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*]

***For correspondence:** phillips@pboc.caltech.edu (RP)

Present address: †Howard Hughes Medical Institute and Department of Biology, University of Washington, Seattle, WA 98195

[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

Abstract

While advances in DNA sequencing have enabled widespread access to this groundbreaking technology, even in the case of arguably one of biology's best understood organisms, the bacterium *Escherichia coli*, for ≈ 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 for unraveling the regulatory genome of a broad array of genes in *E. coli*. Specifically, we show that we can extract previously known regulatory information and unmask the regulatory architectures for more than 80 uncharacterized promoters, and in many of those cases, which transcription factors mediate this 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.

Introduction

DNA sequencing is as important to biology as the telescope was (and is) to astronomy. We are now living in the age of genomics, where the ability to sequence DNA is cheap and easy. 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 tell us very little about how organisms use those parts creatively and constructively in space and time. But we know that promoters apply all-important dynamic logical operations that control when and where genetic information is accessed. In this paper, we show 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 base pair in the promoter to downstream gene expression, quite literally determining how much information each promoter base pair carries about the level of gene expression.

The advent of RNA-Seq (***Mortazavi et al., 2008***) 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. (***Stuart and Satija, 2019***). In contrast to whole genome screening sequencing approaches, such as Tn-Seq (***Goodall et al., 2018***) and ChIP-Seq (***Gao et al., 2018***) 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 (***Sharon et al., 2012***; ***Patwardhan et al., 2012***; ***Fulco et al., 2019***). One elegant study relevant to the bacterial case of interest here by ***Kosuri et al.*** (***2013***) screened more than 10^4 combinations of promoter and ribosome binding sites (RBS). Even more recently, they have utilized this system to search for regulation across the genome (***Urtecho et al., 2019***, ***2020***), in a way we as being complementary to our own. While their approach allows for a course-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 recent X-Seq tools based on massively-parallel reporter assays with broad biophysical reach is the Sort-Seq approach developed by ***Kinney et al.*** (***2010***). The great utility of this technique is that it not only reveals where transcription factors are binding (which many other techniques are capable of as well), but also that this sequencing-as-read-out approach provides a base pair resolution mapping of how sequence controls the level of gene expression. 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 as described by energy matrices (***Kinney et al., 2010***) that allow us to unleash 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 ***Razo-Mejia et al.*** (***2018***) and ***Chure et al.*** (***2019***). 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 provides an interpretable scheme to design genetic circuits with a desired expression output (***Barnes et al., 2019***).

Earlier work from ***Belliveau et al.*** (***2018***) illustrated how Sort-Seq, in conjunction with mass spectrometry to figure out what transcription factors are binding the putative binding sites, could be used to dissect a handful of these promoters for which we were previously mired in regulatory ignorance. However, a crucial drawback of the approach of ***Belliveau et al.*** (***2018***) 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 (***Santos-Zavaleta et al., 2019***; ***Belliveau et al., 2018***), a higher-throughput approach is needed. 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. How is the problem solved? 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 this approach as Reg-Seq, which we unleash here on over one hundred promoters. The Reg-Seq method has some key ideas that are critical for making the link between DNA sequence

and regulatory outcomes. The concept of the method is to perturb promoter regions by mutating them and then using sequencing to read out both perturbation and gene expression. We generate a broad diversity of promoter sequences for each promoter of interest experimentally and use mutual information as a metric to measure information flow from that distribution of sequences to gene expression. By doing this, 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 the ability to scale up the power of the Sort-Seq method, we report here the results of applying our Reg-Seq approach to 113 *E. coli* genes, whose promoter architectures we were able to elucidate 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 basis, allowing us to uncover a large swath of the *E. coli* regulome, in detailed 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 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 of the method to further elucidating the *E. coli* genome and opening up the quantitative dissection of other non-model organisms. Lastly, in the Methods section and flushed out further in the Supplemental 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 methods.

Results

Selection of genes and methodology

As shown in Figure 1, we have considered more than 100 genes from across the entire *E. coli* genome. Our choices were based on a number of different factors including the desire to have a subset of genes that serve as “gold standard” genes for which the hard work of generations of molecular biologists have yielded deep insights into their regulation. By exploiting our method on these genes, we were able to demonstrate that our Reg-Seq approach recovers not only what was already known about binding sites and transcription factors for well-characterized promoters, but also to reveal whether there are any important differences between the results of the methods presented here and the previous generation 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 using mass spectrometry that measured the copy number of different proteins over a wide range (22 distinct) of growth conditions (***Schmidt et al., 2015***). 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 was based on the use of fluorescence activated cell sorting as a way to uncover putative binding sites for previously uncharacterized promoters (***Belliveau et al., 2018***). However, our previous method required a gene-by-gene approach to regulatory discovery. As a result, as shown in Figure 1 we have formulated a second generation version that permits a high-throughput interrogation of the genome. Figure 1 provides an overview of the method used in the work presented here. For each promoter we interrogate, 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 can 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 ($p < 0.01$). Further details are found in the Methods section. Blue regions of the histogram 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. Given the putative binding sites, we can construct oligonucleotides that serve as fishing hooks to fish out the transcription factors that bind to those putative binding sites. 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 1, the experimental pipeline yields a complete picture of a simple repression architecture.

Visual tools for data presentation

Throughout our investigation of the more than 100 genes explored in this study, we repeatedly rely 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 DNA promoter sequence relates to gene expression level in the form of the information footprint shown in the middle of Figure 1. The result is nicely summarized in our information footprints, where each bar corresponds to an individual base pair and the height of the bar represents how large of an effect mutations at this location have on the gene expression level. Specifically, the value is the mutual information at base pair b between mutation of a base pair at the site and expression level and is given by

$$I_b = \sum_{m,exp} p(m, exp) log_2(\frac{p(m, exp)}{p(m)p(exp)}) \quad (1)$$

where m represents whether or not the target base is mutated, and exp refers to reads from the DNA library ($exp = 0$) or the mRNA transcripts ($exp = 1$). $p(m)$ in equation 1 refers to the probability that a given sequencing read will be from a mutated base. $p(exp)$ 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 on average lowered gene expression (in blue, indicating an activating role) or increased gene expression (in red, indicating a repressing role). Within these footprints, we look for regions of contiguous base pairs which

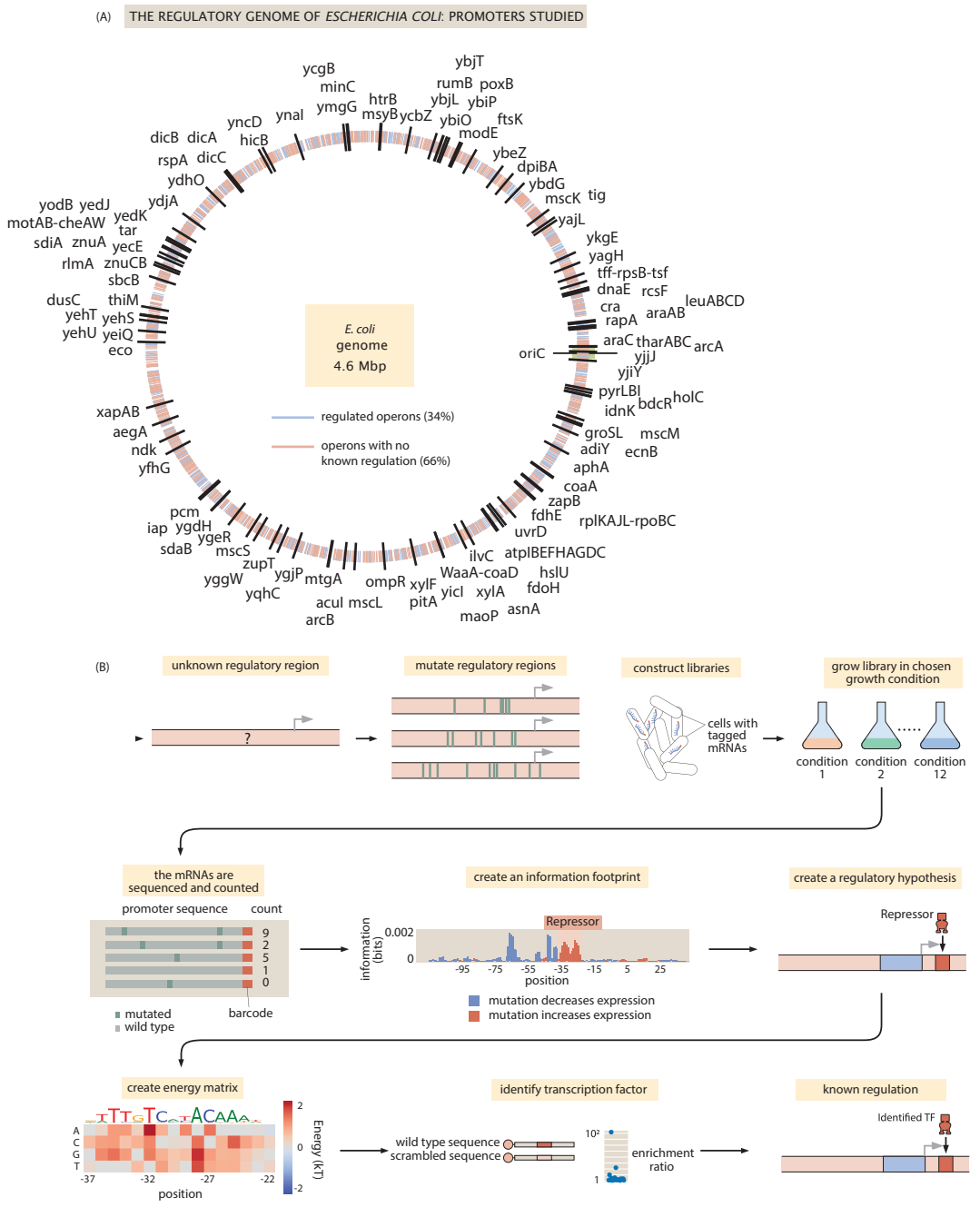


Figure 1. Dissection of the regulatory genome of *Escherichia coli*. (A) Illustration of current ignorance with respect to how genes are regulated in *E. coli*, with genes with previously annotated regulation (as reported on RegulonDB) denoted with blue ticks and genes with no previously annotated regulation denoted with red ticks. The 113 genes explored in this study are labeled in black, with (B) showing the Reg-Seq procedure by which we 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.

impact gene expression similarly (either increasing or decreasing), as these regions implicate the influence of a transcription factor binding site. As can be seen throughout the paper (see Figure 3 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). 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

$$\text{binding energy} = \sum_{i=1}^{N} \varepsilon_i, \tag{2}$$

where the energy matrix is predicated on a linear binding model in which each base within the binding site region contributes a specific value (ε_i for the i^{th} 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 RNAP site, are assigned kT energy units following the procedure in ***Kinney et al.*** (***2010***) and validated on the *lac* operon in ***Barnes et al.*** (***2019***).

