Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time

William T. Ireland[1], Suzannah M. Beeler[2] Emanuel Flores-Bautista[2], Nathan M. Belliveau[2,†], Michael J. Sweredoski[3], Annie Moradian[3], Justin B. Kinney[4], Rob Phillips[1,2,5,*]

[1] Department of Physics, California Institute of Technology, Pasadena, CA 91125
[2] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125
[3] Proteome Exploration Laboratory, Beckman Institute, California Institute of Technology, Pasadena, CA 91125
[4] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[5] Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125
[†] Present address: Howard Hughes Medical Institute and Department of Biology, University of Washington, Seattle, WA 98195
[*] Corresponding author: phillips@pboc.caltech.edu

Abstract

Advances in DNA sequencing have revolutionized our ability to read genomes. However, even in the most well-studied of organisms, the bacterium *Escherichia coli*, for $\approx 65\%$ of the promoters we remain completely ignorant of their regulation. Until we have cracked this regulatory Rosetta Stone, efforts to read and write genomes will remain haphazard. We introduce a new method (Reg-Seq) linking a massively-parallel reporter assay and mass spectrometry to produce a base pair resolution dissection of more than 100 promoters in *E. coli* in 12 different growth conditions. First, we show that our method recapitulates regulatory information from known sequences. Then, we examine the regulatory architectures for more than 80 promoters in the *E. coli* genome which previously had no known regulation. In many cases, we also identify which transcription factors mediate their regulation. The method introduced here clears a path for fully characterizing the regulatory genome of model organisms, with the potential of moving on to an array of other microbes of ecological and medical relevance.

1 Introduction

DNA sequencing is as important to biology as the telescope is to astronomy. We are now living in the age of genomics, where DNA sequencing has become cheap and routine. However, despite these incredible advances, how all of this genomic information is regulated and deployed remains largely enigmatic. Organisms must respond to their environments through regulation of genes. Genomic methods often provide a "parts" list but often leave us uncertain about how those parts are used creatively and constructively in space and time. Yet, we know that promoters apply all-important dynamic logical operations that control when and where genetic information is accessed. In this paper, we demonstrate how we can infer the logical and regulatory interactions that

control bacterial decision making by tapping into the power of DNA sequencing as a biophysical tool. The method introduced here provides a framework for solving the problem of deciphering the regulatory genome by connecting perturbation and response, mapping information flow from individual nucleotides in a promoter sequence to downstream gene expression, determining how much information each promoter base pair carries about the level of gene expression.

The advent of RNA-Seq [1] launched a new era in which sequencing could be used as an experimental read-out of the biophysically interesting counts of mRNA, rather than simply as a tool for collecting ever more complete organismal genomes. The slew of 'X'-Seq technologies that are available continues to expand at a dizzying pace, each serving their own creative and insightful role: RNA-Seq, ChIP-Seq, Tn-Seq, SELEX, 5C, etc. [2]. In contrast to whole genome screening sequencing approaches, such as Tn-Seq [3] and ChIP-Seq [4] which give a coarse-grained view of gene essentiality and regulation respectively, another class of experiments known as massively-parallel reporter assays (MPRA) has been used to study gene expression in a variety of contexts [5, 6, 7, 8, 9, 10, 11, 12]. One elegant study relevant to the bacterial case of interest here by [13] screened more than 10^4 combinations of promoter and ribosome binding sites (RBS). Even more recently, they have utilized MPRAs in sophisticated ways to search for regulated genes across the genome [14, 15], in a way we see as being complementary to our own. While their approach yields a coarse-grained view of where regulation may be occurring, our approach yields a base-pair-by-base-pair view of how exactly that regulation is being enacted.

One of the most exciting X-Seq tools based on MPRAs with broad biophysical reach is the Sort-Seq approach developed by [6]. Sort-Seq uses fluorescence activated cell sorting (FACS) based on changes in the fluorescence due to mutated promoters to identify the specific locations of transcription factor binding in the genome. Importantly, it also provides a readout of how promoter sequences control the level of gene expression with single base-pair resolution. The results of such a massively-parallel reporter assay make it possible to build a biophysical model of gene regulation to uncover how previously uncharacterized promoters are regulated. In particular, high-resolution studies like those described here yield quantitative predictions about promoter organization and protein-DNA interactions as described by energy matrices [6]. This allows us to employ the tools of statistical physics to describe the input-output properties of each of these promoters which can be explored much further with in-depth experimental dissection like those done by [16] and [17] and summarized in [18]. In this sense, the Sort-Seq approach can provide a quantitative framework to not only discover and quantitatively dissect regulatory interactions at the promoter level, but also provides an interpretable scheme to design genetic circuits with a desired expression output [19].

Earlier work from [20] illustrated how Sort-Seq, used in conjunction with mass spectrometry can be used to identify which transcription factors bind to a given binding site, thus enabling the mechanistic dissection of promoters which previously had no regulatory annotation. However, a crucial drawback of the approach of [20] is that while it is high-throughput at the level of a single gene and the number of promoter variants it accesses, it was unable to readily tackle multiple genes at once, still leaving much of the unannotated genome untouched. Given that even in one of biology's best understood organisms, the bacterium *Escherichia coli*, for more than 65% of its genes, we remain completely ignorant of how those genes are regulated [21, 20]. If we hope to some day have a complete base pair resolution mapping of how genetic sequences relate to biological function, we must first be able to do so for the promoters of this "simple" organism.

What has been missing in uncovering the regulatory genome in organisms of all kinds is a large scale method for inferring genomic logic and regulation. Here we replace the low-throughput fluorescence-based Sort-Seq approach with a scalable RNA-Seq based approach that makes it possible to attack multiple promoters at once, setting the stage for the possibility of, to first approximation, uncovering the entirety of the regulatory genome. Accordingly, we refer to the entirety of our approach (MPRA, information footprints and energy matrices, mass spectrometry for transcription factor identification) as Reg-Seq, which we employ here on over one hundred promoters. The concept of MPRA methods is to perturb promoter regions by mutating them and then using sequencing to read out both perturbation and the resulting gene expression [5, 6, 7, 8, 9, 10, 11, 12]. We generate a broad diversity of promoter sequences for each promoter of interest and use mutual information as a metric to measure information flow from that distribution of sequences to gene expression. Thus, Reg-Seq is able to collect causal information about candidate regulatory sequences that is then complemented by mass spectrometry which allows us to find which transcription factors mediate the action of those newly discovered candidate regulatory sequences. Hence, Reg-Seq solves the causal problem of linking DNA sequence to regulatory logic and information flow.

To demonstrate our ability to scale up Sort-Seq with the sequencing based Reg-Seq protocol, we report here our results for 113 *E. coli* genes, whose regulatory architectures (i.e. gene-by-gene distributions of transcription factor (TF) binding sites and identities of TFs that bind those sites) were determined in parallel. By taking the Sort-Seq approach from a gene-by-gene method to a more whole-genome approach, we can begin to piece together not just how individual promoters are regulated, but also the nature of gene-gene interactions by revealing how certain transcription factors serve to regulate multiple genes at once. This approach has the benefits of a high-throughput assay while sacrificing little of the resolution afforded by the previous gene-by-gene approach, allowing us to uncover a large swath of the *E. coli* regulome, with base-pair resolution, in one set of experiments.

The organization of the remainder of the paper is as follows. In the Results section, we provide a global view of the discoveries we made in our exploration of more than 100 promoters in *E. coli* using Reg-Seq. These results are described in summary form in the paper itself, with a full online version of the results (www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) showing how different growth conditions elicit different regulatory responses. This section also follows the overarching view of our results by examining several biological stories that emerge from our data and serve as case studies in what has been revealed in our efforts to uncover the regulatory genome. The Discussion section summarizes the method and the current round of discoveries it has afforded with an eye to future applications to further elucidate the *E. coli* genome and opening up the quantitative dissection of other non-model organisms. Lastly, in the Methods section and fleshed out further in the Supplementary Information, we describe our methodology and benchmark it against our own earlier Sort-Seq experiments to show that using RNA-Seq as a readout of the expression of mutated promoters is equally reliable as the fluorescence-based approach.

2 Results

2.1 Selection of genes and methodology

As shown in Figure 1, we have considered more than 100 genes from across the *E. coli* genome. Our choices were based on a number of factors (see Sections 1.1 and 1.2 of the SI for more details); namely, we wanted a subset of genes that served as a "gold standard" for which the hard work of generations of molecular biologists have yielded deep insights into their regulation. The set includes *lacZYA*, *znuCB*, *znuA*, *ompR*, *araC*, *marR*, *relBE*, *dgoR*, *dicC*, *ftsK*, *xylA*, *xylF*, *dpiBA*, *rspA*, *dicA*, and *araAB*. By using Reg-Seq on these genes we were able to demonstrate that this method recovers not only what was already known about binding sites and transcription factors for well-characterized promoters, but also whether there are any important differences between the results of the methods presented here and the previous generation of experiments based on fluorescence and cell-sorting as a readout of gene expression. These promoters of known regulatory architecture are complemented by an array of previously uncharacterized genes that we selected in part using data from a recent proteomic study, in which mass spectrometry was used to measure the copy number of different proteins in 22 distinct growth conditions [22]. We selected genes that exhibited a wide variation in their copy number over the different growth conditions considered, reasoning that differential expression across growth conditions implies that those genes are under regulatory control.

As noted in the introduction, the original formulation of Reg-Seq termed Sort-Seq was based on the use of fluorescence activated cell sorting one gene at a time as a way to uncover putative binding sites for previously uncharacterized promoters [20]. As a result, as shown in Figure 2 we have formulated a second generation version that permits a high-throughput interrogation of the genome. A comparison between the Sort-Seq and Reg-Seq approaches for the same genes is shown in Supplemental Figure S1. In the Reg-Seq approach, for each promoter interrogated, we generate a library of mutated variants and design each variant to express an mRNA with a unique sequence barcode. By counting the frequency of each expressed barcode using RNA-Seq, we can assess the differential expression from our promoter of interest based on the base-pair-by-base-pair sequence of its promoter. Using the mutual information between mRNA counts and sequences, we develop an information footprint that reveals the importance of different bases in the promoter region to the overall level of expression. We locate potential transcription factor binding regions by looking for clusters of base pairs that have a significant effect on gene expression. Further details on how potential binding sites are identified are found in the Methods section. Blue regions of the histogram shown in the information footprints of Figure 2 correspond to hypothesized activating sequences and red regions of the histogram correspond to hypothesized repressing sequences. With the information footprint in hand, we can then determine energy matrices and sequence logos (described in the next section). Given putative binding sites, we construct oligonucleotides that serve as fishing hooks to fish out the transcription factors that bind to those putative binding sites using DNA-affinity chromatography and mass spectrometry [24]. Given all of this information, we can then formulate a schematized view of the newly discovered regulatory architecture of the previously uncharacterized promoter. For the case schematized in Figure 2, the experimental pipeline yields a complete picture of a simple repression architecture (i.e. a gene regulated by a single binding site for a repressor).

