of the binding site directly overlaps with where we would expect to find a -35 RNAP binding site. This type of interaction by CRP has been previously observed and is defined as class II CRP dependent activation (50), though this sequence-specificity has not been previously described.

In order to isolate and better identify this putative CRP binding site we repeated Sort-Seq in *E. coli* strain JK10, grown in 500 μ M cAMP. Strain JK10 lacks adenylate cyclase (*cyaA*) and phosphodiesterase (*cpdA*), which are needed for cAMP synthesis and degradation, respectively, and is thus unable to control intracellular cAMP levels necessary for activation by CRP (derivative of TK310 (37)). Growth in the presence of 500 μ M cAMP provided strong induction from the *dgoRKADT*

431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453

454 promoter and resulted in a sequence logo at the putative CRP
455 binding site that even more clearly resembled binding by CRP
456 (Fig. S8E). This is likely because expression is now dominated
457 by the CRP activated RNAP binding site. Importantly, this
458 data allowed us to further infer the interaction energy between
459 CRP and RNAP, which we estimate to be $-7.3 k_B T$ (further
460 detailed in Supplemental Information Section H.3.4). We
461 summarize the identified regulatory features in Fig. 7E.

462 Discussion

463 We have established a systematic procedure for dissecting the
464 functional mechanisms of previously uncharacterized regula-
465 tory sequences in bacteria. A massively parallel reporter assay,
466 Sort-Seq (12), is used to first elucidate the locations of func-
467 tional transcription factor binding sites. DNA oligonucleotides
468 containing these binding sites are then used to enrich the
469 cognate transcription factors and identify them by mass spec-
470 trometry analysis. Information-based modeling and inference
471 of energy matrices that describe the DNA binding specificity
472 of regulatory factors provide further quantitative insight into
473 transcription factor identity and the growth condition depen-
474 dent regulatory architectures.

475 To validate this approach we examined four previously
476 annotated promoters of *lac*, *rel*, *mar*, and *yebG*, with our results
477 consistent with established knowledge (12, 27, 29, 30, 35, 39).
478 For the *yebG* promoter, however, our approach corrected an
479 error in a previous annotation. Importantly, we find that
480 DNA affinity chromatography experiments on these promoters
481 were highly sensitive. In particular, LacI was unambiguously
482 identified with the weak O3 binding site, even though LacI is
483 present in only about 10 copies per cell (30). Emboldened by
484 this success, we then studied promoters having little or no prior
485 regulatory annotation: *purT*, *xylE*, and *dgoR*. Here our analysis
486 led to a collection of new regulatory hypotheses. For the *purT*
487 promoter, we identified a simple repression architecture (47),
488 with repression by PurR. The *xylE* promoter was found to
489 undergo activation only when cells are grown in xylose, likely
490 due to allosteric interaction between the activator XylR and
491 xylose, and activation by CRP (49, 51). Finally, in the case
492 of *dgoR*, the base-pair resolution allowed us to tease apart
493 overlapping regulatory binding sites, identify multiple RNAP
494 binding sites along the length of the promoter, and infer further
495 quantitative detail about the interaction between the newly
496 identified binding sites for CRP and RNAP. We view these
497 results as a critical first step in the quantitative dissection of
498 transcriptional regulation, which will ultimately be needed for
499 a predictive understanding of how such regulation works.

500 An important aspect of the presented approach is that it
501 is readily parallelized and scalable. There are a number of
502 ways to increase the resolution and throughput. Microarray-
503 synthesized promoter libraries should allow multiple loci to
504 be studied simultaneously. Landing pad technologies for chro-
505 mosomal integration (53) should enable massively parallel
506 reporter assays to be performed in chromosomes instead of on
507 plasmids. Techniques that combine these assays with transcrip-
508 tion start site readout (54) may further allow the molecular
509 regulators of overlapping RNAP binding sites to be decon-
510 volved, or the contributions from separate RNAP binding
511 sites, like those observed on the *dgoR* promoter, to be better
512 distinguished. Although our work was directed toward reg-
513 ulatory regions of *E. coli*, there are no intrinsic limitations

514 that restrict the analysis to this organism. Rather, it should
515 be applicable to any bacterium that supports efficient trans-
516 formation by plasmids. And although we have focused on
517 bacteria, our general strategy should be feasible in a number
518 of eukaryotic systems – including human cell culture – using
519 massively parallel reporter assays (13–15) and DNA-mediated
520 protein pull-down methods (20, 21) that have already been
521 established.