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 (***Schneider and Stephens, 1990***). 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 by human eyes 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 described above. While the details of the DNA chromatin 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 1(B). In these plots, the relative abundance of each protein binding 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 ($p < 10^{-4}$).

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. 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. 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.

Specific results

Figure 2 provides a summary of the discoveries made in the work done here using our next generation Reg-Seq approach. Figure 2(A) provides a shorthand notation that conveniently characterizes the different kinds of regulatory motifs found in bacteria. In previous work, 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 hypothesize that this is not a statement about the facts of the *E. coli* genome, but rather a reflection of our collective regulatory ignorance. 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 and to now see retrospectively the far-reaching importance of that architecture across the *E. coli* genome (***Jacob and Monod, 1961***).

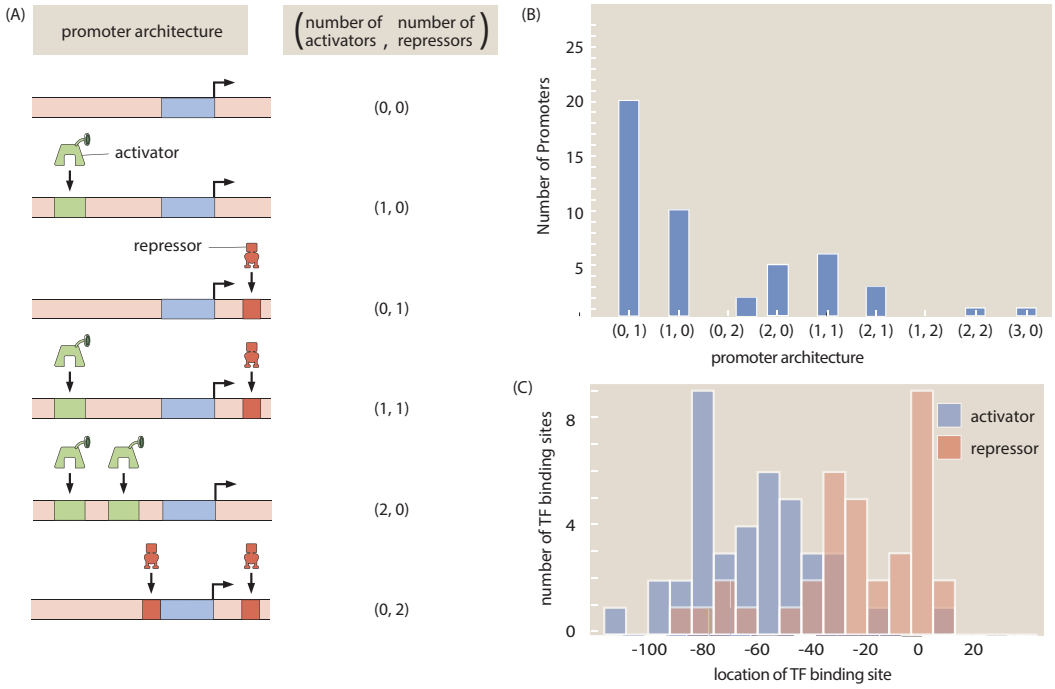


Figure 2. A summary of regulatory architectures discovered in this study. (A) The cartoons display a representative example of each type of architecture, along with the corresponding shorthand notation. (B) Counts of the different regulatory architectures discovered in this study as compared to known regulation found on RegulonDB. (C) Distribution of positions of binding sites discovered in this study for activators and repressors, respectively.

Figure 2(B) summarizes, for the 113 genes we considered, how many of them are simple activation motifs, how many are simple repression motifs and so on. 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 2(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 binding-site position "knob".

Figure 3 delves more deeply into the various regulatory architectures described in Figure 2 (B) by showing some of the promoters for which these architectures were discovered. In each of the cases shown in the figure, prior to the work presented here, these promoters had no regulatory information in knowledge bases like, Ecocyc (***Keseler et al., 2016***) and RegulonDB (***Santos-Zavaleta et al., 2019***). Now, using the sequencing methods explained above in conjunction with mass spectrometry we were able to identify candidate binding sites that were then used to construct oligonucleotides that were used to construct biotinylated DNA "bait" probes to pull down their corresponding putative transcription factors. Given this regulatory information, we are now poised to carry out a quantitative dissection of these promoters using theoretical tools from statistical physics to predict their input-output function and precision biophysical measurements to test those predictions.

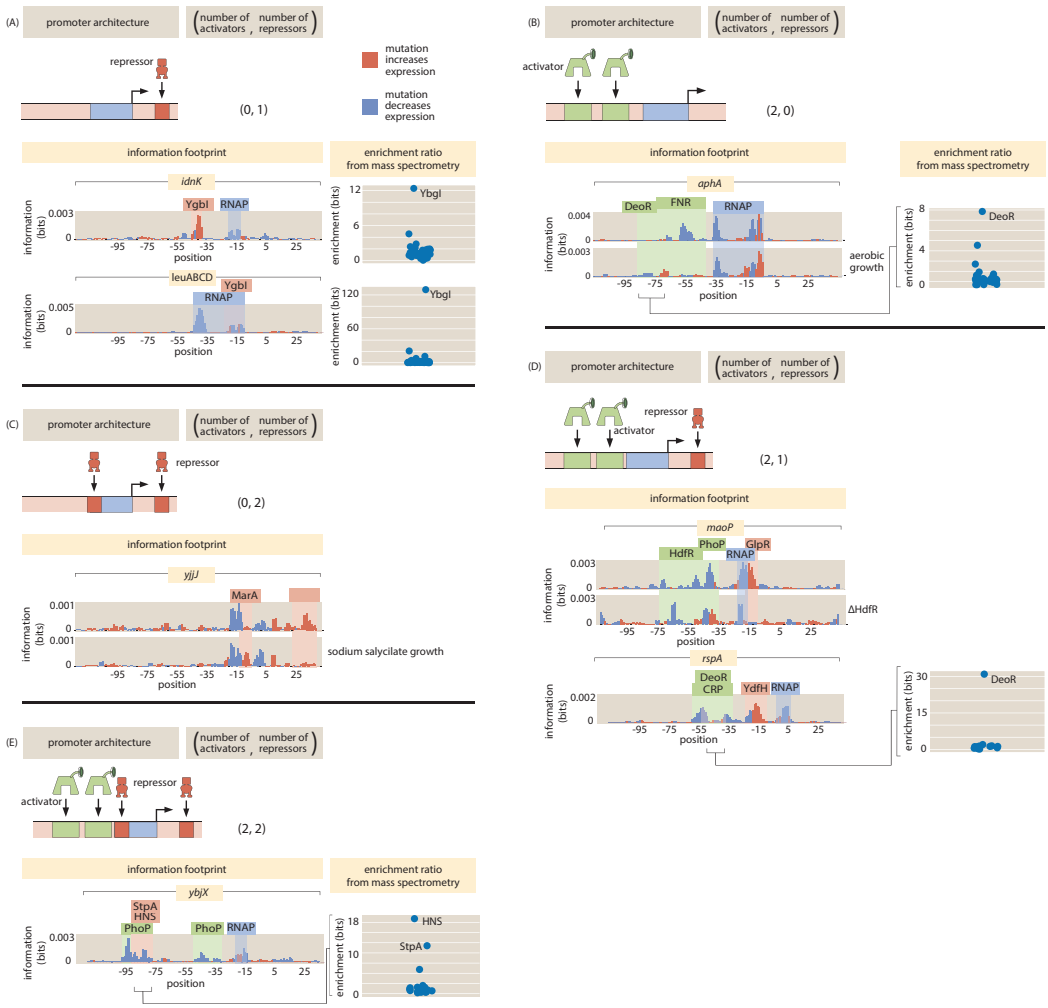


Figure 3. Newly discovered or updated regulatory architectures. Examples of information footprints, gene knockouts, and mass spectrometry data used to identify transcription factors for five genes. (A) Examples of simple repression architectures where the locations of the putative binding sites are highlighted in red and the identities of the bound transcription factors are revealed in the mass spectrometry data. (B) An example of a 2 activator architecture. (C) An example of a 2 repressor architecture. (D) An example of a 2 activator and 1 repressor architecture. (E) An example of a 2 activator and 2 repressor architecture.

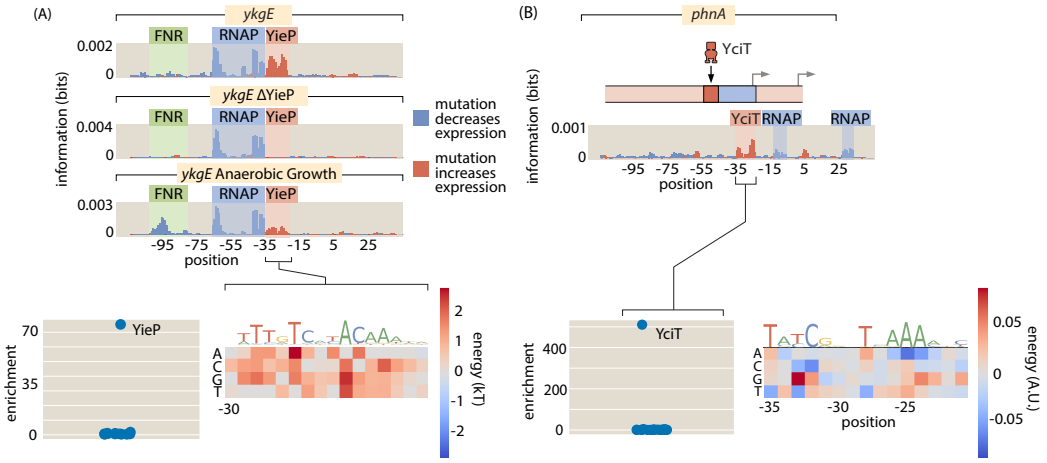


Figure 4. Examples of the power of Reg-Seq in the context of promoters with no previously known regulatory information. (A) From the information footprint of the *ykgE* promoter under different growth conditions, we can identify a repressor binding site downstream of the RNAP binding site. From the enrichment 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) illustrates a comparable dissection for the *phnA* promoter.