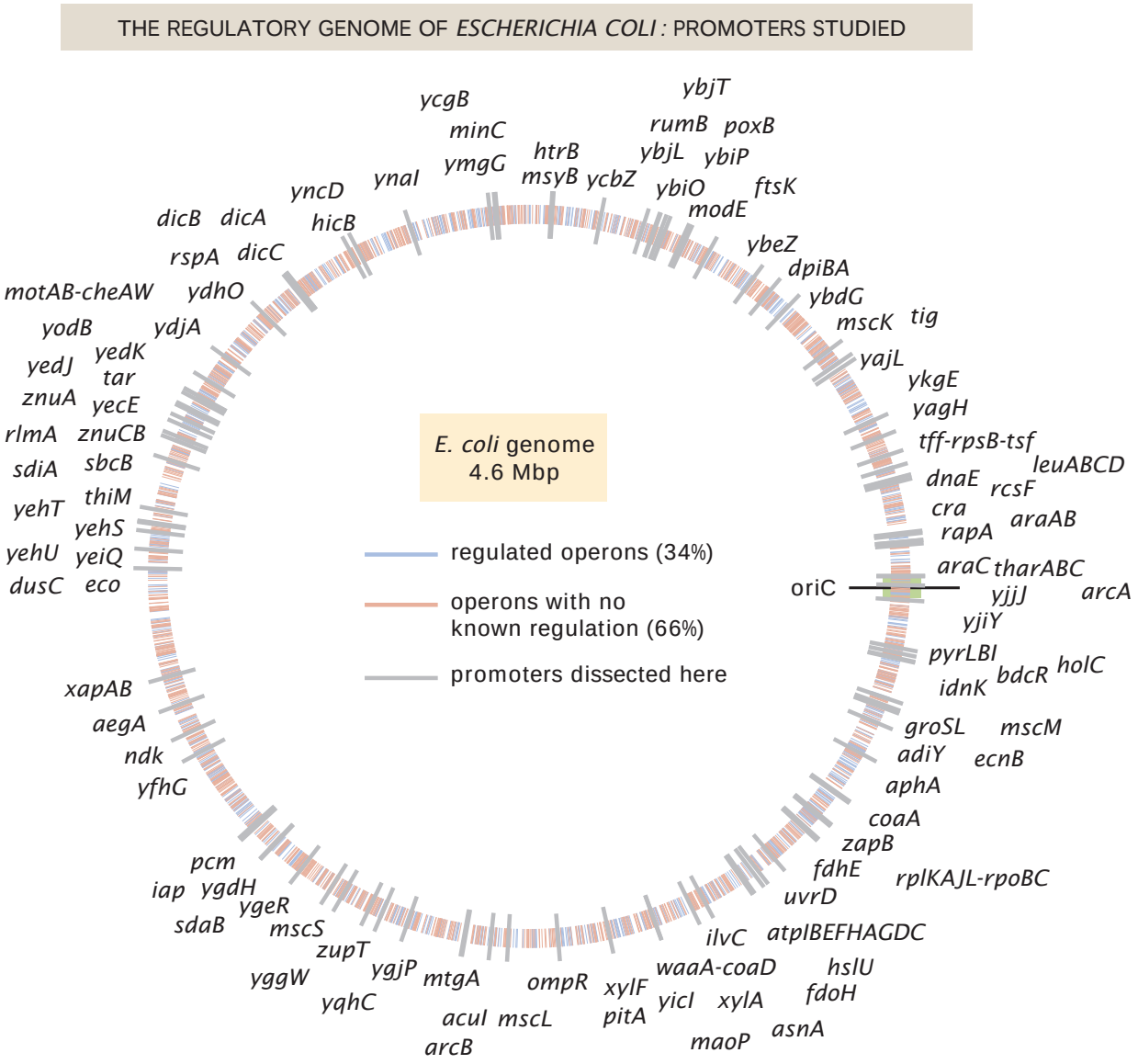

Figure 1: The *E. coli* regulatory genome. Illustration of the current ignorance with respect to how genes are regulated in *E. coli*, with genes with previously annotated regulation (as reported on RegulonDB [23]) denoted with blue ticks and genes with no previously annotated regulation denoted with red ticks. The 113 genes explored in this study are labeled in gray.

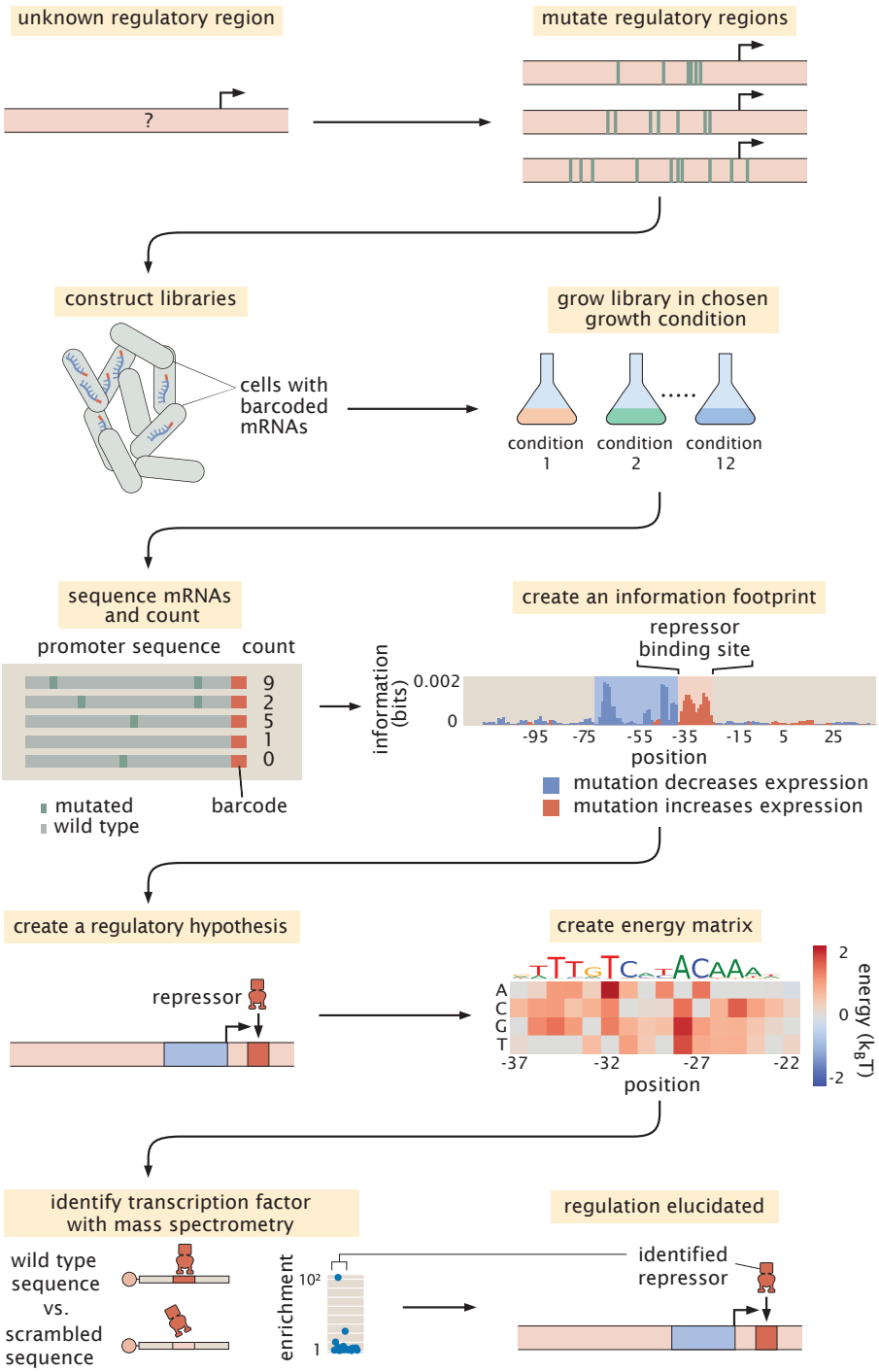

Figure 2: The Reg-Seq procedure used to determine how a given promoter is regulated. This process is as follows: After constructing a promoter library driving expression of a randomized barcode (an average of 5 for each promoter), RNA-Seq is conducted to determine frequency of these mRNA barcodes across different growth conditions. By computing the mutual information between DNA sequence and mRNA barcode counts for each base pair in the promoter region, an "information footprint" is constructed yielding a regulatory hypothesis for the putative binding sites. Energy matrices, which describe the effect any given mutation has on DNA binding energy, and sequence logos are inferred for the putative transcription factor binding sites. Next, we identify which transcription factor preferentially binds to the putative binding site via DNA affinity chromatography followed by mass spectrometry. Finally, this procedure culminates in a coarse-grained cartoon-level view of our regulatory hypothesis for how this given promoter is regulated.

2.2 Visual tools for data presentation

Throughout our investigation of the more than 100 genes explored in this study, we repeatedly relied on several key approaches to help make sense of the immense amount of data generated in these experiments. As these different approaches to viewing the results will appear repeatedly throughout the paper, here we familiarize the reader with five graphical representations referred to respectively as information footprints, energy matrices, sequence logos, mass spectrometry enrichment plots and regulatory cartoons, which taken all together provide a quantitative description of previously uncharacterized promoters.

Information footprints: From our mutagenized libraries of promoter regions, we can build up a base-pair-by-base-pair graphical understanding of how the promoter sequence relates to level of gene expression in the form of the information footprint shown in the middle of Figure 2. In this plot, the bar above each base pair position represents how large of an effect mutations at this location have on the level of gene expression. Specifically, the quantity plotted is the mutual information I_b at base pair b between mutation of a base pair at that position and the level of expression. In mathematical terms, the mutual information measures how much the joint probability $p(m, \mu)$ differs from the product of the probabilities $p_{mut}(m)p_{expr}(\mu)$ which would be produced if mutation and gene expression level were independent. Formally, the mutual information between having a mutation at position b and level of expression is given by

$$I_b = \sum_{m=0}^{1} \sum_{\mu=0}^{1} p(m, \mu) \log_2 \left(\frac{p(m, \mu)}{p_{mut}(m) p_{expr}(\mu)} \right). \tag{1}$$

Note that both m and μ are binary variables that characterize the mutational state of the base of interest and the level of expression, respectively. Specifically, m can take the values

$$m = \begin{cases} 0, & \text{if } b \text{ is a mutated base} \\ 1, & \text{if } b \text{ is a wild-type base.} \end{cases} \tag{2}$$

and μ can take on values

$$\mu = \begin{cases} 0, & \text{for sequencing reads from the DNA library} \\ 1, & \text{for sequencing reads originating from mRNA,} \end{cases} \tag{3}$$

where both m and μ are index variables that tell us whether the base has been mutated and if so, how likely that the read at that position will correspond to an mRNA, reflecting gene expression or a promoter, reflecting a member of the library. The higher the ratio of mRNA to DNA reads at a given base position, the higher the expression. $p_{mut}(m)$ in equation 1 refers to the probability that a given sequencing read will be from a mutated base. $p_{expr}(\mu)$ is a normalizing factor that gives the ratio of the number of DNA or mRNA sequencing counts to total number of counts.