522 Materials and Methods

523 See Supplemental Information Section I for extended experi-
524 mental details.

525 Bacterial strains.

526 All *E. coli* strains used in this work were derived from K-12
527 MG1655, with deletion strains generated by the lambda red
528 recombinase method (55). In the case of deletions for *lysA*
529 ($\Delta lysA::kan$), *purR* ($\Delta purR::kan$), and *xylE* ($\Delta xylE::kan$),
530 strains were obtained from the Coli Genetic Stock Center
531 (CGSC, Yale University, CT, USA) and transferred into a
532 fresh MG1655 strain using P1 transduction. The others were
533 generated in house and include the following deletion strains:
534 $\Delta lacIZYA$, $\Delta relBE::kan$, $\Delta marR::kan$, $\Delta dgoR::kan$ (see Sup-
535 plemental Information Section I.1 for details on strain con-
536 struction).

537 Sort-Seq.

538 Mutagenized single-stranded oligonucleotide pools were pur-
539 chased from Integrated DNA Technologies (Coralville, IA),
540 with a target mutation rate of 9%. Note that in the case of
541 the *lacZ* promoter, the library is identical to that used in the
542 experiments of Razo-Mejia *et al.* (56), and had a mutation
543 rate of approximately 3%. Library oligonucleotides were PCR
544 amplified and inserted into the PCR amplified plasmid back-
545 bone (i.e. vector) of pJK14 (SC101 origin) (12) by Gibson
546 assembly and electroporated into cells following drop dialysis
547 in water.

548 Cells were grown to saturation in LB and then diluted
549 1:10,000 into the appropriate growth media for the promoter
550 under consideration. Upon reaching an OD600 of about 0.3,
551 the cells were washed two times with chilled PBS by spinning
552 down the cells at 4000 rpm for 10 minutes at 4°C and diluted
553 to an OD of 0.1–0.15. A Beckman Coulter MoFlo XDP cell
554 sorter was used to sort cells by fluorescence, with 500,000 cells
555 collected into each of the four bins. Sorted cells were then
556 re-grown overnight in 10 ml of LB media, under kanamycin
557 selection. The plasmid in each bin were miniprepmed following
558 overnight growth (Qiagen, Germany) and PCR was used to
559 amplify the mutated region from each plasmid for Illumina
560 sequencing (see Supplemental Information Section I.3 and I.4
561 for additional Sort-Seq and sequencing details, respectively).
562 Details on constructing expression shift plots and the model
563 inference that was performed are provided in Supplemental
564 Information Section H.

565 DNA affinity chromatography.

566 SILAC labeling (26) was implemented by growing cells in
567 either the stable isotopic form of lysine ($^{13}C_6H_{14}^{15}N_2O_2$),
568 referred to as the heavy label, or natural lysine, referred to as
569 the light label. Cell lysates were prepared using $\Delta lysA$ cells.
570 For each heavy and light labelled cells, 500 ml M9 minimal

571 media was inoculated 1:5,000 with an overnight LB culture of
572 $\Delta lysA$ cells, and grown to an OD600 of ≈ 0.6 (supplemented
573 with the appropriate lysine; 40 $\mu\text{g}/\text{ml}$). Cultures were pelleted,
574 lyse using a Cell Disruptor (CF Range, Constant Systems Ltd.,
575 UK) and concentrated to ~ 150 mg/ml using Amicon Ultra-15
576 centrifugation units (3kDa MWCO, Millipore).