A recent paper identified what they christened the y-ome in *E. coli* (***Ghatak et al., 2019***), which they defined as the set of genes that don't have the annotation necessary to have an effect in a systems biology model. Their surprising result is the finding that roughly 35% of the genes in the *E. coli* genome are functionally unannotated. The situation is likely worse for other organisms, demonstrating the need for methods such as Reg-Seq to come to terms with these parts of genomes. For many of the genes in the y-ome, we remain similarly ignorant of how they are regulated. Figures 3 and 4 provide several examples from the y-ome of genes and transcription factors for which little to nothing was previously known. As shown in Figure 4, 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 to see many examples of the regulation of y-ome genes.

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 4(A)-(B), in both cases, 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 (***Kinney et al., 2010***; ***Bintu et al., 2005***; ***Barnes et al., 2019***; ***Garcia and Phillips, 2011***).

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. 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 regulators with the potential to

change the entire landscape of *E. coli* gene regulation, 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 5(A) more than double the number of operons known to be regulated by GlpR as reported in RegulonDB (***Santos-Zavaleta et al., 2019***). We found 5 newly regulated operons in our data set, even though we were not specifically targeting GlpR regulation. Although the numbers are too small to make good estimates, this does support the notion that GlpR widely regulates and many of its sites would be found in a full search of the genome. The regulatory roles revealed in Figure 5(A) also reinforce the evidence that GlpR is a repressor.

GlpR is known to bind the inducer glycerol-3-phosphate (G3P) (***Larsons et al., 1987***). When G3P binds to GlpR, similarly to the way that the classic LacI repressor is induced by allolactose, the probability of finding GlpR bound to DNA is reduced. In the three growth conditions in our study containing glucose, of which a representative set of examples are shown in Figure 5(A), the repressor binds strongly while in all other growth conditions the binding is greatly diminished but not entirely abolished. 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. None of the genes found to be regulated by GlpR are also under CRP regulation, and we cannot be sure what the exact cause of the indirect interaction is. However, *gpsA* is activated by CRP (***Seoh and Tai, 1999***). Its product synthesizes the inducer of GlpR (G3P) and so intracellular G3P levels will be lowered by the presence of glucose, which would repress those genes under GlpR control.

Prior to this study, there were 4 operons known to be regulated by GlpR, each with between 4 and 8 GlpR binding sites (***Gama-Castro et al., 2016***). In these operons, the absence of glucose and the partial induction of GlpR was not enough to prompt a notable change in gene expression (***Lin, 1976***). In these operons, GlpR acts 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 a glucose sensor, as only the absence of glucose is needed to relieve repression by GlpR.

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 (***Gama-Castro et al., 2016***). However, the newly discovered FNR sites displayed in Figure 5(B) demonstrate that for even 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 (***Myers et al., 2013***). 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 5(B).

To end the summary of our results, we would like to further emphasize the value we gained from having taken a hundred-gene approach rather than the gene-by-gene approach taken before. We observe quantitatively how FNR affects the expression of *fdhE* both directly through transcription factor binding (Figure 6(A)) and indirectly though increased expression of ArcA (Figure 6(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 (***Conway et al., 2014***). Regulation within an operon is even more poorly studied than regulation in general. The main promoter for *fdoGHI-fdhE* is repressed by a transcription factors. We see in Figure 6(B) that, as *fdhE* is regulated by ArcA and FNR, it

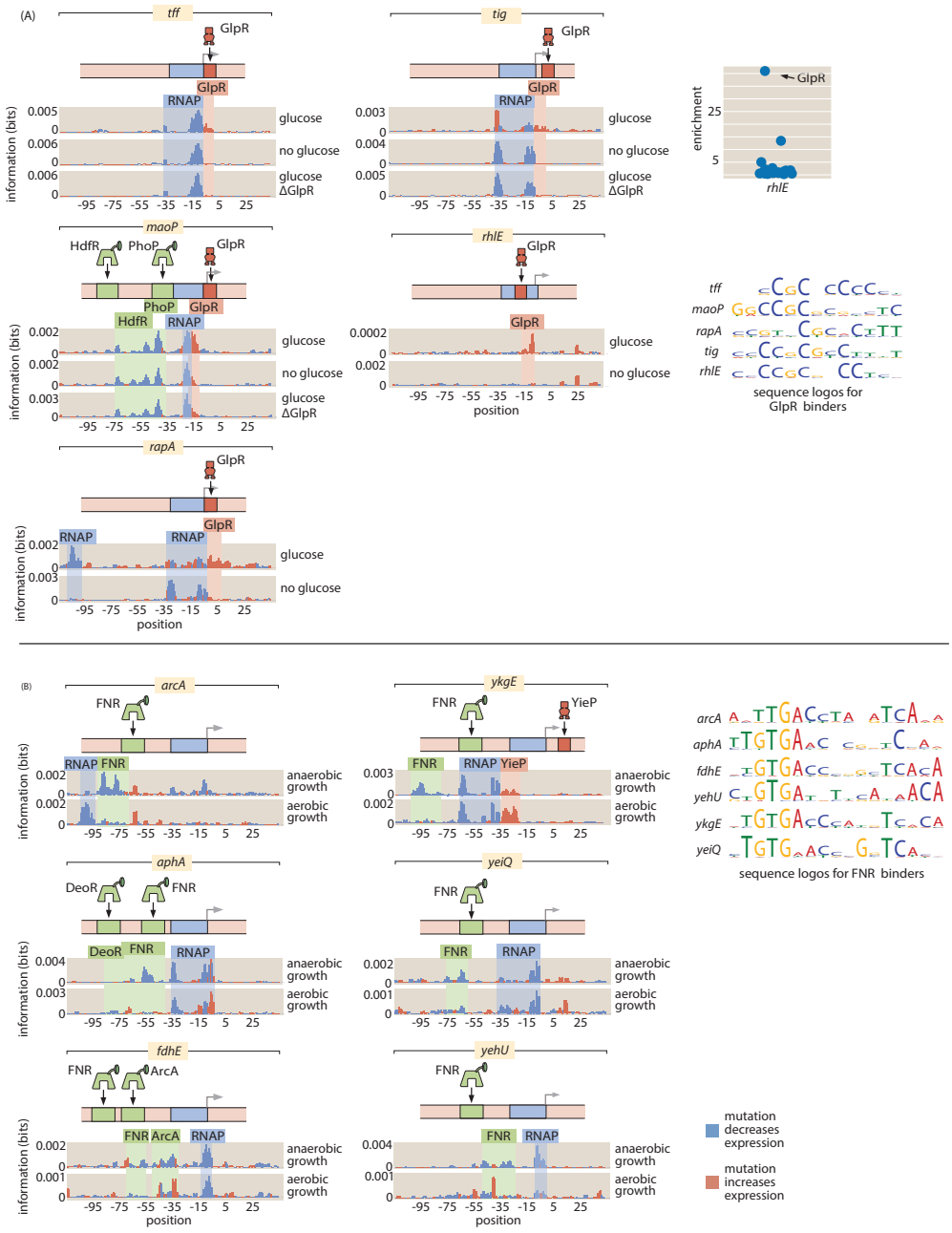


Figure 5. Reg-Seq analysis of broadly-acting transcription factors. By using Reg-Seq to dissect over 100 promoters at once, we can begin to understand not only how individual promoters are regulated, but also how individual transcription factors serve to regulate a suite of genes. (A) GlpR as a widely-acting 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.

will be upregulated in anaerobic conditions (***Compan and Touati, 1994***). The main TSS transcribes all four genes in the operon, while the secondary site shown in Figure 6(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 the typical operon assumption breaks down.

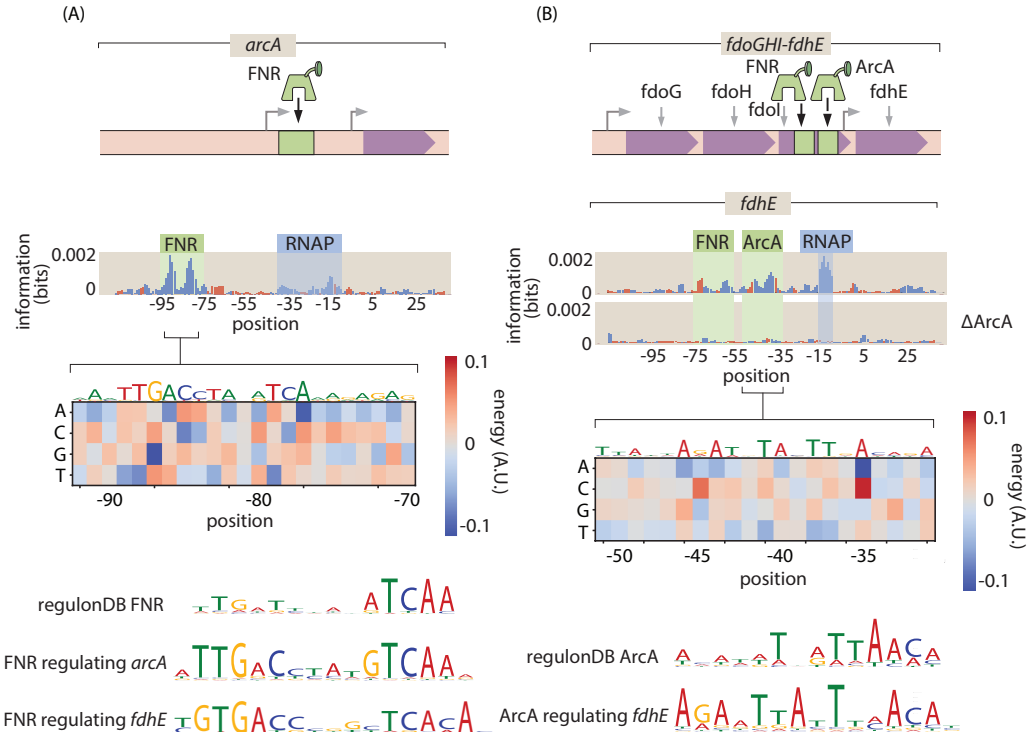


Figure 6. Inspection of an anaerobic respiration genetic circuit. By exploring over a 100 promoters at once, we can begin to piece together some genetic circuitry of the cell. A) Here we see not only how the *arcA* promoter is regulated, but also the role this transcription factor plays in the regulation of another promoter. B) Intra-operon regulation of *fdhE* by both FNR and ArcA. A TOMTOM 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.