Furthermore, we color the bars based on whether mutations at this location lowered gene expression on average (in blue, indicating an activating role) or increased gene expression (in red, indicating a repressing role). Within these footprints, we look for regions of approximately 10 to 20 contiguous base pairs which impact gene expression similarly (either increasing or decreasing), as these regions implicate the influence of a transcription factor binding site. In this experiment, we targeted the regulatory regions based on a guess of where a transcription start site (TSS) will

be, based on experimentally confirmed sites contained in regulonDB [21], a 5′ RACE experiment [25], or by targeting small intergenic regions. After completing the Reg-Seq experiment, we note that many of the presumed TSS sites are not in the locations assumed, the promoters have multiple active RNA polymerase (RNAP) sites and TSS, or the primary TSS shifts with growth condition. To simplify the data presentation, the '0' base pair in all information footprints is set to the originally assumed base pair for the primary TSS, rather than one of the TSS that was found in the experiment. As can be seen throughout the paper (see Figure 4 for several examples of each of the main types of regulatory architectures) and the online resource, we present such information footprints for every promoter we have considered, with one such information footprint for every growth condition.

Energy matrices: Focusing on an individual putative transcription factor binding site as revealed in the information footprint, we are interested in a more fine-grained, quantitative understanding of how the underlying protein-DNA interaction is determined. An energy matrix displays this information using a heat map format, where each column is a position in the putative binding site and each row displays the effect on binding that results from mutating to that given nucleotide (given as a change in the DNA-TF interaction energy upon mutation) [26, 27, 6]. These energy matrices are scaled such that the wild type sequence is colored in white, mutations that improve binding are shown in blue, and mutations that weaken binding are shown in red. These energy matrices encode a full quantitative picture for how we expect sequence to relate to binding for a given transcription factor, such that we can provide a prediction for the binding energy of every possible binding site sequence as

$$\text{binding energy} = \sum_{i=1}^{N} \varepsilon_i, \tag{4}$$

where the energy matrix is predicated on an assumption of a linear binding model in which each base within the binding site region contributes a specific value (ε_i for the ith base in the sequence) to the total binding energy. Energy matrices are either given in A.U. (arbitrary units), or if the gene has a simple repression or activation architecture with a single RNA polymerase (RNAP) site, are assigned k_BT energy units following the procedure in [6] and validated on the *lac* operon in [19].

Sequence logos: From an energy matrix, we can also represent a preferred transcription factor binding site with the use of the letters corresponding to the four possible nucleotides, as is often done with position weight matrices [28]. In these sequence logos, the size of the letters corresponds to how strong the preference is for that given nucleotide at that given position, which can be directly computed from the energy matrix. This method of visualizing the information contained within the energy matrix is more easily digested and allows for quick comparison among various binding sites.

Mass spectrometry enrichment plots: As the final piece of our experimental pipeline, we wish to determine the identity of the transcription factor we suspect is binding to our putative binding site that is represented in the energy matrix and sequence logo. While the details of the DNA affinity chromatography and mass spectrometry can be found in the methods, the results of these experiments are displayed in enrichment plots such as is shown in the bottom panel of Figure 2. In these plots, the relative abundance of each protein bound to our site of interest is quantified relative to a scrambled control sequence. The putative transcription factor is the one we find to be

highly enriched compared to all other DNA binding proteins.

Regulatory cartoons: The ultimate result of all these detailed base-pair-by-base-pair resolution experiments yields a cartoon model of how we think the given promoter is being regulated. A complete set of cartoons for all the architectures considered in our study is presented in Supplemental Figure S6. While the cartoon serves as a convenient visual way to summarize our results, it's important to remember that these cartoons are a shorthand representation of all the data in the four quantitative measures described above and are in fact backed by quantitative predictions of how we expect the system to behave which can be tested experimentally. Throughout this paper we use consistent iconography to illustrate the regulatory architecture of promoters, with activators and their binding sites in green, repressors in red, and RNAP in blue.

2.3 Newly discovered *E. coli* regulatory architectures

Figure 3 (and Tables 1 and 2) provides a summary of the discoveries made in the work done here using our next generation Reg-Seq approach. Figure 3(A) provides a shorthand notation that conveniently characterizes the different kinds of regulatory architectures found in bacteria. In previous work [29], we have explored the entirety of what is known about the regulatory genome of *E. coli*, revealing that the most common motif is the (0,0) constitutive architecture, though we hypothesized that this is not a statement about the facts of the *E. coli* genome, but rather a reflection of our collective regulatory ignorance in the sense that we suspect that with further investigation, many of these apparent constitutive architectures will be found to be regulated under the right environmental conditions. The two most common regulatory architectures that emerged from our previous database survey are the (0,1) and (1,0) architectures, the simple repression motif and the simple activation motif, respectively. It is interesting to consider that the (0,1) architecture is in fact the repressor-operator model originally introduced in the early 1960s by Jacob and Monod as the concept of gene regulation emerged [30]. Now we see retrospectively the far-reaching importance of that architecture across the *E. coli* genome.

For the 113 genes we considered, Figure 3(B) summarizes the number of simple repression $(0,1)$ architectures discovered, the number of simple activation $(1,0)$ architectures discovered and so on. A comparison of the frequency of the different architectures found in our study to the frequencies of all the known architectures in the RegulonDB database is provided in Supplemental Figure S7. Tables 1 and 2 provide a more detailed view of our results. As seen in Table 1, of the 113 genes we considered, 32 of them revealed no signature of any transcription factor binding sites and they are labeled as $(0,0)$. The simple repression architecture $(0,1)$ was found 26 times, the simple activation architecture $(1,0)$ was found 13 times, and more complex architectures featuring multiple binding sites (e.g. $(1,1)$, $(0,2)$, $(2,0)$, etc.) were revealed as well. Further, for 18 of the genes that we label "inactive", Reg-Seq didn't even reveal an RNAP binding site. The lack of observable RNAP site could be because the proper growth condition to get high levels of expression was not used, or because the mutation window chosen for the gene does not capture a highly transcribing TSS. The tables also include our set of 16 "gold standard" genes for which previous work has resulted in a knowledge (sometimes only partial) of their regulatory architectures. We find that our method recovers the regulatory elements of these gold standard cases fully in 12 out of 16 cases, and the majority of regulatory elements in 2 of the remainder. Overall the performance of Reg-Seq in these gold-standard cases (for more details see

Supplemental Figure S2) builds confidence in the approach. Further, the failure modes inform us of the blind spots of Reg-Seq. For example, we find it challenging to observe weaker binding sites when multiple strong binding sites are also present such as in the *marRAB* operon. Additionally the method will fail when there is no active TSS in the mutation window, as occurred in the case of *dicA*. Further details on the comparison to gold standard genes can be found in SI Section 2.2.

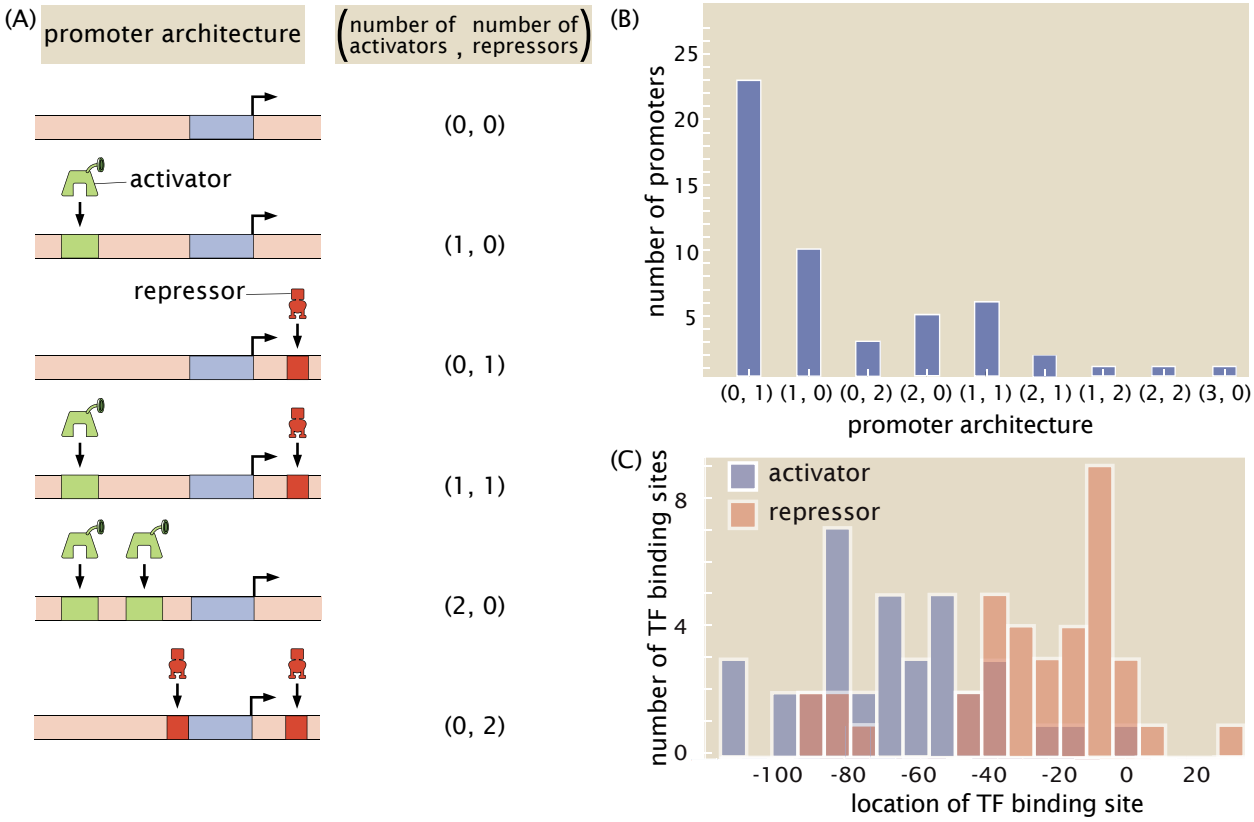

Figure 3: A summary of regulatory architectures discovered in this study. (A) The cartoons display a representative example of each type of architecture, along with the corresponding shorthand notation. (B) Counts of the different regulatory architectures discovered in this study. Only those promoters where at least one new binding site was discovered are included in this figure. If one repressor was newly discovered and two activators were previously known, then the architecture is still counted as a (2,1) architecture. (C) Distribution of positions of binding sites discovered in this study for activators and repressors. Only newly discovered binding sites are included in this figure. The position of the TF binding sites are calculated relative to the estimated TSS location, which is based on the location of the associated RNAP site.

We observe that the most common motif to emerge from our work is the simple repression motif. Another relevant regulatory statistic is shown in Figure 3(C) where we see the distribution of binding site positions. Our own experience in the use of different quantitative modeling approaches to consider transcriptional regulation reveal that, for now, we remain largely ignorant of how to account for transcription factor binding site position, and datasets like that presented here will begin to provide data that can help us uncover how this parameter dictates gene expression. Indeed, with binding site positions and energy matrices in hand, we can systematically move these binding sites and explore the implications for the level of gene expression, providing a

systematic tool to understand the role of binding-site position.