577 DNA affinity chromatography was performed by incubat-
578 ing cell lysate with magnetic beads (Dynabeads MyOne T1,
579 ThermoFisher, Waltham, MA) containing tethered DNA. The
580 DNA was tethered through a linkage between streptavidin on
581 the beads and biotin on the DNA. Single-stranded DNA was
582 purchased from Integrated DNA Technologies with the biotin
583 modification on the 5' end of the oligonucleotide sense strand.
584 Cell lysates were incubated on a rotating wheel with the DNA
585 tethered beads overnight at 4°C. Beads were washed three
586 times using lysis buffer and once more with NEB Buffer 3.1
587 (New England Biolabs, MA, USA). Both purifications (with
588 the target DNA and reference control) were combined by resus-
589 pending in 50 μL NEB Buffer 3.1, and the DNA was cleaved by
590 adding 10 μl of the restriction enzyme PstI (100,000 units/ml,
591 New England Biolabs targeting a CTGCAG sequence on the
592 DNA) and incubating for 1.5 hours at 25°C. The beads were
593 then removed and the samples prepared for mass spectrometry
594 by in-gel digestion with endoproteinase Lys-C.

595 LC-MS/MS analysis and protein quantitation.

596 Liquid chromatography tandem-mass spectrometry (LC-
597 MS/MS) experiments were carried out as previously described
598 (57) and further detailed in supplemental experimental de-
599 tails. Thermo RAW files were processed using MaxQuant (v.
600 1.5.3.30) (58). Spectra were searched against the UniProt *E.*
601 *coli* K-12 database (4318 sequences) as well as a contaminant
602 database (256 sequences). Additional details are provided in
603 Supplemental Information Section I.5. To calculate the overall
604 protein ratio, the non-normalized protein replicate ratios were
605 log transformed and then shifted so that the median protein
606 log ratio within each replicate was zero (i.e., the median pro-
607 tein ratio was 1:1). The overall experimental log ratio was
608 then calculated from the average of the replicate ratios.

609 Code and data availability.

610 All code used for processing data and plotting, as well as the
611 final processed data are available upon request. Thermo RAW
612 files for mass spectrometry are available on the jPOSTrepo
613 repository (59) under accession code PXD007892. Sort-Seq
614 sequencing files are available on the Sequence Read Archive
615 under accession code SRP121362.

616 Acknowledgements.

617 We thank David Tirrell, Bradley Silverman, and Seth Lieblich
618 for access and training for use of their Beckman Coulter MoFlo
619 XDP cell sorter. We thank Jost Vielmetter and Nina Budaeva
620 for access and training for use on their Cell Disruptor. We
621 also thank Hernan Garcia, Manuel Razo-Mejia, Griffin Chure,
622 Suzannah Beeler, Heun Jin Lee, Justin Bois, and Soichi Hi-
623 rokawa for useful advice and discussion. This work was sup-
624 ported by La Fondation Pierre-Gilles de Gennes, the Rosen
625 Center at Caltech, and the National Institutes of Health DP1
626 OD000217 (Director's Pioneer Award), R01 GM085286, and
627 1R35 GM118043-01 (MIRA), the Gordon and Betty Moore
628 Foundation through GBMF227, the National Institutes of