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 7, 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

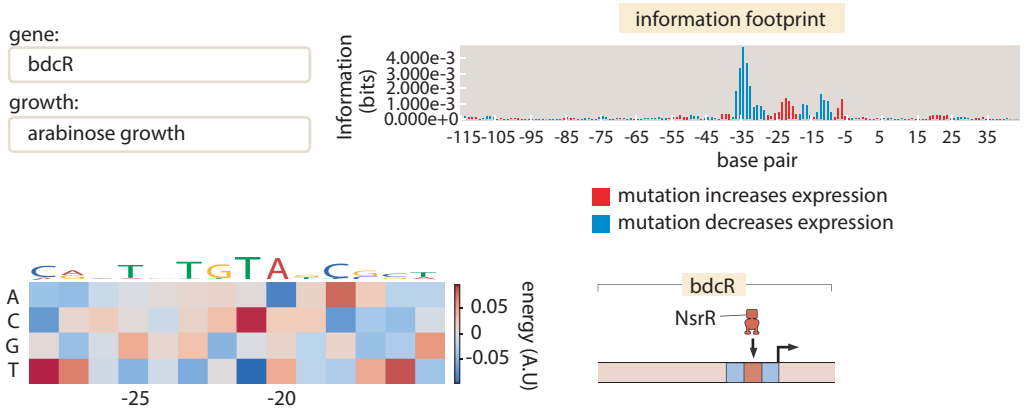


Figure 7. Representative view of the interactive figure that is available online. This interactive figure captures the entirety of our dataset. Each figure features a drop-down menu of genes and growth conditions. For each such gene and growth condition, there is a corresponding information footprint 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.

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.

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 even of our best-understood model organisms remains so vast. Despite understanding the regulation of transcription initiation in bacterial promoters (***Browning and Busby, 2016***), and how to tune their expression, we lack an experimental framework to unravel understudied promoter architectures at scale. Until the work presented here, even in the case of one of biology's best understood organisms, the humble bacterium *E. coli*, over 65% of its genes have no known regulatory architecture (***Belliveau et al., 2018***). 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 (***Lindsley and Rutter, 2006***). The work presented here provides a general methodology for attacking the former problem and even shows how by performing Reg-Seq in different growth conditions, we can make headway on the latter problem.

The advent of cheap DNA sequencing offers the promise of beginning to achieve that grand challenge goal in the form of massively-parallel reporter assays reviewed in ***Kinney and McCandlish (2019)***. A particular implementation of such methods was christened Sort-Seq (***Kinney et al., 2010***) and was demonstrated in the context of well understood regulatory architectures. A next generation of the Sort-Seq method (***Belliveau et al., 2018***) established experiments through the use of affinity chromatography and mass spectrometry which made it possible to identify the transcription factors that bind the putative binding sites discovered in the Sort-Seq analysis. 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. The case studies described in the main text of the paper provide evidence of the ability of the method to deliver on the promise of beginning to systematically uncover the regulatory genome and 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.

Several shortcomings remain in the approach introduced here. First, we are well aware of and excited about 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 (***Melnikov et al., 2012***). The current incarnation of the methods described here has 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 sequencing technology to use Reg-Seq to identify putative DNA binding sites across the genome already exists, and this technique could be applied in parallel to recover putative binding sites across the genome without losing resolution. However, 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 next-generation experiments that will make it possible to accomplish transcription factor identification at higher throughput. We are exploring multiplexed mass spectrometry measurements and multiplexed Reg-Seq on libraries of gene knockouts as ways to break the identification bottleneck.

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 in short order 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.

Methods

Library construction

Promoter variants were synthesized on a microarray (TWIST Bioscience, San Franscisco, CA). The promoters were designed computationally to have a 10 % rate and to cover 160 bp. The mutations were generated randomly, and the library was regenerated if the mutation rate did not fall within 0.5% of the target mutation rate. There are an average of 2200 unique promoter sequences per gene. 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. After the barcode there is a RBS and a GFP coding region. Mutated promoters were PCR amplified and inserted by Gibson assembly into the

plasmid backbone of pJK14 (SC101 origin) (***Kinney et al., 2010***). Constructs were electroporated into *E. coli* K-12 MG1655 (***Blattner, 1997***).

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 preformed 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.

Analysis of sequencing results

To determine putative transcription factor binding sites, we first computed the effect of mutations on gene expression at a base pair-by-base pair level using information footprints. To initially capture all regions important for gene expression, regardless of whether or not they are transcription factors, we identify regions where gene expression is changed significantly up or down by mutation ($p < 0.01$). We find a region of effect by averaging over 15 base pairs, as this is a reasonable size for a transcription factor binding site. We include binding sites that are significant in any of the tested growth conditions.

Many false positive 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 (for example, does the site resemble the consensus TGNTATAAT extended minus 10 for sigma 70 sites), and the TOMTOM tool to computationally compare the potential site to examples of sigma 70, sigma 38, and sigma 54 sites that we determined in this experiment. We discard any sites that have a p-value of similarity with an RNAP site of less than $5x10^{-3}$. 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 leaves 83 identified transcription factor binding regions. We adjust the edges of binding sites to only be wide enough to contain all base-pairs with a significant effect on gene expression ($p < 0.01$).

For primary RNAP sites, we include a list of probable sigma factor identities as a Supplementary Table. Sites are judged by similarity to consensus binding sites, with the most similar sigma factor site reported based on TOMTOM comparison (***Gupta et al., 2007***). Those sites where there is significant uncertainty due to overlapping sites are omitted. Overlapping sites in general can pose issues for this method. In many cases, looking at growth conditions where only one of the overlapping sites is active is effective to establish site boundaries and 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.

DNA affinity chromatography and mass spectrometry

Upon identifying a putative transcription factor binding site, we used DNA affinity chromatography, as done in (***Belliveau et al., 2018***) to isolate and enrich for the transcription factor of interest. In brief, we order biotinylation 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 then tether these oligos to magnetic streptavidin beads

(Dynabeads MyOne T1; ThermoFisher, Waltham, MA), which we then incubate 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 were 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 spec by in-gel digestion using the Lys-C protease (Wako Chemicals, Osaka, Japan) or in-solution digestion. Liquid chromatography coupled mass spectrometry (LC-MS) was performed as previously described by (***Belliveau et al., 2018***), and are further discussed in the SI appendix. 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 doesn't 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 (***Pappireddi et al., 2019***) 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 (20) of all identified transcription factor targets were analyzed by mass spec 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 ($p < 10^{-4}$).

Construction of knockout strains

Conducting DNA affinity chromatography followed by mass spectrometry on our putative binding sites resulted in likely candidates for the transcription factors that are responsible for the information contained at a given promoter region. To verify that a given transcription factor is, in fact, acting to regulate 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 (***Yamamoto et al., 2009***) from the Coli Genetic Stock Center. These knockouts were then 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 these constructed knockout strains, allowing us to perform the RNA sequencing in the same context as the original experiments.

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) All raw sequencing data is available at the Sequence Read Archive (accession no.SUB6723530). All inferred information footprints and energy matrices can be found on the CalTech data repository doi:10.22002/D1.1331.

Acknowledgments

Experimentally, 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. We also thank Stephanie Barnes, Griffin Chure, Hernan Garcia, Soichi Hirokawa, Heun Jin Lee, Nicholas McCarty, Muir Morrison, Manuel Razo-Mejia and Gabe Salmon for useful discussion. 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 Guillaume Urtecho, Hernan Garcia, Steve Quake, Matt Thompson, Curt Callan

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 training grant,.

References

Barnes SL, Belliveau NM, Ireland WT, Kinney JB, Phillips R. Mapping DNA sequence to transcription factor binding energy *in vivo*. PLoS Computational Biology. 2019; 15(2):1–29. doi: 10.1371/journal.pcbi.1006226.

Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, Hess S, Kinney JB, Phillips R. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2018; 115(21):E4796–E4805. doi: 10.1073/pnas.1722055115.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. Transcriptional regulation by the numbers: models. Current Opinion in Genetics & Development. 2005 Apr; 15(2):116–124. doi: 10.1016/j.gde.2005.02.007.

Blattner FR. The Complete Genome Sequence of Escherichia coli K-12. Science. 1997 Sep; 277(5331):1453–1462. doi: 10.1126/science.277.5331.1453.

Browning DF, Busby SJW. Local and global regulation of transcription initiation in bacteria. Nature Reviews Microbiology. 2016; p. 638–650. doi: 10.1038/nrmicro.2016.103.

Chure G, Razo-Mejia M, Belliveau NM, Einav T, Kaczmarek ZA, Barnes SL, Lewis M, Phillips R. Predictive shifts in free energy couple mutations to their phenotypic consequences. Proceedings of the National Academy of Sciences of the United States of America. 2019; 116(37):18275–18284. doi: 10.1073/pnas.1907869116.

Compan I, Touati D. Anaerobic activation of arcA transcription in *Escherichia coli* : roles of Fnr and ArcA. Molecular Microbiology. 1994 Mar; 11(5):955–964.

Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, Mori H, Wanner BL. Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. mBio. 2014 Jul; 5(4):e01442–14. doi: 10.1128/mBio.01442-14.

Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology. 2008 Dec; 26(12):1367–1372. doi: 10.1038/nbt.1511.

Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen JV, Mann M. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nature Protocols. 2009 May; 4(5):698–705. doi: 10.1038/nprot.2009.36.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, Nguyen TH, Kane M, Perez EM, Durand NC, Lareau CA, Stamenova EK, Aiden EL, Lander ES, Engreitz JM. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019; 51(12):1664–1669.

Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Del Moral-Chavez V, Rinaldi F, et al. RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Research. 2016; 44(D1):D133–D143. doi: 10.1093/nar/gkv1156.

Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Chen K, Sastry AV, Fang X, Mih N, Yang L, Eichner J, Cho Bk, Kim D, Palsson BO. Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. Nucleic Acids Research. 2018; 46(20):10682–10696. doi: 10.1093/nar/gky752.

Garcia HG, Phillips R. Quantitative dissection of the simple repression input-output function. Proceedings of the National Academy of Sciences. 2011 Jul; 108(29):12173–12178. doi: 10.1073/pnas.1015616108.

Ghatak S, King ZA, Sastry A, Palsson BO. The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. Nucleic Acids Research. 2019; 47(5):2446–2454. doi: 10.1093/nar/gkz030.

Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. The Essential Genome of Escherichia coli K-12. mBio. 2018; 9(1).

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biology. 2007; 8(2). doi: 10.1186/gb-2007-8-2-r24.

Jacob F, Monod J. On the Regulation of Gene Activity. . 1961; p. 19.

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, et al. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. Nucleic Acids Research. 2016; 45(D1):D543–D550. https://doi.org/10.1093/nar/gkw1003, doi: 10.1093/nar/gkw1003.

Kinney JB, McCandlish DM. Massively Parallel Assays and Quantitative Sequence–Function Relationships. Annual Review of Genomics and Human Genetics. 2019; 20(1):99–127. doi: 10.1146/annurev-genom-083118-014845.

Kinney JB, Murugan A, Callan CG, Cox EC. 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. 2010; 107(20):9158–9163. doi: 10.1073/pnas.1004290107.

Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(34). doi: 10.1073/pnas.1301301110.

Larsons TJ, Ye S, Weissenborn DL, Hoffmann HJ. Purification and Characterization of the Repressor for the sn-Glycerol 3-Phosphate Regulon of *Escherichia coli* K12. Journal of Biological Chemistry. 1987; 262(33):15869–15874.

Lin ECC. Glycerol Dissimilation and its Regulation in Bacteria. . 1976; p. 44.

Lindsley JE, Rutter J. Whence cometh the allosterome? Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(28):10533–10535. doi: 10.1073/pnas.0604452103.

Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011 Nov; 27(21):2957–2963. doi: 10.1093/bioinformatics/btr507.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Jr CGC, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature Biotechnology. 2012; 30(3):271–277. doi: 10.1038/nbt.2137.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5(7):621–628. doi: 10.1038/nmeth.1226.

Myers KS, Yan H, Ong IM, Chung D, Liang K, Tran F, Keleş S, Landick R, Kiley PJ. Genome-scale Analysis of *Escherichia coli* FNR Reveals Complex Features of Transcription Factor Binding. PLOS Genetics. 2013 06; 9(6):1–24. https://doi.org/10.1371/journal.pgen.1003565, doi: 10.1371/journal.pgen.1003565.

Pappireddi N, Martin L, Wühr M. A Review on Quantitative Multiplexed Proteomics. ChemBioChem. 2019; 20(10):1210–1224. doi: 10.1002/cbic.201800650.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, Ahituv N, Pennacchio LA, Shendure J. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30(3):265–70.

Razo-Mejia M, Barnes SL, Belliveau NM, Chure G, Einav T, Lewis M, Phillips R. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. Cell Systems. 2018; 6(4):456–469.e10. doi: 10.1016/j.cels.2018.02.004.

Santos-Zavaleta A, Salgado H, Gama-castro S, Laura G, Ledezma-tejeida D, Mishael S, Garc S, Alquicira-hern K, Jos L, Pe P, Ishida-guti C, Vel DA, Moral-ch D, Galagan J, Collado-vides J. RegulonDB v 10 . 5 : tackling challenges to unify classic and high throughput knowledge of gene regulation in E . coli K-12. Nucleic Acids Research. 2019; 47(November 2018):212–220. doi: 10.1093/nar/gky1077.

Schmidt A, Kochanowski K, Vedelaar S, Ahrne E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M. The quantitative and condition-dependent *Escherichia coli* proteome. Nature biotechnology. 2015; 34(1):104–110. doi: 10.1038/nbt.3418.

Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research. 1990; 18(20):6097–6100. doi: 10.1093/nar/18.20.6097.

Seoh HK, Tai PC. Catabolic repression of secB expression is positively controlled by cyclic AMP (cAMP) receptor protein-cAMP complexes at the transcriptional level. Journal of Bacteriology. 1999 Mar; 181(6):1892–1899.

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

Stuart T, Satija R. Integrative single-cell analysis. Nature Reviews Genetics. 2019; 20(May):257–272. doi: 10.1038/s41576-019-0093-7.

Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S. Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. Biochemistry. 2019; 58(11):1539–1551.

Urtecho G, Insigne K, Tripp AD, Brinck M, Lubock NB, Kim H, Chan T, Kosuri S. Genome-wide Functional Characterization of *Escherichia col* Promoters and Regulatory Elements Responsible for their Function. bioRxiv. 2020; https://www.biorxiv.org/content/early/2020/01/06/2020.01.04.894907, doi: 10.1101/2020.01.04.894907.

Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, Datsenko KA, Nakayashiki T, Tomita M, Wanner BL, Mori H. Update on the Keio collection of Escherichia coli single-gene deletion mutants. Molecular systems biology. 2009; 5:335–335. doi: 10.1038/msb.2009.92.

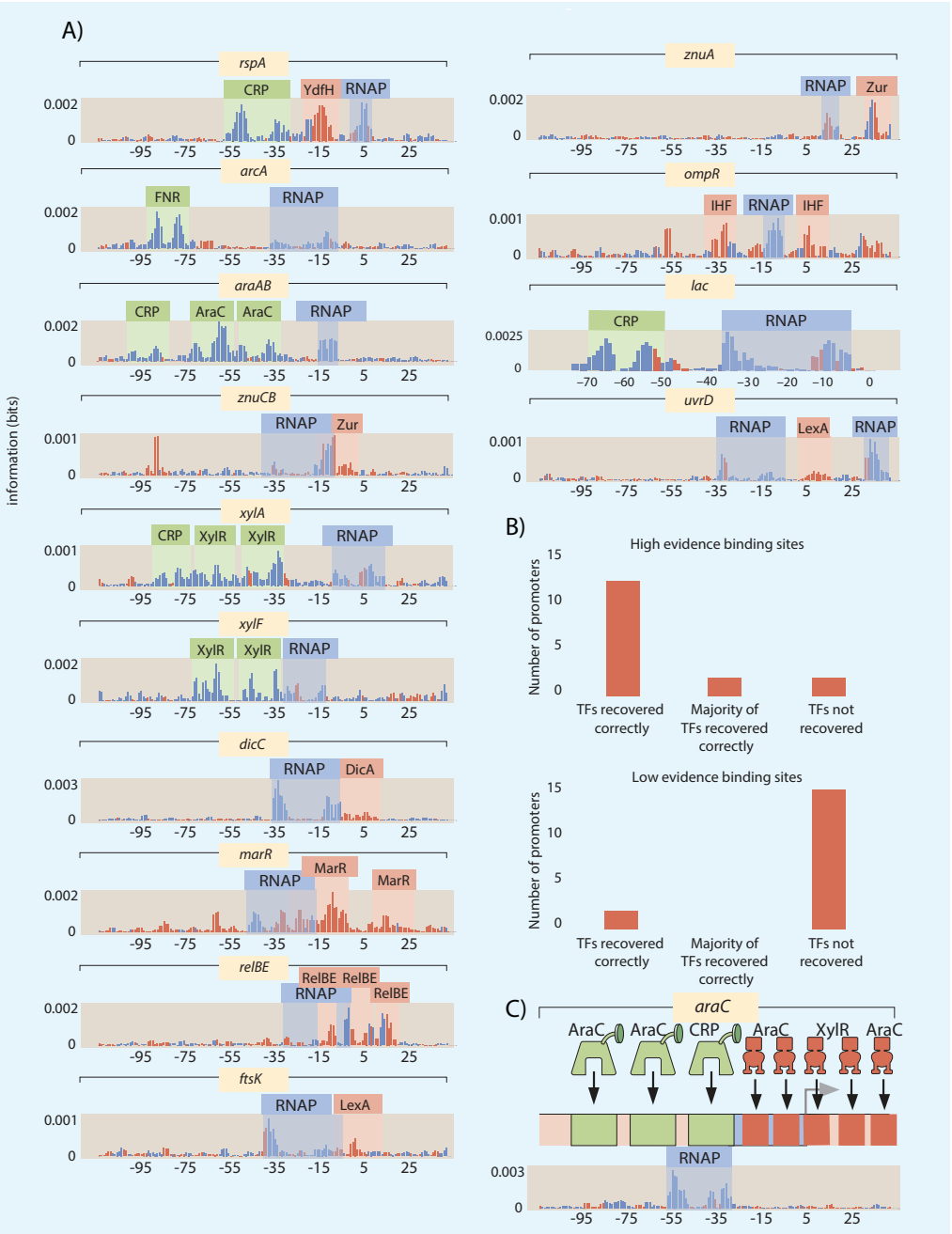
Appendix 1

Comparison of results for known binding sites

In the sections that follow, we present a number of extra details about both our methodology and our results whose aim is to aid the reader in getting a more rigorous view of the work we did to ensure that our methods were coherent and to validate our biological insights.

The work presented here is effectively a third-generation of the use of Sort-Seq methods for the discovery of regulatory architecture. The primary difference between the present work and previous generations is the use of RNA-Seq rather than fluorescence and cell sorting as a readout of the level of expression of our promoter libraries. As such, there are many important questions to be asked about the comparison between the earlier methods and this work. We attack that question in several ways. First, as shown in Appendix 1 Figure 2, we have performed a head-to-head comparison of the two approaches to be described further in this section. Second, as shown in the next section, our list of candidate promoters included roughly 20% for which the last 50 years of molecular biology has resulted in a knowledge (sometimes only partial) of their regulatory architecture. In these cases, we examined the extent to which our methods recover the known features of regulatory control about those promoters.

We have tested over 20 genes for which there is already some substantial regulatory knowledge reported in the literature. The successes and failures of this test are detailed in Appendix 1 Figure 1. For those promoters which have strong evidence of a binding site, as determined by RegulonDB, we recover all relevant transcription factor binding sites for 12 out of 16 cases, the majority of relevant binding sites for 2 out of 16 cases, and miss all or most of the regulation for 2 promoters. We identify a total of 22 previously known high evidence binding sites.



Appendix 1 Figure 1. An analysis of binding sites in the literature. (A) Information footprints for known and properly recovered binding sites. (B) A summary of how well the SortSeq results conform to literature results. The sites that are low evidence in the literature are determined by RegulonDB. (C) The information footprint and known binding sites for the *araC* promoter. Despite all the binding sites present, the only binding signature that appears is for a RNAP.

These results showcase that our method largely agrees with the established literature but also highlights several areas in which our method is prone to missing regulatory elements. One failure mode is caused by the presence of strong secondary binding sites. For example, in the *araC* promoter, as shown in Appendix 1 Figure 1(C), the only binding signatures that appear in the information footprint are from a secondary RNAP site. The secondary site seems to be expressed constitutively, and in the cases where the primary start site is even

partially repressed, the secondary start site will dominate transcription and obscure the many binding sites that are in this promoter.

If there are large numbers of regulatory elements, the data will often only show the few most important elements. If we look at the *mar* promoter in Appendix 1 Figure 1 C), we can only see the signature of the two MarR sites even though CpxR, Fis, and CRP are all known to bind to the promoter. MarR is a strong enough repressor that mutating any of the other transcription factor sites is unlikely to meaningfully change gene expression unless the MarR site is also mutated. This illustrates that the regulatory architectures discovered in this study represent a lower bound on what exists in each promoter.