Architecture	Total number of promoters	Number of promoters with at least one newly discovered binding site
All Architectures	113	52
(0,0)	32	0
(0,1)	26	23
(1,0)	13	10
(1,1)	6	6
(0,2)	4	3
(2,0)	6	5
(2,1)	2	2
(1,2)	1	1
(2,2)	1	1
(3,0)	3	1
(0,4)	1	0
inactive	18	0

Table 1: All promoters examined in this study, categorized according to type of regulatory architecture. Those promoters which have no recognizable RNAP site are labeled as inactive rather than constitutively expressed (0, 0).

Figure 4 delves more deeply into the various regulatory architectures described in Figure 3(B) by showing several example promoters for each of the different architecture types. In each of the cases shown in the figure, prior to the work presented here, these promoters had no regulatory information in relevant databases such as Ecocyc [31] and RegulonDB [21]. Now, using the sequencing methods explained above we were able to identify candidate binding sites. For a number of cases, these putative binding sites were then used to synthesize oligonuceotide probes to enrich and identify their corresponding putative transcription factor using mass spectrometry. While Figure 4 gives a sense of the kinds of regulatory architectures we discovered in this study, our entire collection of regulatory cartoons can be found in Supplementary Figure S6.

A recent paper christened that part of the *E. coli* genome for which the function of the genes is unknown the y-ome [32]. Their surprising finding is that roughly 35% of the genes in the *E. coli* genome are functionally unannotated. The situation is likely worse for other organisms. For many of the genes in the y-ome, we remain similarly ignorant of how those genes are regulated. Figures 4 and 5 provide several examples from the y-ome, of genes and transcription factors for which little to nothing was previously known. As shown in Figure 5, our study has found the first examples that we are aware of in the entire *E. coli* genome of a binding site for YciT. These examples are intended to show the outcome of the methods developed here and to serve as an invitation to browse the online resource (www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) to see many examples of the regulation of y-ome genes.

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 0)	*acuI*	0	0	0	
	adiY	0	0	0	
	arcB	0	0	0	
	coaA	0	0	0	
	dnaE	0	0	0	
	ecnB	0	0	0	
	holC	0	0	0	
	hslUI	0	0	0	
	htrB	0	0	0	
	modE	0	0	0	
	motAB-cheAW	0	0	0	
	poxB	0	0	0	
	rcsF	0	0	0	
	rumB	0	0	0	
	sbcB	0	0	0	
	sdaB	0	0	0	
	ybdG	0	0	0	
	ybiP	0	0	0	
	ybjL	0	0	0	
	ybjT	0	0	0	
	yehS	0	0	0	
	yehT	0	0	0	
	yfhG	0	0	0	
	ygdH	0	0	0	
	ygeR	0	0	0	
	yggW	0	0	0	
	ygjP	0	0	0	
	ynaI	0	0	0	
	yqhC	0	0	0	
	zapB	0	0	0	
	zupT	0	0	0	
	amiC	0	0	0	
(0, 1)	*aegA*	1	0	0	
	bdcR	1	0	1	Known binding location (NsrR) [33]
	dicC	0	1	0	
	fdoH	1	0	0	
	groSL	1	0	0	
	idnK	1	0	1	Mass-Spectrometry (YgbI)
	leuABCD	1	0	1	Mass-Spectrometry (YgbI)

Table 2: continued on next page

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
	pcm	1	0	0	
	yedK	1	0	1	Mass-Spectrometry (TreR)
	rapA	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR)
	sdiA	1	0	0	
	tar	1	0	0	
	tff-rpsB-tsf	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
	thiM	1	0	0	
	tig	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
	ycgB	1	0	0	
	ydjA	1	0	0	
	yedJ	1	0	0	
	ycbZ	1	0	0	
	phnA	1	0	1	Mass-Spectrometry (YciT)
	mutM	1	0	0	
	rhlE	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Mass-Spectrometry (GlpR)
	uvrD	1	0	1	Bioinformatic (LexA)
	dusC	1	0	0	
	ftsK	0	1	0	
	znuA	0	1	0	
(1, 0)	*waaA-coaD*	1	0	0	
	cra	1	0	0	
	iap	1	0	0	
	araC	0	1	0	
	minC	1	0	0	
	ybeZ	1	0	0	
	mscM	1	0	0	
	mscS	1	0	0	
	rlmA	1	0	1	Bioinformatic (GcvA)
	thrLABC	1	0	0	
	yeiQ	1	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)

Table 2: continued on next page

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
	dgoR	0	1	0	Mass-Spectrometry (DgoR)
	lac	0	1	0	Mass-Spectrometry (LacI)
(0, 2)	*yecE*	2	0	1	Mass-Spectrometry (HU)
	yjjJ	2	0	1	Growth condition Knockout (MarA), Bioinformatic (MarA)
	dcm	2	0	1	Mass-Spectrometry (HNS)
	marR	0	2	0	Mass-Spectrometry (MarR)
(1, 1)	*ilvC*	2	0	1	Mass-Spectrometry (IlvY) [34]
	ybiO	2	0	0	
	yehU	2	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)
	ykgE	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry(YieP) Knockout (YieP)
	ymgG	2	0	0	
	znuCB	1	1	0	
(2, 0)	*aphA*	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (DeoR)
	arcA	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (FNR, CpxR)
	asnA	2	0	0	
	fdhE	2	0	2	Growth condition Knockout (FNR, ArcA), Bioinformatic (FNR, ArcA), Knockout (ArcA)
	xylF	0	2	0	
	mscL	2	0	0	

Table 2: continued on next page

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(2, 1)	*maoP*	3	0	3	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (PhoP, HdfR, GlpR)
	rspA	1	2	1	Mass-Spectrometry (DeoR)
(1, 2)	*dinJ*	3	0	0	
(2, 2)	*ybjX*	4	0	4	Bioinformatic (PhoP), Mass-Spectrometry (HNS, StpA)
(3, 0)	*araAB*	0	3	0	
	xylA	0	3	0	
	yicI	3	0	0	
(0, 4)	*relBE*	0	4	0	Mass-Spectrometry (RelBE)

Table 2: All genes investigated in this study categorized according to their regulatory architecture, given as (number of activators, number of repressors). The table also lists the number of newly discovered binding sites, previously known binding sites, and number of identified transcription factors. The evidence used for the transcription factor identification is given in the final column.

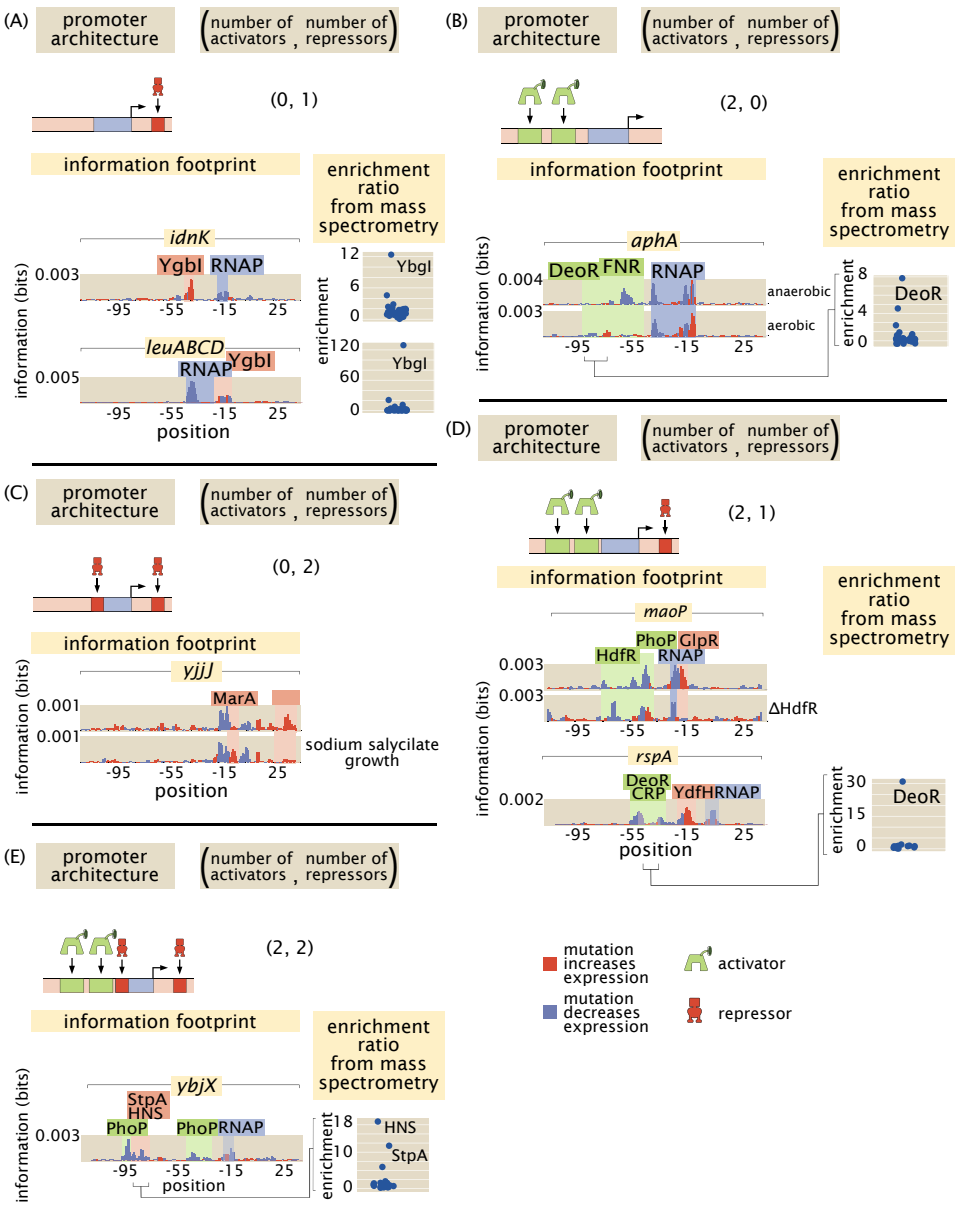

Figure 4: Newly discovered or updated regulatory architectures. Examples of information footprints, gene knockouts, and mass spectrometry data used to identify transcription factors for five genes. (A) Examples of simple repression, i.e. (0, 1) architectures where the locations of the putative binding sites are highlighted in red and the identities of the bound transcription factors are revealed in the mass spectrometry data. (B) An example of a (2, 0) architecture. During aerobic growth FNR is inactive, but the DeoR site now has a significant effect on expression. (C) An example of a (0, 2) architecture. *yjjJ* is regulated by MarA, which is only active in growth with sodium salycilate, and an unknown repressor. (D) An example of a (2, 1) architecture. (E) An example of a (2, 2) architecture.

The ability to find binding sites for both widely acting regulators and transcription factors which may have only a few sites in the whole genome allows us to get an in-depth and quantitative view of any given promoter. As indicated in Figures 5(A) and (B), we were able to perform the relevant search and capture for the transcription factors that bind our putative binding sites. In both of these cases, we now hypothesize that these newly discovered binding site-transcription factor pairs exert their control through repression. The ability to extract the quantitative features of regulatory control through energy matrices means that we can take a nearly unstudied gene such as *ykgE*, which is regulated by an understudied transcription factor YieP, and quickly get to the point at which we can do quantitative modeling in the style that we and many others have performed on the *lac* operon [35, 36, 37, 6, 38, 39, 19, 18].