631 References

- 632 1. Gama-Castro S, et al. (2016) RegulonDB version 9.0: high-level
633 integration of gene regulation, coexpression, motif clustering
634 and beyond. *Nucleic Acids Research* 44(D1):D133–D143.
- 635 2. Keseler IM, et al. (2013) EcoCyc: fusing model organism
636 databases with systems biology. *Nucleic Acids Research*
637 41(D1):D605–D612.
- 638 3. Münch R, et al. (2003) PRODORIC: prokaryotic database of
639 gene regulation. *Nucleic Acids Research* 31(1):266–269.
- 640 4. Cipriano MJ, et al. (2013) RegTransBase – a database of
641 regulatory sequences and interactions based on literature: a
642 resource for investigating transcriptional regulation in prokary-
643 otes. *BMC Genomics* 14(1):213–221.
- 644 5. Kılıç S, White ER, Sagitova DM, Cornish JP, Erill I (2013)
645 CollecTF: a database of experimentally validated transcrip-
646 tion factor-binding sites in Bacteria. *Nucleic Acids Research*
647 42(D1):D156–D160.
- 648 6. Minchin SD, Busby SJW (2009) Analysis of mechanisms of
649 activation and repression at bacterial promoters. *Methods*
650 47(1):6–12.
- 651 7. Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJW
652 (2005) Studies of the distribution of *Escherichia coli* cAMP-
653 receptor protein and RNA polymerase along the *E. coli* chro-
654 mosome. *PNAS* 102(49):17693–17698.
- 655 8. Bonocora RP, Wade JT (2015) *ChIP-Seq for genome-scale*
656 *analysis of bacterial DNA-binding proteins*. ((New York, Hu-
657 mana Press)), pp. 327–340.
- 658 9. Zheng D, Constantinidou C, Hobman JL, Minchin SD (2004)
659 Identification of the CRP regulon using in vitro and in vivo
660 transcriptional profiling. *Nucleic Acids Research* 32(19):5874–
661 5893.
- 662 10. Singh SS, et al. (2014) Widespread suppression of intragenic
663 transcription initiation by H-NS. *Genes & Development*
664 28(3):214–219.
- 665 11. Wade JT (2015) ChIP-Seq for Genomic-Scale Analysis of Bacter-
666 ial DNA-Binding Proteins. *Prokaryotic Systems Biology*
667 883(Chapter 7):119–134.
- 668 12. Kinney JB, Murugan A, Callan CG, Cox EC (2010) Using
669 deep sequencing to characterize the biophysical mechanism of a
670 transcriptional regulatory sequence. *PNAS* 107(20):9158–9163.
- 671 13. Melnikov A, et al. (2012) Systematic dissection and optimiza-
672 tion of inducible enhancers in human cells using a massively
673 parallel reporter assay. *Nature Biotechnology* 30(3):271–277.
- 674 14. Kheradpour P, et al. (2013) Systematic dissection of regulatory
675 motifs in 2000 predicted human enhancers using a massively
676 parallel reporter assay. *Genome Research* 23(5):800–811.
- 677 15. Patwardhan RP, et al. (2012) Massively parallel functional dis-
678 section of mammalian enhancers in vivo. *Nature Biotechnology*
679 30(3):265–270.
- 680 16. Sharon E, et al. (2012) inferring gene regulatory logic from
681 high-throughput measurements of thousands of systematically
682 designed promoters. *Nature Biotechnology* 30(6):521–530.
- 683 17. Kosuri S, et al. (2013) Composability of regulatory sequences
684 controlling transcription and translation in *Escherichia coli*.
685 *PNAS* 110(34):14024–14029.
- 686 18. Maricque BB, Dougherty JD, Cohen BA (2017) A genome-
687 integrated massively parallel reporter assay reveals DNA se-
688 quence determinants of cis-regulatory activity in neural cells.
689 *Nucleic Acids Research* 45(4):e16–e16.
- 690 19. Fulco CP, et al. (2016) Systematic mapping of functional en-
691 hancer–promoter connections with CRISPR interference. *Sci-*
692 *ence* 354(6313):769–773.
- 693 20. Mittler G, Butter F, Mann M (2009) A SILAC-based DNA pro-
694 tein interaction screen that identifies candidate binding proteins
695 to functional DNA elements. *Genome Research* 19(2):284–293.
- 696 21. Mirzaei H, et al. (2013) Systematic measurement of trans-