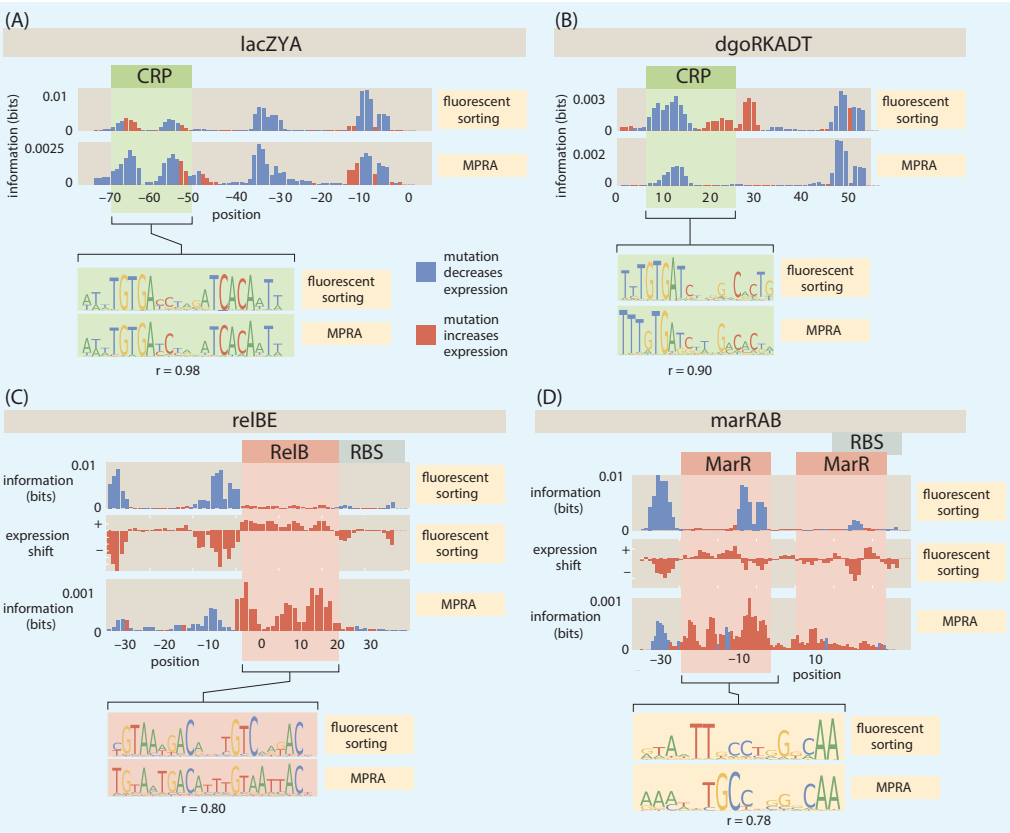
Finally, for some genes such as *dicA* there was no known TSS prior to the experiment. Although there is a small regulatory region between *dicA* and its neighboring gene, this does not insure that we will include the strongest RNAP sites. Better mapping of transcription start sites could improve our method.

We next consider low evidence binding sites. Other research determined the locations of the low evidence sites through gene expression analysis and sequence comparison to consensus sequences. For 5 promoters in our list, the binding sites location itself is not known, only that the TF in question regulates the gene. For these promoters we recover the known regulation in only 2 out of 15 cases. Comparison to consensus sequences can be unreliable can generate false positives when the entirety of the *E. Coli* genome is considered. Gene expression analysis alone has difficulty ruling out indirect effects of a given transcription factor on gene expression and regulation determined by this method may occur outside of the 160 bp mutation window we consider. As our results recover high evidence sites well, the poor recovery of sites based on sequence gazing and gene expression analysis most likely indicates that these methods are unreliable for determining binding locations.

Comparison between Reg-Seq by RNA-Seq and fluorescent sorting

As the basis for comparing the results of the fluorescence-based Sort-Seq approach with our RNA-Seq-based approach, we use information footprints, expression shifts and sequence logos as our metrics. Appendix 1 Figure 2 shows examples of this comparison for four distinct genes of interest. Appendix 1 Figure 2(A) shows the results of the two methods for the *lacZYA* promoter with special reference to the CRP binding site. Both the information footprint and the sequence logo identify the same binding site.

Appendix 1 Figure 2(B) provides a similar analysis for the *dgoRKADT* promoter where once again the information footprints and the sequence logos from the two methods are in reasonable accord. Appendix 1 Figure 2(C) provides a quantitative dissection of the *relBE* promoter which is repressed by RelBE. Here we use both information footprints and expression shifts as a way to quantify the significance of mutations to different binding sites across the promoter. Finally, Appendix 1 Figure 2(D) shows a comparison of the two methods for the *marRAB* promoter. The two approaches both identify a MarR binding site.



Appendix 1 Figure 2. A summary of four direct comparisons of measurements using fluorescence and sorting and using RNA-Seq. (A) CRP binds upstream of RNAP in the *lacZYA* promoter. Despite the different measurement techniques the two inferred energy matrices, for the CRP binding site have a Pearson correlation coefficient $r = 0.98$. (B) The *dgoRKADT* promoter is activated by CRP in the presence of galactonate. The FACS measurements were taken in the JK10 strain in the presence of 500mM cAMP. In both cases, a type II activator binding site can be identified based on the signals in the information footprint and expression shift plot in the area indicated in orange. Additionally the quantitative agreement between the CRP binding preference matrices are strong, with $r = 0.9$.(C) The *relBE* promoter is repressed by RelBE. The inferred matrices between the two measurement methods have $r = 0.8$. (D) The *marRAB* promoter is repressed by MarR. The features we can observe in the information footprint reflect this under measurement with both FACS or RNAseq. The inferred binding matrices shown have $r = 0.78$. The right most MarR site overlaps with a ribosome binding site. The overlap has a stronger obscuring effect on the sequence specificity of the FACS measurement, which measures protein levels directly, than it does on the output of the RNAseq measurement.

We note that the first aim of our methods is regulatory discovery. We would like to be able to determine how previously uncharacterized promoters are regulated and ultimately, this is a question of binding-site and transcription factor identification. For that task, we do not require perfect correspondence between the two methods. With regulatory sites identified, our next objective is the determination of energy matrices that will allow us to turn binding site strength into a tunable knob that can nearly continuously tune the strength of transcription factor binding, thus altering gene expression in predictable ways as already shown in our earlier work (***Barnes et al., 2019***). The r-values between energy matrices range from 0.78 to 0.96, indicating reasonable to very good agreement. Reg-Seq appears to be, if anything, more accurate than previous methods as it has higher relative information content in known areas of transcription factor binding and also does not have repressor-like bases on CRP sites as in Appendix 1 Figure 2 (A) and (B).

Appendix 2

Extended details of analysis methods

Information footprints

We use information footprints as a tool for hypothesis generation to identify regions which may contain transcription factor binding sites. In general, a mutation within a transcription factor site is likely to severely weaken that site. We look for groups of positions where mutation away from wild type has a large effect on gene expression. Our data sets consist of nucleotide sequences, the number of times we sequenced the construct in the plasmid library, and the number of times we sequenced its corresponding mRNA. A simplified data set on a 4 nucleotide sequence then might look like

Sequence	Library Sequencing Counts	mRNA Counts
ACTA	5	23
ATTA	5	3
CCTG	11	11
TAGA	12	3
GTGC	2	0
CACA	8	7
AGGC	7	3

One possible calculation is to take all sequences which have base b at position i and determine the number of mRNAs produced per read in the sequencing library. By comparing the values for different bases we could determine how large of an effect mutation has on gene expression. However, in this paper we will use mutual information to quantify the effect of mutation, as ***Kinney et al.*** (***2010***) demonstrated could be done successfully. In Table 1 the frequency of the different nucleotides in the library at position 2 is 40% A, 32% C, 14% G and 14% T. Cytosine is enriched in the mRNA transcripts over the original library, as it now composes 68% of all mRNA sequencing reads while A, G, and T only compose only 20%, 6%, and 6% respectively. Large enrichment of some bases over others occurs when base identity is important for gene expression. We can quantify how important using the mutual information between base identity and gene expression level. Mutual information is given at position i by

$$I_b = \sum_{m,exp} p(m,exp) log_2(\frac{p(m,exp)}{p(m)p(exp)}) \tag{3}$$

$p(b)$ in equation 3 refers to the probability that a given sequencing read will be from a mutated base. $p(exp)$ is a normalizing factor that gives the ratio of the number of DNA or mRNA sequencing counts to total number of counts.

The mutual information quantifies how much a piece of knowledge reduces the entropy of a distribution. At a position where base identity matters little for expression level, there would be little difference in the frequency distributions for the library and mRNA transcripts. The entropy of the distribution would decrease only by a small amount when considering the two types of sequencing reads separately.

We are interested in quantifying the degree to which mutation away from a wild type sequence affects expression. Although their are obviously 4 possible nucleotides, we can classify each base as either wild-type or mutated so that b in equation 3 represents only

these two possibilities.

If mutations at each position are not fully independent, then the information value calculated in equation 1 will also encode the effect of mutation at correlated positions. If having a mutation at position 1 is highly favorable for gene expression and is also correlated with having a mutation at position 2, mutations at position 2 will also be enriched amongst the mRNA transcripts. Position 2 will appear to have high mutual information even if it has minimal effect on gene expression. Due to the DNA synthesis process used in library construction, mutation in one position can make mutation at other positions more likely by up to 10 percent. This is enough to cloud the signature of most transcription factors in an information footprint calculated using equation 1.