One of the revealing case studies that demonstrates the broad reach of our approach for discovering regulatory architectures is offered by the insights we have gained into two widely acting regulators, GlpR [40] and FNR [41, 42]. In both cases, we have expanded the array of promoters that they are now known to regulate. Further, these two case studies illustrate that even for widely acting transcription factors, there is a large gap in regulatory knowledge and the approach advanced here has the power to discover new regulatory motifs. The newly discovered binding sites in Figure 6(A) more than double the number of operons known to be regulated by GlpR as reported in RegulonDB [21]. We found 5 newly regulated operons in our data set, even though we were not specifically targeting GlpR regulation. Although the number of example promoters across the genome that we considered is too small to make good estimates, finding 5 regulated operons out of approximately 100 examined operons supports the claim that GlpR widely regulates and many more of its sites would be found in a full search of the genome. The regulatory roles revealed in Figure 6(A) also reinforce the evidence that GlpR is a repressor.

For the GlpR-regulated operons newly discovered here, we found that this repressor binds strongly in the presence of glucose while all other growth conditions result in greatly diminished, but not entirely abolished, binding (Figure 6(A)). As there is no previously known direct molecular interaction between GlpR and glucose and the repression is reduced but not eliminated, the derepression in the absence of glucose is likely an indirect effect. As a potential mechanism of the indirect effect, *gpsA* is known to be activated by CRP [43], and GpsA is involved in the synthesis of glycerol-3-phosphate (G3P), a known binding partner of GlpR which disables its repressive activity [44]. Thus in the presence of glucose GpsA and consequently G3P will be found in low concentration, ultimately allowing GlpR fulfill its role as a repressor.

Prior to this study, there were 4 operons known to be regulated by GlpR, each with between 4 and 8 GlpR binding sites [23], where the absence of glucose and the partial induction of GlpR was not enough to prompt a notable change in gene expression [45]. These previously explored operons seemingly are regulated as part of an AND gate, where high G3P concentration *and* an absence of glucose is required for high gene expression. By way of contrast, we have discovered operons whose regulation appears to be mediated by a single GlpR site per operon. With only a single site, GlpR functions as an indirect glucose sensor, as only the absence of glucose is needed to relieve repression by GlpR.

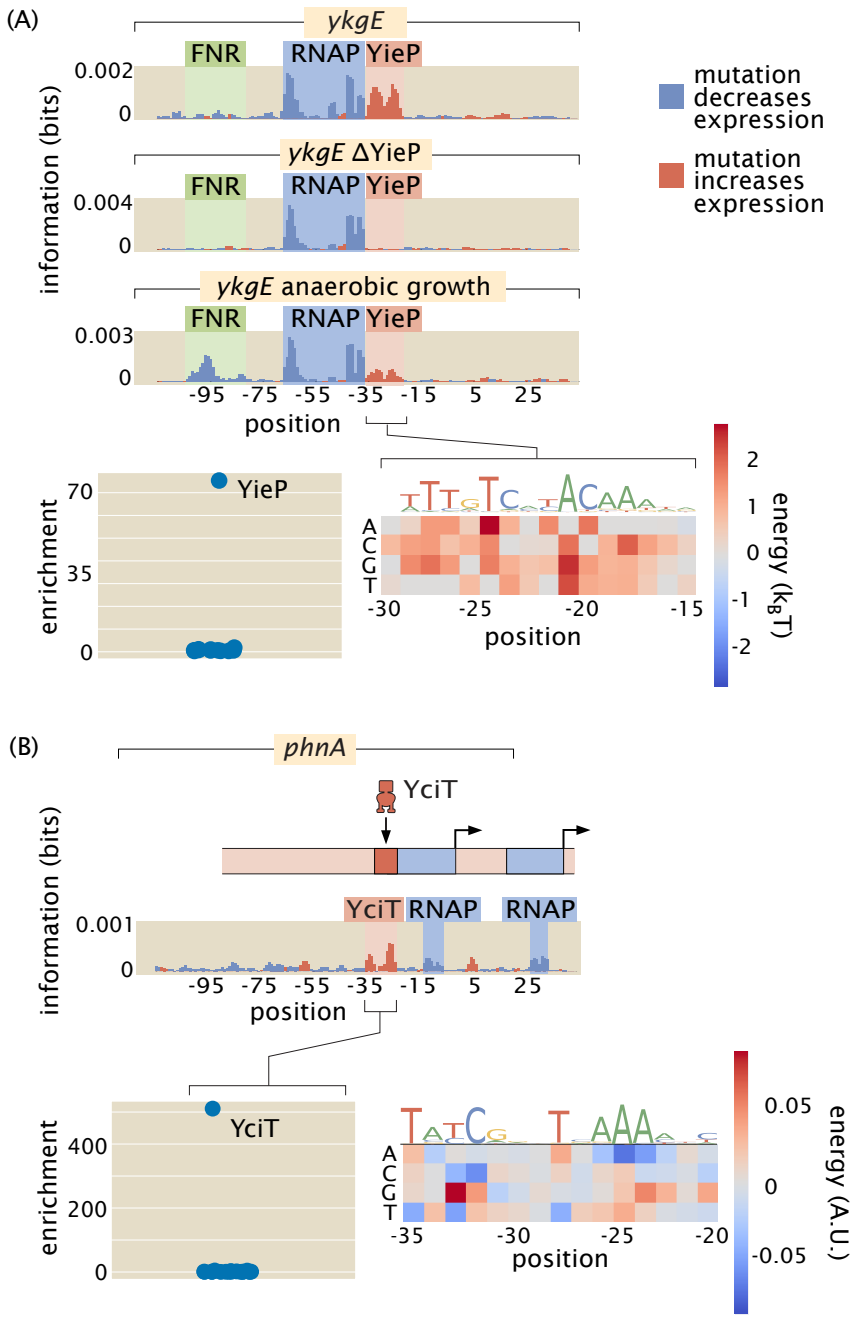

Figure 5: Examples of the insight gained by Reg-Seq in the context of promoters with no previously known regulatory information. (A) From the information footprint of the *ykgE* promoter under different growth conditions, we can identify a repressor binding site downstream of the RNAP binding site. From the enrichment of proteins bound to the DNA sequence of the putative repressor as compared to a control sequence, we can identify YieP as the transcription factor bound to this site as it has a much higher enrichment ratio than any other protein. Lastly, the binding energy matrix for the repressor site along with corresponding sequence logo shows that the wild type sequence is the strongest possible binder and it displays an imperfect inverted repeat symmetry. (B) Illustration of a comparable dissection for the *phnA* promoter.

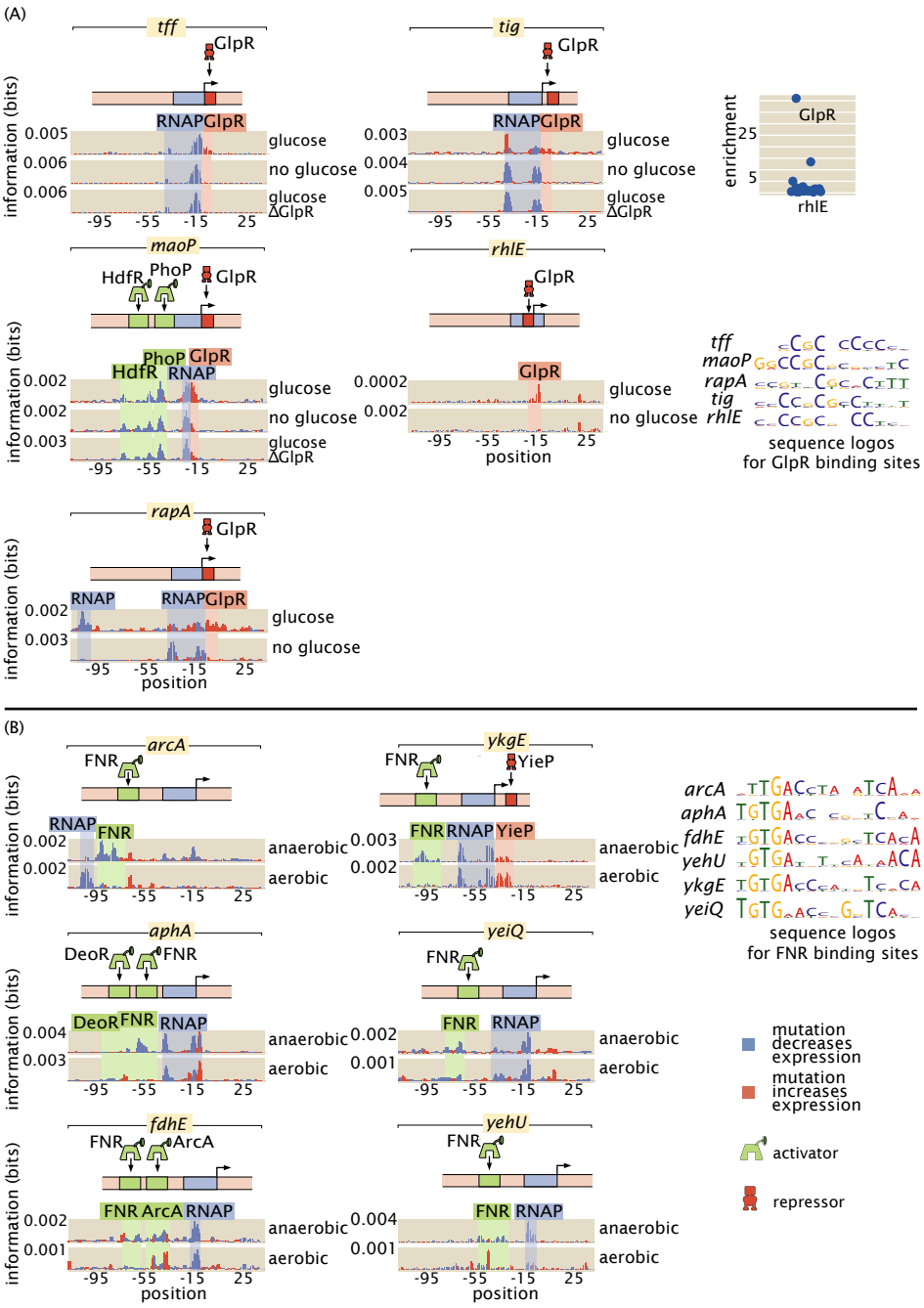

Figure 6: Reg-Seq analysis of broadly-acting transcription factors. (A) GlpR as a widely-acting regulator. Here we show the many promoters which we found to be regulated by GlpR, all of which were previously unknown. GlpR was demonstrated to bind to *rhlE* by mass spectrometry enrichment experiments as shown in the top right. Binding sites in the *tff*, *tig*, *maoP*, *rhlE*, and *rapA* have similar DNA binding preferences as seen in the sequence logos and each TF binding site binds strongly only in the presence of glucose. These similarities suggest that the same TF binds to each site. To test this hypothesis we knocked out GlpR and ran the Reg-Seq experiments for *tff*, *tig*, and *maoP*. We see that knocking out GlpR removes the binding signature of the TF. (B) FNR as a global regulator. FNR is known to be upregulated in anaerobic growth, and here we found it to regulate a suite of six genes. In growth conditions with prevalent oxygen the putative FNR sites are weakened, and the DNA binding preference of the six sites are shown to be similar from the sequence logos displayed on the right.