697 cription factor-DNA interactions by targeted mass spectrom-
etry identifies candidate gene regulatory proteins. *PNAS* 698
110(9):3645–3650. 699
22. Lutz R, Bujard H (1997) Independent and tight regulation of
700 transcriptional units in *Escherichia coli* via the LacR/O, the
701 TetR/O and AraC/I1-I2 regulatory elements. *Nucleic acids*
702 *research* 25(6):1203–1210. 703
23. Mustonen V, Kinney J, Callan CG, Lassig M (2008) Energy-
704 dependent fitness: A quantitative model for the evolu-
705 tion of yeast transcription factor binding sites. *PNAS*
706 105(34):12376–12381. 707
24. Ireland WT, Kinney JB (2016) MPAthic: quantitative mod-
708 eling of sequence-function relationships for massively parallel
709 assays. *bioRxiv* p. 054676. 710
25. Schneider TD, Stephens RM (1990) Sequence logos: a new
711 way to display consensus sequences. *Nucleic Acids Research*
712 18(20):6097–6100. 713
26. Ong SE, et al. (2002) Stable isotope labeling by amino acids
714 in cell culture, SILAC, as a simple and accurate approach
715 to expression proteomics. *Molecular & Cellular Proteomics*
716 1(5):376–386. 717
27. Oehler S, Eismann ER, Krämer H, Müller-Hill B (1990) The
718 three operators of the lac operon cooperate in repression. *The*
719 *EMBO Journal* 9(4):973–979. 720
28. Gerdes K, Christensen SK, Løbner-Olesen A (2005) Prokary-
721 otic toxin–antitoxin stress response loci. *Nature Reviews Mi-*
722 *crobiology* 2(5):371–382. 723
29. Alekshun MN, Levy SB (1997) Regulation of chromosomally
724 mediated multiple antibiotic resistance: the *mar* regulon.
725 *Journal of Molecular Biology* 41(10):2067–2075. 726
30. Garcia HG, Phillips R (2011) Quantitative dissection
727 of the simple repression input-output function. *PNAS*
728 108(29):12173–12178. 729
31. Maisonneuve E, Gerdes K (2014) Molecular Mechanisms Under-
730 lying Bacterial Persisters. *Cell* 157(3):539–548. 731
32. Overgaard M, Borch J, Gerdes K (2013) Bacterial Toxin
732 RelE: A Highly Efficient Ribonuclease with Exquisite Substrate
733 Specificity Using Atypical Catalytic Residues. *Biochemistry*
734 52(48):8633–8642. 735
33. Overgaard M, Borch J, Gerdes K (2009) RelB and RelE of
736 *Escherichia coli* Form a Tight Complex That Represses Tran-
737 scription via the Ribbon–Helix–Helix Motif in RelB. *Journal*
738 *of Molecular Biology* 394(2):183–196. 739
34. Martin RG, Rosner JL (1997) Fis, an accessory factor for tran-
740 scriptional activation of the *mar* (multiple antibiotic resistance)
741 promoter of *Escherichia coli* in the presence of the activator
742 MarA, SoxS, or Rob. *Journal of Bacteriology* 179(23):7410–
743 7419. 744
35. Seoane AS, Levy SB (1995) Characterization of MarR, the
745 repressor of the multiple antibiotic resistance (*mar*) operon in
746 *Escherichia coli*. *Journal of Bacteriology* 177(12):3414–3419. 747
36. Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quan-
748 tifying Absolute Protein Synthesis Rates Reveals Principles
749 Underlying Allocation of Cellular Resources. *Cell* 157(3):624–
750 635. 751
37. Kuhlman T, Zhang Z, Saier MH, Hwa T (2007) Combinatorial
752 transcriptional control of the lactose operon of *Escherichia coli*.
753 *PNAS* 104(14):6043–6048. 754
38. Li GY, Zhang Y, Inouye M, Ikura M (2008) Structural Mecha-
755 nism of Transcriptional Autorepression of the *Escherichia coli*
756 RelB/RelE Antitoxin/Toxin Module. *Journal of Molecular*
757 *Biology* 380(1):107–119. 758
39. Overgaard M, Borch J, Jørgensen MG, Gerdes K (2008) Mes-
759 senger RNA interferase RelE controls *relBE* transcription by
760 conditional cooperativity. *Molecular Microbiology* 69(4):841–
761 857. 762
40. Marbach D, et al. (2012) Wisdom of crowds for robust gene
763 network inference. *Nature Methods* 9(8):796–804. 764