We need to determine values for $p_i(m|exp)$ when mutations are independent, and to do this we need to fit these quantities from our data. We assert that

$$\langle mRNA \rangle \propto e^{-\beta E_{eff}} \tag{4}$$

is a reasonable approximation to make. $\langle mRNA \rangle$ is the average number of mRNAs produced by that sequence for every cell containing the construct and E_{eff} is an effective energy for the sequence that can be determined by summing contributions from each position in the sequence. There are many possible underlying regulatory architectures, but to demonstrate that our approach is reasonable let us first consider the simple case where there is only a RNAP site in the studied region. We can write down an expression for average gene expression per cell as

$$\langle mRNA \rangle \propto p_{bound} \propto \frac{\frac{p}{N_{NS}} e^{-\beta E_P}}{1 + \frac{p}{N_{NS}} e^{-\beta E_P}} \tag{5}$$

Where p_{bound} is the probability that the RNAP is bound to DNA and is known to be proportional to gene expression in *E. coli* (***Garcia and Phillips, 2011***), E_P is the energy of RNAP binding, N_{NS} is the number of nonspecific DNA binding sites, and p is the number of RNAP. If RNAP binds weakly then $\frac{p}{N_{NS}} e^{-\beta E_P} << 1$. We can simplify equation 5 to

$$\langle mRNA \rangle \propto e^{-\beta E_P}. \tag{6}$$

If we assume that the energy of RNAP binding will be a sum of contributions from each of the positions within its binding site then we can calculate the difference in gene expression between having a mutated base at position i and having a wild type base as

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = \frac{e^{-\beta E_{P_{WT_i}}}}{e^{-\beta E_{P_{Mut_i}}}} \tag{7}$$

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = e^{-\beta (E_{P_{WT_i}} - E_{P_{Mut_i}})}. \tag{8}$$

In this example we are only considering single mutation in the sequence so we can further simplify the equation to

$$\frac{\langle mRNA_{WT_i} \rangle}{\langle mRNA_{Mut_i} \rangle} = e^{-\beta \Delta E_{P_i}}. \tag{9}$$

We can now calculate the base probabilities in the expressed sequences. If the probability of finding a wild type base at position i in the DNA library is $p_i(m = WT|exp = 0)$ then

$$p_i(m = WT|exp = 1) = \frac{p_i(m = WT|exp = 0)\frac{\langle mRNA_{WT_i}\rangle}{\langle mRNA_{Mut_i}\rangle}}{p_i(m = Mut|exp = 0) + p_i(m = WT|exp = 0)\frac{\langle mRNA_{WT_i}\rangle}{\langle mRNA_{Mut}\rangle}} \quad (10)$$

$$p_i(m = WT|exp = 1) = \frac{p_i(m = WT|exp = 0)e^{-\beta\Delta E_{P_i}}}{p_i(m = Mut|exp = 0) + p_i(m = WT|exp = 0)e^{-\beta\Delta E_{P_i}}}. \quad (11)$$

Under certain conditions, we can also infer a value for $p_i(m|exp = 1)$ using a linear model when there are any number of activator or repressor binding sites. We will demonstrate this in the case of a single activator and a single repressor, although a similar analysis can be done when there are greater numbers of transcription factors. We will define $P = \frac{p}{N_{NS}}e^{-\beta E_P}$. We will also define $A = \frac{a}{N_{NS}}e^{-\beta E_A}$ where a is the number of activators, and E_A is the binding energy of the activator. We will finally define $R = \frac{r}{N_{NS}}e^{-\beta E_R}$ where r is the number of repressors and E_R is the binding energy of the repressor. We can write

$$\langle mRNA\rangle \propto p_{bound} \propto \frac{P + PAe^{-\beta\epsilon_{AP}}}{1 + A + P + R + PAe^{-\beta\epsilon_{AP}}} \quad (12)$$

If activators and RNAP bind weakly but interact strongly, and repressors bind very strongly, then we can simplify equation 12. In this case $A << 1$, $P << 1$, $PAe^{-\epsilon_{AP}} >> P$, and $R >> 1$. We can then rewrite equation 12 as

$$\langle mRNA\rangle \propto \frac{PAe^{-\beta\epsilon_{AP}}}{R} \quad (13)$$

$$\langle mRNA\rangle \propto e^{-\beta(-E_P - E_A + E_R)} \quad (14)$$

As we typically assume that RNAP binding energy, activator binding energy, and repressor binding can all be represented as sums of contributions from their constituent bases, the combination of the energies can be written as a total effective energy E_{eff} which is a sum of contributions from all positions within the binding sites.

We fit the parameters for each base using a Markov Chain Monte Carlo Method. Two MCMC runs are conducted using randomly generated initial conditions. We require both chains to reach the same distribution to prove the convergence of the chains. We do not wish for mutation rate to affect the information values so we set the $p(WT) = p(Mut) = 0.5$ in the information calculation. The information values are smoothed by averaging with neighboring values.

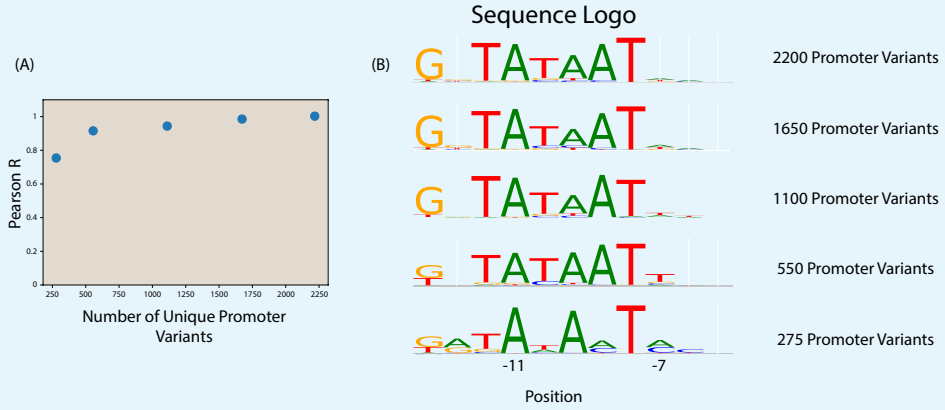
Analysis of mass spectrometry results

Mass spectrometry results were processed using MaxQuant (***Cox and Mann, 2008***) (***Cox et al., 2009***). Spectra were searched against theUniProt *E. coli* K-12 database as well as a contaminant database (256 sequences). LysC was specified as the digestion enzyme. Proteins were considered if they were known to be transcription factors, or were predicted to bind DNA (using gene ontology term GO:0003677, for DNA-binding in BioCyc). The reported binding TFs were enriched above all other other DNA binding proteins with $p < 10^{-4}$. The p-value was calculated with a two sample t-test using the python function scipy.stats.ttest_ind_from_stats. The uncertainty in the background DNA binding protein

ratios was calculated using all the enrichment ratios of proteins not determined to bind. The uncertainty in the enrichment ratio of the binding proteins was calculated from the spread in their enrichment ratios across the many (≈ 10) mass spec runs where the protein in question was not an active TF. Any dataset where the protein in question is not identified is excluded from the analysis.

Uncertainty due to number of independent sequences

1400 promoter variants were ordered from TWIST Bioscience for each promoter studied. Due to errors in synthesis, the final number of variants received was an average of 2200 per promoter. To test whether the number of promoter variants is a significant source of uncertainty in the experiment we computationally reduced the number of promoter variants used in the analysis of the *zapAB* -10 RNAP region. Each sub-sampling was performed 3 times. The results, as displayed in Appendix 2 Figure 1, show that there is only a small effect on the resulting sequence logo until the library has been reduced to approximately 500 promoter variants.



Appendix 2 Figure 1. A comparison of RNAP -10 site sequence logos. (A) This figure shows the Pearson correlation coefficient between the energy matrix models inferred from the full dataset (2200 unique promoter variants) and that from a computationally restricted dataset. (B) Sequence logos of the RNAP -10 region from each sub-sampled dataset.

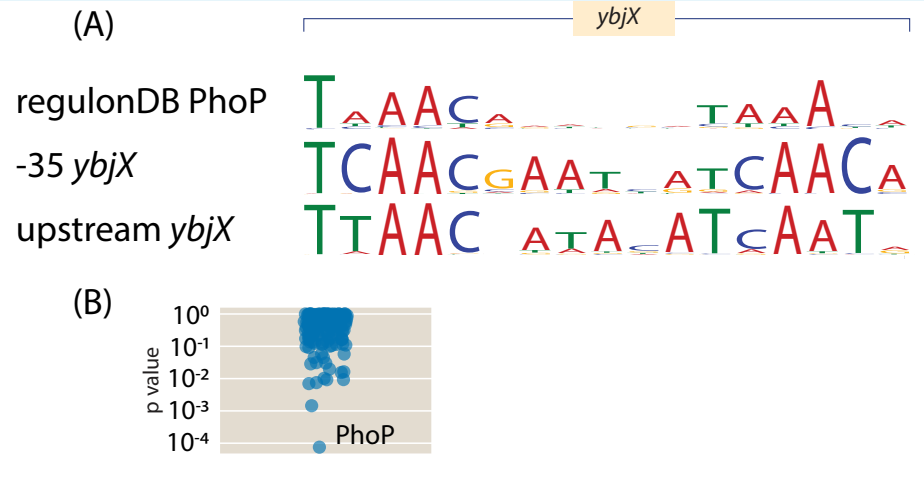
TOMTOM motif comparison

We used an alternative approach to discover the TF identity regulating a given promoter, using a motif comparison tool. TOMTOM (***Gupta et al., 2007***) is a tool that uses a statistical method to infer if a putative motif resembles to any previously discovered motif in a database. Of interest, it accounts for all possible offsets between the motifs. Moreover, it uses a suite of metrics to compare between motifs such as Kullback-Leibler divergence, Pearson correlation, euclidean distance, among others.

We performed comparisons of the motifs generated from our energy matrices to those generated from all known transcription factor binding sites in RegulonDB. Appendix 2 Figure 2 shows a result of TOMTOM, where we compared the motif derived from the -35 region of the *ybjX* promoter and found a good match with the motif of PhoP from RegulonDB.

The information derived from this approach was then used to guide some of the TF knockout experiments, in order to validate its interaction with a target promoter characterized by the loss of the information footprint. Furthermore, we also used TOMTOM to search for

similarities between our own database of motifs, in order to generate regulatory hypotheses in tandem. This was particularly useful when looking at the group of GlpR binding sites found in this experiment.



Appendix 2 Figure 2. Motif comparison using TOMTOM. Searching our energy motifs against the RegulonDB database using TOMTOM allowed us to guide our TF knockout experiments. Here we show the sequence logos of the PhoP transcription factor from RegulonDB (top) and the one generated from the ybjX promoter energy matrix. E-value = 0.01 using Euclidean distance as a similarity matrix.

Appendix 3

Extended details of experimental design

Choosing target genes

Genes in this experiment were chosen to cover several different categories. 29 genes had some information on their regulation already known to validate our method under a number of conditions. 37 were chosen because the work of ***Schmidt et al.*** (***2015***) demonstrated that gene expression changed significantly under different growth conditions. A handful of genes such as *minC*, *maoP*, or *fdhE* were chosen because we found either their physiological significance interesting, as in the case of the cell division gene *minC* or that we found the gene regulatory question interesting, such for the intra-operon regulation demonstrated by *fdhE*. The remainder of the genes were chosen because they had no regulatory information, often had minimal information about the function of the gene, and had an annotated transcription start site (TSS) in RegulonDB.

Choosing transcription start sites

A known limitation of the experiment is that the mutational window is limited to 160 bp. As such, it is important to correctly target the mutation window to the location around the most active TSS. To do this we first prioritized those TSS which have been extensively experimentally validated and catalogued in RegulonDB. Secondly we selected those sites which had evidence of active transcription from RACE experiments and were listed in RegulonDB. If the intergenic region was small enough, we covered the entire region with our mutation window. If none of these options were available, we used computationally predicted start sites.