The second widely acting regulator our study revealed, FNR, has 151 binding sites already reported in RegulonDB and is well studied compared to most transcription factors [23]. However, the newly discovered FNR sites displayed in Figure 6(B) demonstrate that even for well-understood transcription factors there is much still to be uncovered. Our information footprints are in agreement with previous studies suggesting that FNR acts as an activator. In the presence of O_2, dimeric FNR is converted to a monomeric form and its ability to bind DNA is greatly reduced [46]. Only in low oxygen conditions did we observe a binding signature from FNR, and we show a representative example of the information footprint from one of 11 growth conditions with plentiful oxygen in Figure 6(B).

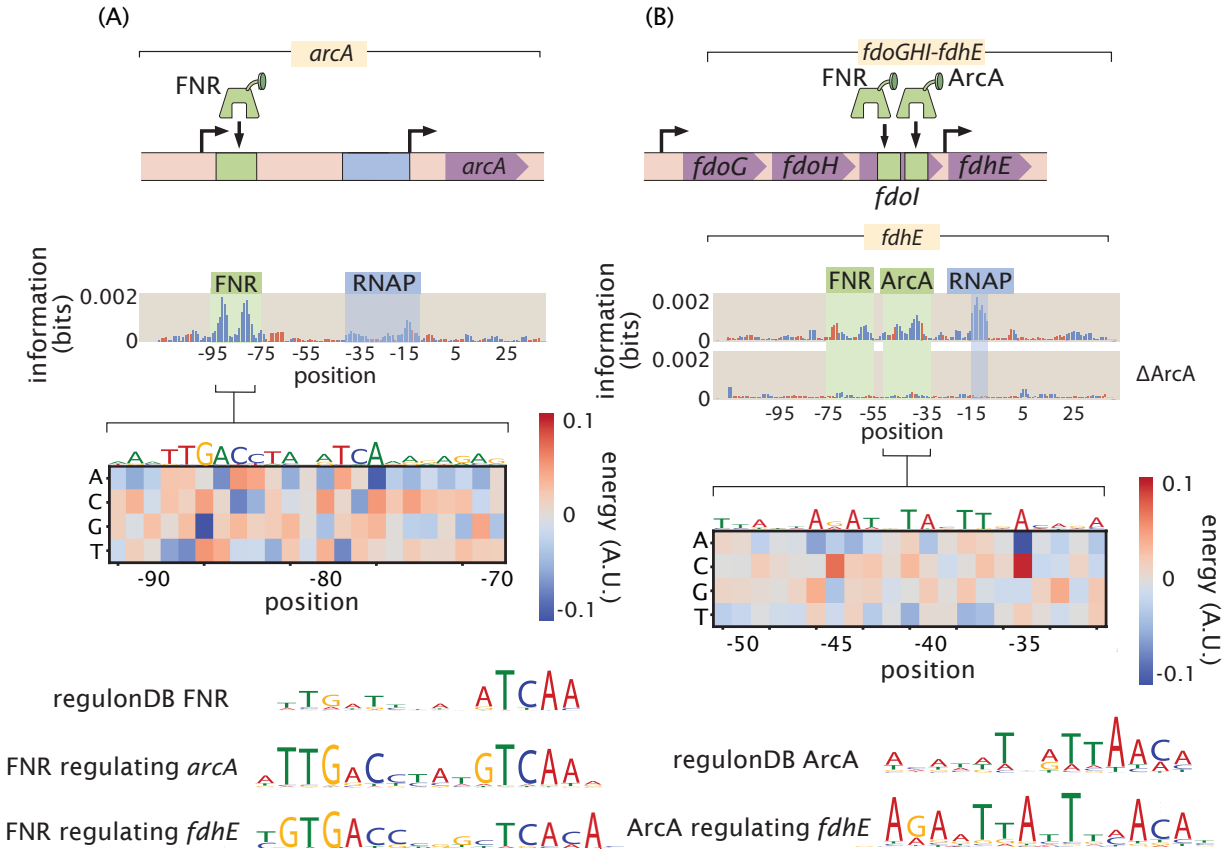

Figure 7: Inspection of an anaerobic respiration genetic circuit. (A) Here we see not only how the *arcA* promoter is regulated, but also the role this transcription factor plays in the regulation of another promoter. (B) Intra-operon regulation of *fdhE* by both FNR and ArcA. A TOMTOM [47] search of the binding motif found that ArcA was the most likely candidate for the transcription factor. A knockout of ArcA demonstrates that the binding signature of the site, and its associated RNAP site, are no longer significant determinants of gene expression.

We observe quantitatively how FNR affects the expression of *fdhE* both directly through transcription factor binding (Figure 7(A)) and indirectly through increased expression of ArcA (Figure 7(B)). Also, fully understanding even a single operon often requires investigating several regulatory regions as we have in the case of *fdoGHI-fdhE* by investigating the main promoter for the operon as well as the promoter upstream of *fdhE*. 36% of all multi-gene operons have at least one TSS which transcribes only a subset of the genes in the operon [48]. Regulation

within an operon is even more poorly studied than regulation in general. The main promoter for *fdoGHI-fdhE* has a repressor binding site, which demonstrates that there is regulatory control of the entire operon. However, we also see in Figure 7(B) that there is control at the promoter level, as *fdhE* is regulated by both ArcA and FNR and will therefore be upregulated in anaerobic conditions [49]. The main TSS transcribes all four genes in the operon, while the secondary site shown in Figure 7(B) only transcribes *fdhE*, and therefore anaerobic conditions will change the stoichiometry of the proteins produced by the operon. At the higher throughput that we use in this experiment it becomes feasible to target multiple promoters within an operon as we have done with *fdoGHI-fdhE*. We can then determine under what conditions an operon is internally regulated. Figure 7 also makes it clear that for cases such as *fdoGHI-fdhE*, there are many subtleties both in the interpretation of the information footprints and in the construction of regulatory cartoons that are simultaneously accurate and transparent. A crucial next step in the development of these analyses is to move from manual curation of the data to automated statistical analyses that can help make sense of these complicated datasets.

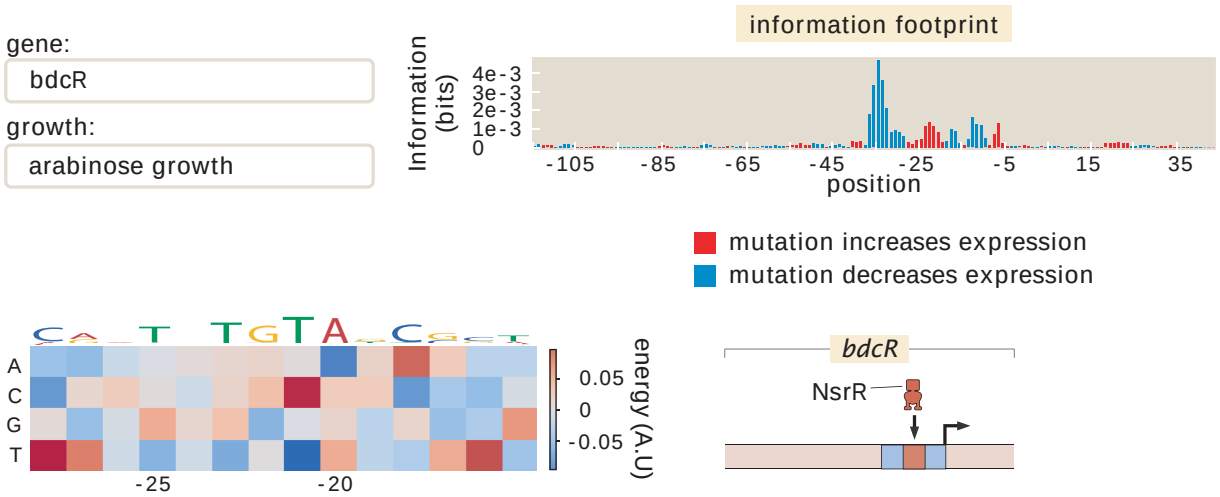

Figure 8: Representative view of the interactive figure that is available online. This interactive figure captures the entirety of our dataset. Each figure features a drop-down menu of genes and growth conditions. For each such gene and growth condition, there is a corresponding information footprint revealing putative binding sites, an energy matrix that shows the strength of binding of the relevant transcription factor to those binding sites and a cartoon that schematizes the newly-discovered regulatory architecture of that gene.

By examining the over 100 promoters considered here, grown under 12 growth conditions, we have a total of more than 1000 information footprints and data sets. In this age of big data, methods to explore and draw insights from that data are crucial. To that end, as introduced in Figure 8, we have developed an online resource (see www.rpgroup.caltech.edu/RNAseq_SortSeq/interactive_a) that makes it possible for anyone who is interested to view our data and draw their own biological conclusions. Information footprints for any combination of gene and growth condition are displayed via drop down menus. Each identified transcription factor binding site or transcription start site is marked, and energy matrices for all transcription factor binding sites are displayed. In addition, for each gene, we feature a simple cartoon-level schematic that captures our now current best understanding of the regulatory architecture and resulting mechanism.

The interactive figure in question was invaluable in identifying transcription factors, such as GlpR, whose binding properties vary depending on growth condition. As sigma factor availability also varies greatly depending on growth condition, studying the interactive figure identified many of the secondary RNAP sites present. The interactive figure provides a valuable resource both to those who are interested in the regulation of a particular gene and those who wish to look for patterns in gene regulation across multiple genes or across different growth conditions.

3 Discussion

The study of gene regulation is one of the centerpieces of modern biology. As a result, it is surprising that in the genome era, our ignorance of the regulatory landscape in even the best-understood model organisms remains so vast. Despite understanding the regulation of transcription initiation in bacterial promoters [50], and how to tune their expression, we lack an experimental framework to unravel understudied promoter architectures at scale. As such, in our view one of the grand challenges of the genome era is the need to uncover the regulatory landscape for each and every organism with a known genome sequence. Given the ability to read and write DNA sequence at will, we are convinced that to make that reading of DNA sequence truly informative about biological function and to give that writing the full power and poetry of what Crick christened "the two great polymer languages", we need a full accounting of how the genes of a given organism are regulated and how environmental signals communicate with the transcription factors that mediate that regulation – the so-called "allosterome" problem [51]. The work presented here provides a general methodology for making progress on the former problem and also demonstrates that, by performing Reg-Seq in different growth conditions, we can make headway on the latter problem as well.

The advent of cheap DNA sequencing offers the promise of beginning to achieve that grand challenge goal in the form of MPRAs reviewed in [12]. A particular implementation of such methods was christened Sort-Seq [6] and was demonstrated in the context of well understood regulatory architectures. A second generation of the Sort-Seq method [20] established experiments through the use of DNA-affinity chromatography and mass spectrometry which made it possible to identify the transcription factors that bind the putative binding sites discovered by Sort-Seq. But there were critical shortcomings in the method, not least of which was that it lacked the scalability to uncover the regulatory genome on a genome-wide basis.