- 765 41. Schmidt A, et al. (2016) The quantitative and condition-
766 dependent *Escherichia coli* proteome. *Nature Biotechnology*
767 34(1):104–111.
- 768 42. Rolfes RJ (2006) Regulation of purine nucleotide biosynthesis:
769 in yeast and beyond. *Biochemical Society transactions* 34(Pt
770 5):786–790.
- 771 43. Cho BK, et al. (2011) The PurR regulon in *Escherichia coli*
772 K-12 MG1655. *Nucleic Acids Research* 39(15):6456–6464.
- 773 44. Lomba MR, Vasconcelos AT, Pacheco ABF, Almeida DF (1997)
774 Identification of *yebG* as a DNA damage-inducible *Escherichia*
775 *coli* gene. *FEMS Microbiology Letters* 156(1):119–122.
- 776 45. Wade JT, Reppas NB, Church GM, Struhl K (2005) Genomic
777 analysis of LexA binding reveals the permissive nature of the
778 *Escherichia coli* genome and identifies unconventional target
779 sites. *Genes and Development* 19(21):2619–2630.
- 780 46. Choi KY, Zalkin H (1992) Structural characterization and
781 corepressor binding of the *Escherichia coli* purine repressor.
782 *Journal of Bacteriology* 174(19):6207–6214.
- 783 47. Bintu L, et al. (2005) Transcriptional regulation by the num-
784 bers: models. *Current Opinion in Genetics and Development*
785 15(2):116–124.
- 786 48. Atwal GS, Kinney JB (2016) Learning Quantitative Sequence-
787 Function Relationships from Massively Parallel Experiments.
788 *Journal of Statistical Physics* 162(5):1203–1243.
- 789 49. Song S, Park C (1997) Organization and regulation of the
790 D-xylose operons in *Escherichia coli* K-12: XylR acts as a
791 transcriptional activator. *Journal of Bacteriology* 179(22):7025–
792 7032.
- 793 50. Browning DF, Busby SJW (2016) Local and global regula-
794 tion of transcription initiation in bacteria. *Nature Reviews*
795 *Microbiology* 14(10):638–650.
- 796 51. Laikova ON, Mironov AA, Gelfand MS (2001) Computational
797 analysis of the transcriptional regulation of pentose utilization
798 systems in the gamma subdivision of Proteobacteria. *FEMS*
799 *microbiology letters* 205(2):315–322.
- 800 52. Cooper R (1978) The utilisation of D-galactonate and D-2-oxo-
801 3-deoxygalactonate by *Escherichia coli* K-12. Biochemical and
802 genetical studies. *Archives of Microbiology* 1(118):199–206.
- 803 53. Kuhlman TE, Cox EC (2010) Site-specific chromosomal inte-
804 gration of large synthetic constructs. *Nucleic Acids Research*
805 38(6):e92–e92.
- 806 54. Vvedenskaya IO, et al. (2015) Massively Systematic Transcript
807 End Readout, “MASTER”: Transcription Start Site Selection,
808 Transcriptional Slippage, and Transcript Yields. *Molecular*
809 *Cell* 60(6):953–965.
- 810 55. Datsenko KA, Wanner BL (2000) One-step inactivation of chro-
811 mosomal genes in *Escherichia coli* K-12 using PCR products.
812 *PNAS* 97(12):6640–6645.
- 813 56. Razo-Mejia M, et al. (2014) Comparison of the theoretical and
814 real-world evolutionary potential of a genetic circuit. *Physical*
815 *Biology* 11(2):026005.
- 816 57. Kalli A, Hess S (2011) Effect of mass spectrometric param-
817 eters on peptide and protein identification rates for shotgun
818 proteomic experiments on an LTQ-orbitrap mass analyzer.
819 *Proteomics* 12(1):21–31.
- 820 58. Cox J, et al. (2009) A practical guide to the MaxQuant com-
821 putational platform for SILAC-based quantitative proteomics.
822 *Nature Protocols* 4(5):698–705.
- 823 59. Okuda S, et al. (2017) jPOSTrepo: an international stan-
824 dard data repository for proteomes. *Nucleic Acids Research*
825 45(D1):D1107–D1111.