Sequencing

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. The total library was first sequenced by PCR amplifying the region containing the variant promoters as well as the corresponding barcodes. This allowed us to uniquely associate each random 20 bp tag with a promoter variant. Any tag which was associated with a promoter variant with insertions or deletions was removed from further analysis. Similarly, any tag that was associated with multiple promoter variants was also removed from the analysis. The paired end reads from this sequencing step were then assembled using the FLASH tool ***Magoč and Salzberg*** (***2011***). Any sequence with PHRED score less than 20 was removed using the FastX toolkit. Additionally, when sequencing the initial library, sequences which only appear in the dataset once were not included in further analysis in order to remove possible sequencing errors.

For all the MPRA experiments, only the region containing the random 20 bp tag was sequenced, since the tag can be matched to a specific promoter variant using the initial library sequencing run described above. For a given growth condition, each promoter yielded 50,000 to 500,000 usable sequencing reads. Under some growth conditions, genes were not analyzed further if they did not have at least 50,000 reads.

To determine which base pair regions were statistically significant a 99 % confidence interval was constructed using the MCMC inference to determine the uncertainty.

Growth Conditions

The growth conditions studied in this experiment include differing carbon sources such as growth in M9 with 0.5% Glucose, M9 with acetate (0.5%), M9 with arabinose (0.5%), M9 with Xylose (0.5%) and arabinose (0.5%), M9 with succinate (0.5%), M9 with fumarate (0.5%), M9 with Trehalose (0.5%), and LB. In each case cell harvesting was done at an OD of 0.3. These growth conditions were chosen so as to span a wide range of growth rates, as well as to illuminate any carbon source specific regulators.

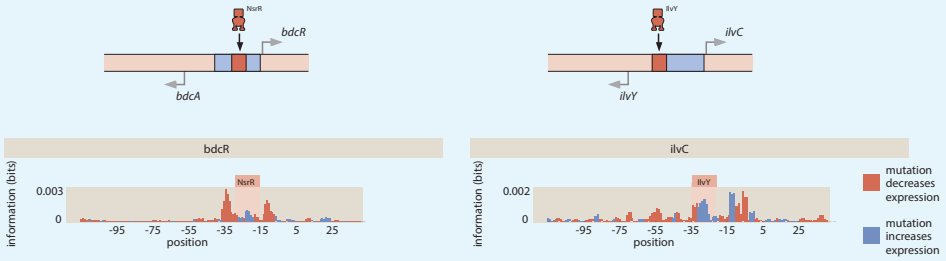
We also used several stress conditions such as heat shock, where cells were grown in M9 and were subjected to a heat shock of 42 degrees for 5 minutes before harvesting RNA. We grew in low oxygen conditions. Cells were grown in LB in a container with minimal oxygen, although some will be present as no anaerobic chamber was used. This level of oxygen stress was still sufficient to activate FNR binding, and so activated the anaerobic metabolism. We also grew cells in M9 with Glucose and 5mM sodium salycilate.

Growth with zinc was preformed at a concentration of 5mM $ZnCl_2$ and growth with iron was preformed by first growing cells to an OD of 0.3 and then adding $FeCL_2$ to a concentration of 5mM and harvesting RNA after 10 minutes. Growth without cAMP was accomplished by the use of the JK10 strain which does not maintain its cAMP levels.

All knockout experiment were preformed in M9 with Glucose except for the knockouts for *arcA*, *hdfR*, and *phoP* which were grown in LB.

Additional discussion of results

Binding sites regulating divergent operons



Appendix 3 Figure 1. A figure displaying two cases in which we see transcription factor binding sites that we have found to regulate both of the two divergently transcribed genes.

In addition to discovering new binding sites, we have discovered additional functions of known binding sites. In particular, in the case of *bdcR*, the repressor for the divergently transcribed gene *bdcA*, is also shown to repress *bdcR* in Appendix 3 Figure 2 (A). Similarly in Appendix 3 Figure 2 B) IvlY is shown to repress *ilvC* in the absence of inducer. Divergently transcribed operons that share regulatory regions are plentiful in *E. coli*, and although there are already many known examples of transcription factor binding sites regulating several different operons, there are almost certainly many examples of this type of transcription that have yet to be discovered.

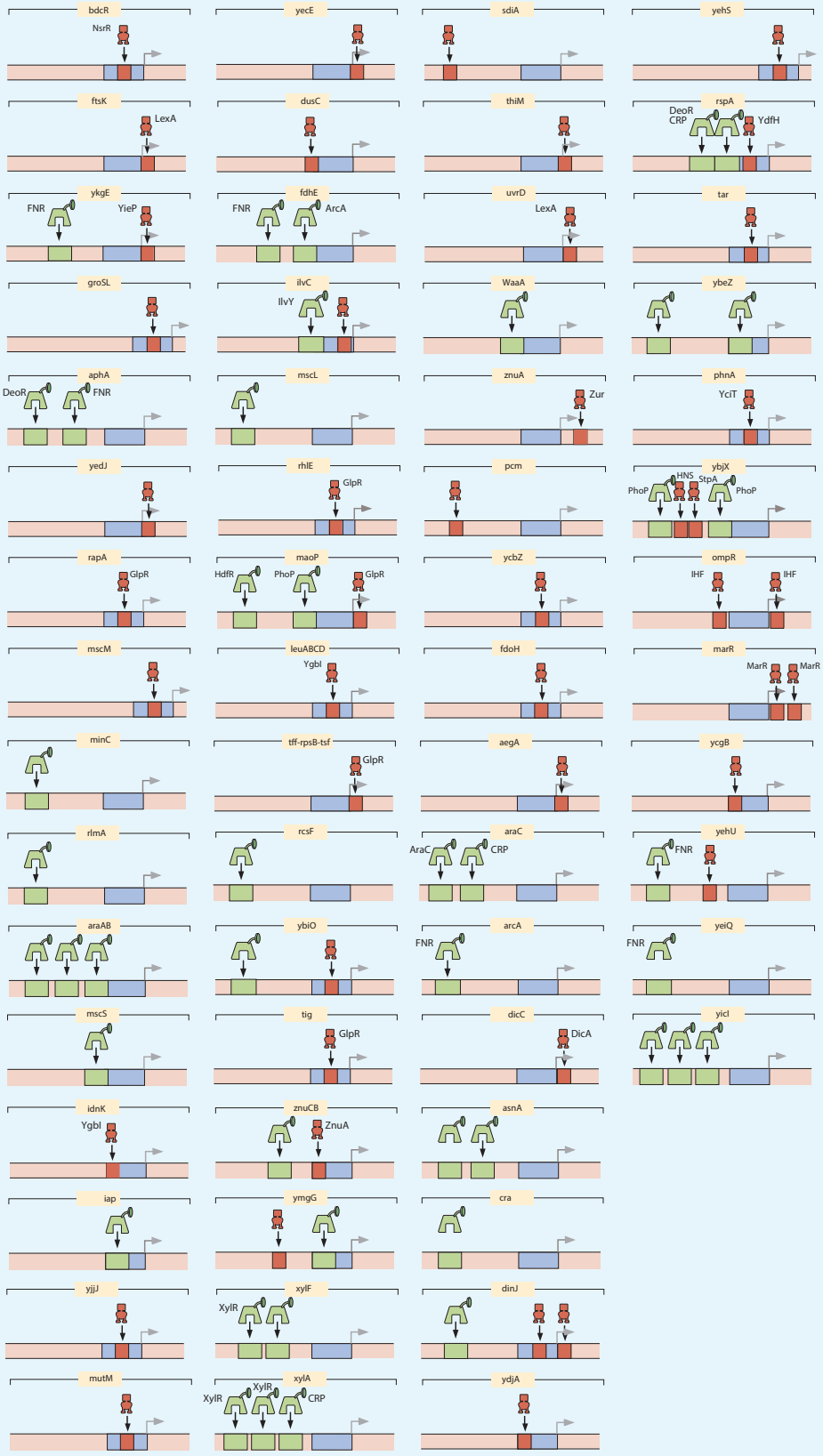
Multi-purpose binding sites allow for more genes to be regulated with fewer binding sites. However, they can also serve to sharpen the promoter's response to environmental cues. In

the case of *ilvC*, IlvY is known to activate *ilvC* in the presence of inducer. However, we now see that it also represses the promoter in the absence of that inducer. The production of *ilvC* is known to increase by approximately a factor of 100 in the presence of inducer. The magnitude of the change is attributed to the cooperative binding of two IlvY binding sites, but the lowered expression of the promoter due to IlvY repression in the absence of inducer is also a factor.

Appendix 4

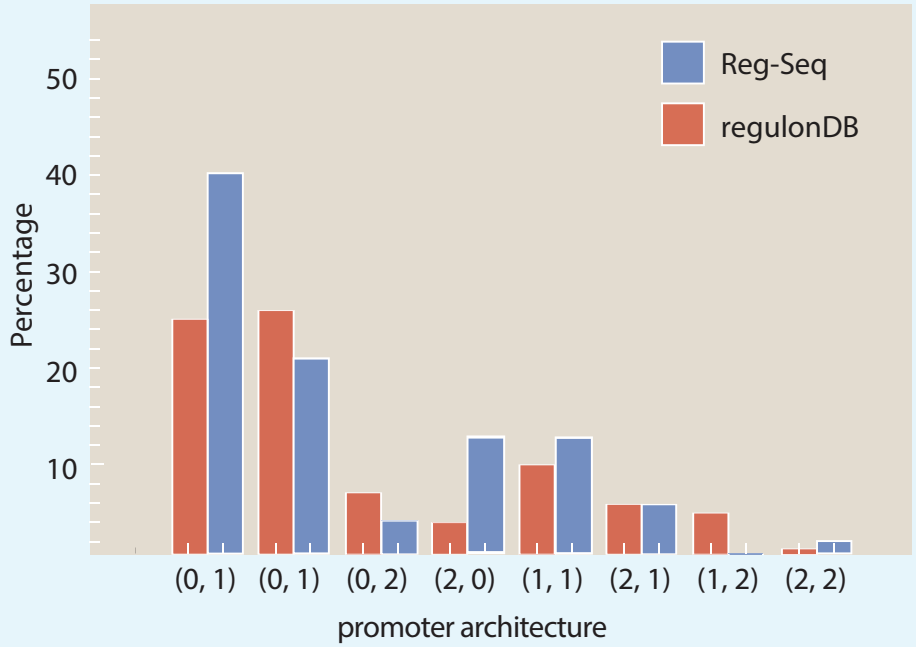
Further results.

Regulatory cartoons



Appendix 4 Figure 1. All regulatory cartoons for genes studied in this experiment.

Comparison of results to regulonDB



Appendix 4 Figure 2. A comparison of the types of architectures found in regulonDB to the architectures with newly discovered binding sites found in the Reg-Seq experiment.