The work presented here builds on the foundations laid in the previous studies by invoking RNA-Seq as a readout of the level of expression of the promoter mutant libraries needed to infer information footprints and their corresponding energy matrices and sequence logos followed by a combination of mass spectrometry and gene knockouts to identify the transcription factors that bind those sites. The case studies described in the main text showcase the ability of the method to deliver on the promise of beginning to uncover the regulatory genome systematically. The extensive online resources hint at a way of systematically reporting those insights in a way that can be used by the community at large to develop regulatory intuition for biological function and to design novel regulatory architectures using energy matrices.

However, several shortcomings remain in the approach introduced here. First, the current

implementation of Reg-Seq still largely relies on manual curation as the basis of using information footprints to generate testable regulatory hypotheses. As described in the methods section, we have also used statistical testing as a way to convert information footprints into regulatory hypotheses, but there clearly remains much work to be done on the data analysis pipeline to improve both the power and the accuracy of this approach. In addition, these regulatory hypotheses can also be converted into gene regulatory models using statistical physics [52, 37]. However, here too, as the complexity of the regulatory architectures increases, it will be of great interest to use automated model generation as suggested in a recent biophysically-based neural network approach [53].

A second key challenge faced by the methods described here is that the mass spectrometry and the gene knockout confirmation aspects of the experimental pipeline remain low-throughput. To overcome this, we have begun to explore a new generation of experiments such as *in vitro* binding assays that will make it possible to accomplish transcription factor identification at higher throughput. Specifically, we are exploring multiplexed mass spectrometry measurements and multiplexed Reg-Seq on libraries of gene knockouts as ways to break the identification bottleneck.

Another shortcoming of the current implementation of the method is that it would miss regulatory action at a distance. Indeed, our laboratory has invested a significant effort in exploring such long-distance regulatory action in the form of DNA looping in bacteria and VDJ recombination in jawed vertebrates. It is well known that transcriptional control through enhancers in eukaryotic regulation is central in contexts ranging from embryonic development to hematopoiesis [9]. The current incarnation of the methods described here have focused on contiguous regions in the vicinity of the transcription start site. Clearly, to go further in dissecting the entire regulatory genome, these methods will have to be extended to non-contiguous regions of the genome.

The findings from this study provide a foundation for systematically performing genome-wide regulatory dissections. We have developed a method to pass from complete regulatory ignorance to designable regulatory architectures and we are hopeful that others will adopt these methods with the ambition of uncovering the regulatory architectures that preside over their organisms of interest.

4 Methods

4.1 Library construction

Promoter variants were synthesized on a microarray (TWIST Bioscience, San Francisco, CA). The sequences were designed computationally such that each base in the 160 bp promtoter region has a 10% probability of being mutated. For each given promoter's library, we ensured that the mutation rate as averaged across all sequences was kept between 9.5% and 10.5%, otherwise the library was regenerated. There are an average of 2200 unique promoter sequences per gene (for an analysis of how our results depend upon number of unique promoter sequences see Supplementary Figure S3). An average of 5 unique 20 base pair barcodes per variant promoter was used for the purpose of counting transcripts. The barcode was inserted 110 base pairs from the 5′ end of the mRNA, containing 45 base pairs from the targeted regulatory region, 64 base pairs containing primer sites used in the construction of the plasmid, and 11 base pairs containing

a three frame stop codon. All the sequences are listed in Supplementary Table 1. Following the barcode there is an RBS and a GFP coding region. Mutated promoters were PCR amplified and inserted by Gibson assembly into the plasmid backbone of pJK14 (SC101 origin) [6]. Constructs were electroporated into *E. coli* K-12 MG1655 [54].

4.2 RNA preparation and sequencing

Cells were grown to an optical density of 0.3 and RNA was then stabilized using Qiagen RNA Protect (Qiagen, Hilden, Germany). Lysis was performed using lysozyme (Sigma Aldrich, Saint Louis, MO) and RNA was isolated using the Qiagen RNA Mini Kit. Reverse transcription was preformed using Superscript IV (Invitrogen, Carlsbad, CA) and a specific primer for the labeled mRNA. qPCR was preformed to check the level of DNA contamination and the mRNA tags were PCR amplified and Illumina sequenced. Within a single growth condition, all promoter variants for all regulatory regions were tested in a single multiplexed RNA-Seq experiment. All sequencing was carried out by either the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech (HiSeq 2500) on a 100 bp single read flow cell or using the sequencing services from NGX Bio on a 250 bp or 150 base paired end flow cell.

4.3 Analysis of sequencing results

To determine putative transcription factor binding sites, we first compute the effect of mutations on gene expression at a base pair-by-base pair level using information footprints. The information footprints are a hypothesis generating tool and we choose which regions to further investigate using techniques such as mass spectrometry by visually inspecting the data for regions of 10 to 20 base pairs that have high information content compared to background. Our technique currently relies on using human intuition to determine binding sites, but to validate these choices and to capture all regions important for gene expression we computationally identify regions where gene expression is changed significantly up or down by mutation ($p < 0.01$), and discard any potential sites which do not fit this criteria. We infer the effect of mutation using Markov Chain Monte Carlo, and we use the distribution of parameters from the inference to form a 99 % confidence interval for the average effect of mutation across a 15 base pair region. We include binding sites that are statistically significant at the 0.01 level in any of the tested growth conditions.

Many false positives will be secondary RNAP sites and we remove from consideration any sites that resemble RNAP sites. We fit energy matrices to each of the possible binding sites and use the preferred DNA sequence for binding to identify the RNAP sites. We use both visual inspection to compare the preferred sequence to known consensus sequences for each of the *E. coli* sigma factor binding sites (for example, do the preferred bases in the energy matrix have few mismatches to the TGNTATAAT extended minus 10 for σ^{70} sites), and the TOMTOM tool [47] to computationally compare the potential site to examples of σ^{70}, σ^{38}, and σ^{54} sites that we determined in this experiment. For further details see Supplementary Figure S4. We discard any sites that have a p-value of similarity with an RNAP site of less than $5x10^{-3}$ in the TOMTOM analysis or are deemed to be too visually similar to RNAP sites. If a single site contains an RNAP site along with a transcription factor site we remove only those bases containing the probable

RNAP site. This results in 95 identified transcription factor binding regions.

For primary RNAP sites, we include a list of probable sigma factor identities as Supplementary Table 2. Sites are judged by visual similarity to consensus binding sites. Those sites where the true sigma factor is unclear due to overlapping binding sites are omitted. Overlapping binding sites (from multiple TFs or RNAP sites) in general can pose issues for this method. In many cases, looking at growth conditions where only one of the relevant transcription factors is present or active is an effective way to establish site boundaries and infer correct energy matrices. For sites where no adequate growth condition can be found, or when a TF overlaps with an RNAP site, the energy matrix will not be reflective of the true DNA-protein interaction energies. If the TFs in overlapping sites are composed of one activator and one repressor, then we use the point at which the effect of mutation shifts from activator-like to repressor-like as a demarcation point between binding sites. We see a case of a potentially overlooked repressor due to overlapping sites in Figure 4(B), as there are several repressor like bases overlapping the RNAP -10 site and the effect weakens in low oxygen growth. However, due to the effect of the RNAP site, when averaged over a potential 15 base pair region, the repressor-like bases do not have a significant effect on gene expression.

4.4 DNA affinity chromatography and mass spectrometry

Upon identifying a putative transcription factor binding site, we used DNA affinity chromatography, as done in [20] to isolate and enrich for the transcription factor of interest. In brief, we order biotinylated oligos of our binding site of interest (Integrated DNA Technologies, Coralville, IA) along with a control, "scrambled" sequence, that we expect to have no specificity for the given transcription factor. We tether these oligos to magnetic streptavidin beads (Dynabeads MyOne T1; ThermoFisher, Waltham, MA), and incubate them overnight with whole cell lysate grown in the presences of either heavy (with ^{15}N) or light (with ^{14}N) lysine for the experimental and control sequences, respectively. The next day, proteins are recovered by digesting the DNA with the PtsI restriction enzyme (New England Biolabs, Ipswich, MA), whose cut site was incorporated into all designed oligos.

Protein samples were then prepared for mass spectrometry by either in-gel or in-solution digestion using the Lys-C protease (Wako Chemicals, Osaka, Japan). Liquid chromatography coupled mass spectrometry (LC-MS) was performed as previously described by [20], and is further discussed in the SI. SILAC labeling was performed by growing cells (Δ LysA) in either heavy isotope form of lysine or its natural form.

It is also important to note that while we relied on the SILAC method to identify the TF identity for each promoter, our approach doesnt require this specific technique. Specifically, our method only requires a way to contrast between the copy number of proteins bound to a target promoter in relation to a scrambled version of the promoter. In principle, one could use multiplexed proteomics based on isobaric mass tags [55] to characterize up to 10 promoters in parallel. Isobaric tags are reagents used to covalently modify peptides by using the heavy-isotope distribution in the tag to encode different conditions. The most widely adopted methods for isobaric tagging are the isobaric tag for relative and absolute quantitation (iTRAQ) and the tandem mass tag (TMT). This multiplexed approach involves the fragmentation of peptide ions by colliding with an inert gas. The resulting ions are resolved in a second MS-MS scan (MS2).

Only a subset (13) of all transcription factor targets were identified by mass spectrometry due to limitations in scaling the technique to large numbers of targets. The transcription factors identified by this method are enriched more than any other DNA binding protein, with $p < 0.01$ using the outlier detection method as outlined by Cox and Mann [56], with corrections for multiple hypothesis testing using the method proposed by Benjamini and Hochberg [57].

4.5 Construction of knockout strains

Conducting DNA affinity chromatography followed by mass spectrometry on putative binding sites resulted in potential candidates for the transcription factors that are responsible for the information contained at a given promoter region. For some cases, to verify that a given transcription factor is, in fact, regulating a given promoter, we repeated the RNA sequencing experiments on strains with the transcription factor of interest knocked out.

To construct the knockout strains, we ordered strains from the Keio collection [58] from the Coli Genetic Stock Center. These knockouts were put in a MG1655 background via phage P1 transduction and verified with Sanger sequencing. To remove the kanamycin resistance that comes with the strains from the Keio collection, we transformed in the pCP20 plasmid, induced FLP recombinase, and then selected for colonies that no longer grew on either kanamycin or ampicillin. Finally, we transformed our desired promoter libraries into the constructed knockout strains, allowing us to perform the RNA sequencing in the same context as the original experiments.

4.6 Code and Data Availability

All code used for processing data and plotting as well as the final processed data, plasmid sequences, and primer sequences can be found on the GitHub repository (https://github.com/RPGroup-PBoC/RNAseq_SortSeq) doi:10.5281/zenodo.3611914. Energy matrices were generated using the MPAthic software [59]. All raw sequencing data is available at the Sequence Read Archive (accession no.PRJNA599253). All inferred information footprints and energy matrices can be found on the CalTech data repository doi:10.22002/D1.1331. All mass spectrometry raw data is available on the CalTech data repository doi:10.22002/d1.1336

5 Acknowledgments

We are grateful to Rachel Banks, Stephanie Barnes, Curt Callan, Griffin Chure, Ana Duarte, Vahe Galstyan, Hernan Garcia, Soichi Hirokawa, Thomas Lecuit, Heun Jin Lee, Madhav Mani, Nicholas McCarty, Muir Morrison, Steve Quake, Tom Röschinger, Manuel Razo-Mejia, Gabe Salmon, and Guillaume Urtecho for useful discussion and feedback on the manuscript. Guillaume Urtecho and Sri Kosuri have been instrumental in providing key advice and protocols at various stages in the development of this work. We would like to thank Jost Vielmetter and Nina Budaeva for providing access to their Cell Disruptor. Brett Lomenick provided crucial help and advice with protein preparation. We also thank Igor Antoshechkin for his help with sequencing at the Caltech

Genomics Facility.

Funding: We are deeply grateful for support from NIH Grants DP1 OD000217 (Director's Pioneer Award) and 1R35 GM118043-01 (Maximizing Investigators Research Award) which made it possible to undertake this multi-year project. N.M.B. was supported by an HHMI International Student Research Fellowship. S.M.B was supported by the NIH Institutional National Research Service Award (5T32GM007616-38) provided through Caltech.

References

[1] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

[2] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20:257–272, 2019.

[3] Emily C. A. Goodall, Ashley Robinson, Iain G. Johnston, Sara Jabbari, Keith A. Turner, Adam F. Cunningham, Peter A. Lund, Jeffrey A. Cole, and Ian R. Henderson. The essential genome of *Escherichia coli* k-12. *mBio*, 9(1), 2018.

[4] Ye Gao, James T Yurkovich, Sang Woo Seo, Ilyas Kabimoldayev, Ke Chen, Anand V Sastry, Xin Fang, Nathan Mih, Laurence Yang, Johannes Eichner, Byung-kwan Cho, Donghyuk Kim, and Bernhard O Palsson. Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. *Nucleic Acids Research*, 46(20):10682–10696, 2018.

[5] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'er, and J. Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27(12):1173–1175, 2009.

[6] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20):9158–9163, 2010.

[7] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–30, 2012.

[8] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S. I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30(3):265–70, 2012.

[9] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G Callan Jr, Justin B Kinney, Manolis Kellis, Eric S Lander, and Tarjei S Mikkelsen. Systematic dissection and optimization of inducible

enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277, 2012.

[10] J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A*, 109(47):19498–503, 2012.

[11] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669, 2019.

[12] Justin B. Kinney and David M. McCandlish. Massively parallel assays and quantitative sequencefunction relationships. *Annual Review of Genomics and Human Genetics*, 20(1):99–127, 2019.

[13] Sriram Kosuri, Daniel B Goodman, Guillaume Cambray, Vivek K Mutalik, and Yuan Gao. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 110(34), 2013.

[14] G. Urtecho, A. D. Tripp, K. D. Insigne, H. Kim, and S. Kosuri. Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. *Biochemistry*, 58(11):1539–1551, 2019.

[15] Guillaume Urtecho, Kimberly Insigne, Arielle D Tripp, Marcia Brinck, Nathan B Lubock, Hwangbeom Kim, Tracey Chan, and Sriram Kosuri. Genome-wide functional characterization of *Escherichia col* promoters and regulatory elements responsible for their function. *bioRxiv*, 2020.

[16] Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Mitchell Lewis, and Rob Phillips. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. *Cell Systems*, 6(4):456–469.e10, 2018.

[17] Griffin Chure, Manuel Razo-Mejia, Nathan M. Belliveau, Tal Einav, Zofii A. Kaczmarek, Stephanie L. Barnes, Mitchell Lewis, and Rob Phillips. Predictive shifts in free energy couple mutations to their phenotypic consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 116(37):18275–18284, 2019.

[18] R. Phillips, N. M. Belliveau, G. Chure, H. G. Garcia, M. Razo-Mejia, and C. Scholes. Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression. *Annual Review of Biophysics*, 48:121–163, 2019.

[19] Stephanie L. Barnes, Nathan M. Belliveau, William T. Ireland, Justin B. Kinney, and Rob Phillips. Mapping DNA sequence to transcription factor binding energy *in vivo*. *PLoS Computational Biology*, 15(2):1–29, 2019.

[20] Nathan M. Belliveau, Stephanie L. Barnes, William T. Ireland, Daniel L. Jones, Michael J. Sweredoski, Annie Moradian, Sonja Hess, Justin B. Kinney, and Rob Phillips. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 115(21):E4796–E4805, 2018.

[21] Alberto Santos-Zavaleta, Heladia Salgado, Socorro Gama-castro, G Laura, Daniela Ledezma-tejeida, S Mishael, Santiago Garc, Kevin Alquicira-hern, Luis Jos, Pablo Pe, Cecilia Ishida-guti, David A Vel, Del Moral-ch, James Galagan, and Julio Collado-vides. Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli K-12*. *Nucleic Acids Research*, 47:212–220, 2019.

[22] Alexander Schmidt, Karl Kochanowski, Silke Vedelaar, Erik Ahrne, Benjamin Volkmer, Luciano Callipo, Kevin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology*, 34(1):104–110, 2015.

[23] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chavez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1):D133–D143, 2016.

[24] G. Mittler, F. Butter, and M. Mann. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res*, 19(2):284–93, 2009.

[25] Alfredo Mendoza-Vargas, Leticia Olvera, Maricela Olvera, Ricardo Grande, Leticia Vega-Alvarado, Blanca Taboada, Vernica Jimenez-Jacinto, Heladia Salgado, Katy Jurez, Bruno Contreras-Moreira, Araceli M. Huerta, Julio Collado-Vides, and Enrique Morett. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. *PLoS ONE*, 4(10):e7526, October 2009.

[26] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–50, 1987.

[27] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci*, 23(3):109–13, 1998.

[28] Thomas D. Schneider and R.Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[29] M. Rydenfelt, H. G. Garcia, R. S. Cox III, and R. Phillips. The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS One*, 9(12):e114347, 2014.

[30] FRANqOIS Jacob and Jacques Monod. On the regulation of gene activity. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:19, 1961.

[31] Ingrid M. Keseler, Amanda Mackie, Alberto Santos-Zavaleta, Richard Billington, César Bonavides-Martínez, Ron Caspi, Carol Fulcher, Socorro Gama-Castro, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne

Paley, Martin Peralta-Gil, Pallavi Subhraveti, David A. Velázquez-Ramírez, Daniel Weaver, Julio Collado-Vides, Ian Paulsen, and Peter D. Karp. The EcoCyc database: reflecting new knowledge about *escherichia coli* K-12. *Nucleic Acids Research*, 45(D1):D543–D550, 2016.

[32] Sankha Ghatak, Zachary A. King, Anand Sastry, and Bernhard O. Palsson. The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Research*, 47(5):2446–2454, 2019.

[33] Jonathan D. Partridge, Diane M. Bodenmiller, Michael S. Humphrys, and Stephen Spiro. NsrR targets in the Escherichia coli genome: new insights into DNA sequence requirements for binding and a role for NsrR in the regulation of motility. *Molecular Microbiology*, 73(4):680–694, 2009.

[34] Kyu Y. Rhee, Donald F. Senear, and G. Wesley Hatfield. Activation of Gene Expression by a Ligand-induced Conformational Change of a Protein-DNA Complex. *Journal of Biological Chemistry*, 273(18):11257–11266, May 1998.

[35] J. M. Vilar and S. Leibler. DNA looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.

[36] J. M. Vilar, C. C. Guet, and S. Leibler. Modeling network dynamics: the lac operon, a case study. *J Cell Biol*, 161(3):471–6, 2003.

[37] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jan Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124, April 2005.

[38] Hernan G. Garcia and Rob Phillips. Quantitative dissection of the simple repression input-output function. *Proceedings of the National Academy of Sciences*, 108(29):12173–12178, July 2011.

[39] J. M. Vilar and L. Saiz. Reliable prediction of complex phenotypes from a modular design in free energy space: an extensive exploration of the lac operon. *ACS Synth Biol*, 2(10):576–86, 2013.

[40] H Schweizer, W Boos, and T J Larson. Repressor for the sn-glycerol-3-phosphate regulon of Escherichia coli K-12: cloning of the glpR gene and identification of its product. *Journal of Bacteriology*, 161(2):563–566, 1985.

[41] Heinz Krner, Heidi J. Sofia, and Walter G. Zumft. Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiology Reviews*, 27(5):559–592, December 2003.

[42] Manika Kargeti and K. V. Venkatesh. The effect of global transcriptional regulators on the anaerobic fermentative metabolism of Escherichia coli. *Molecular BioSystems*, 13(7):1388–1398, 2017.

[43] H. K. Seoh and P. C. Tai. Catabolic repression of secB expression is positively controlled by cyclic AMP (cAMP) receptor protein-cAMP complexes at the transcriptional level. *Journal of Bacteriology*, 181(6):1892–1899, March 1999.

[44] Timothy J Larsons, Shanzhang Ye, Deborah L Weissenborn, and Heidi J Hoffmann. Purification and Characterization of the Repressor for the sn-Glycerol 3-Phosphate Regulon of *Escherichia coli* K12. *Journal of Biological Chemistry*, 262(33):15869–15874, 1987.

[45] E. C. C. Lin. Glycerol dissimilation and its regulation in bacteria. *Annual Review of Microbiology*, 30(1):535–578, 1976.

[46] Kevin S. Myers, Huihuang Yan, Irene M. Ong, Dongjun Chung, Kun Liang, Frances Tran, Sündüz Keleş, Robert Landick, and Patricia J. Kiley. Genome-scale analysis of *Escherichia coli* fnr reveals complex features of transcription factor binding. *PLOS Genetics*, 9(6):1–24, 06 2013.

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

[48] Tyrrell Conway, James P. Creecy, Scott M. Maddox, Joe E. Grissom, Trevor L. Conkle, Tyler M. Shadid, Jun Teramoto, Phillip San Miguel, Tomohiro Shimada, Akira Ishihama, Hirotada Mori, and Barry L. Wanner. Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *mBio*, 5(4):e01442–14, July 2014.

[49] Ins Compan and Danlle Touati. Anaerobic activation of arcA transcription in Escherichia coli: roles of Fnr and ArcA. *Molecular Microbiology*, 11(5):955–964, 1994.

[50] Douglas F Browning and Stephen J W Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, pages 638–650, 2016.

[51] Janet E. Lindsley and Jared Rutter. Whence cometh the allosterome? *Proceedings of the National Academy of Sciences of the United States of America*, 103(28):10533–10535, 2006.

[52] Nicolas E Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9):5136–5141, April 2003.

[53] A. Tareen and J. B. Kinney. Biophysical models of cis-regulation as interpretable neural networks. *bioRxiv*, 2019.

[54] F. R. Blattner. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, September 1997.

[55] Nishant Pappireddi, Lance Martin, and Martin Wühr. A Review on Quantitative Multiplexed Proteomics. *ChemBioChem*, 20(10):1210–1224, 2019.

[56] Jrgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008.

[57] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995.

[58] Natsuko Yamamoto, Kenji Nakahigashi, Tomoko Nakamichi, Mihoko Yoshino, Yuki Takai, Yae Touda, Akemi Furubayashi, Satoko Kinjyo, Hitomi Dose, Miki Hasegawa, Kirill A Datsenko, Toru Nakayashiki, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Update on the keio collection of escherichia coli single-gene deletion mutants. *Molecular Systems Biology*, 5:335–335, 2009.

[59] William T. Ireland and Justin B. Kinney. MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays. preprint, Bioinformatics, May 2016.