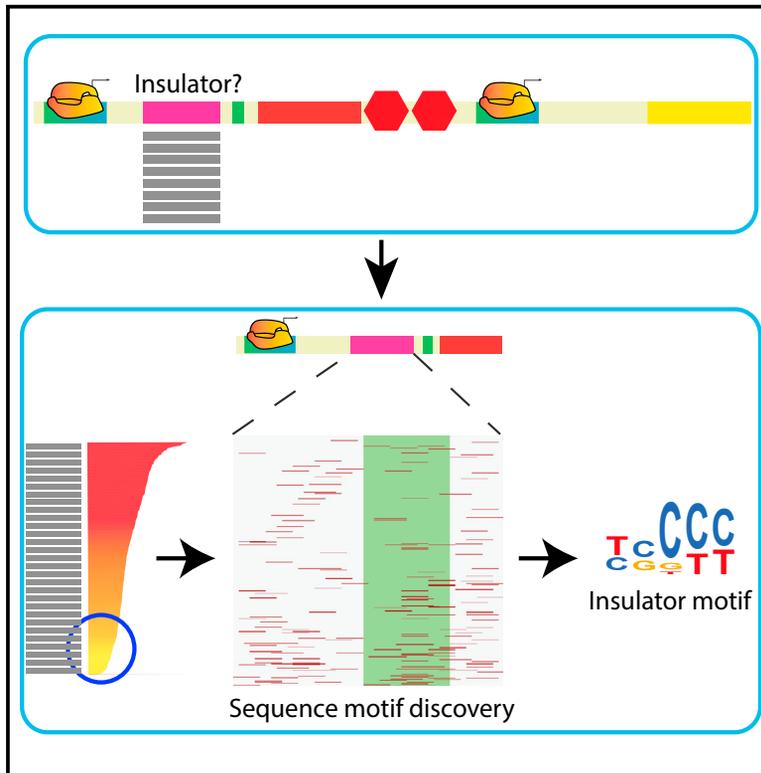


# Cell Reports

## A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial $\sigma^{54}$ Promoters

### Graphical Abstract



### Authors

Lior Levy, Leon Anavy, Oz Solomon, ..., Sarah Goldberg, Zohar Yakhini, Roei Amit

### Correspondence

amitroe4@gmail.com

### In Brief

Levy et al. identify a gene insulation phenomenon encoded within a subset of  $\sigma^{54}$  promoters in *E. coli*. The authors use an oligo library, sequencing, bioinformatics analysis, and a synthetic biology approach to show that a short CT-rich motif (3–5 bp) is responsible for the insulation phenomenon.

### Highlights

- Short CT-rich motifs (3–5 bases) are responsible for insulation effect
- Insulation strength depends on the location and number of the insulator motifs
- The insulator motifs are abundant within  $\sigma^{54}$  promoters in *E. coli*



# A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial $\sigma^{54}$ Promoters

Lior Levy,<sup>1,5</sup> Leon Anavy,<sup>2,5</sup> Oz Solomon,<sup>1,3</sup> Roni Cohen,<sup>1</sup> Michal Brunwasser-Meirom,<sup>1</sup> Shilo Ohayon,<sup>1</sup> Orna Atar,<sup>1</sup> Sarah Goldberg,<sup>1</sup> Zohar Yakhini,<sup>2,3</sup> and Roei Amit<sup>1,4,6,\*</sup>

<sup>1</sup>Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

<sup>2</sup>Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel

<sup>3</sup>School of Computer Science, Interdisciplinary Center, Herzeliya, Israel

<sup>4</sup>Russell Berrie Nanotechnology Institute, Technion - Israel Institute of Technology, Haifa 32000, Israel

<sup>5</sup>These authors contributed equally

<sup>6</sup>Lead Contact

\*Correspondence: [amitroe4@gmail.com](mailto:amitroe4@gmail.com)

<https://doi.org/10.1016/j.celrep.2017.09.063>

## SUMMARY

We use an oligonucleotide library of >10,000 variants to identify an insulation mechanism encoded within a subset of  $\sigma^{54}$  promoters. Insulation manifests itself as reduced protein expression for a downstream gene that is expressed by transcriptional readthrough. It is strongly associated with the presence of short CT-rich motifs (3–5 bp), positioned within 25 bp upstream of the Shine-Dalgarno (SD) motif of the silenced gene. We provide evidence that insulation is triggered by binding of the ribosome binding site (RBS) to the upstream CT-rich motif. We also show that, in *E. coli*, insulator sequences are preferentially encoded within  $\sigma^{54}$  promoters, suggesting an important regulatory role for these sequences in natural contexts. Our findings imply that sequence-specific regulatory effects that are sparsely encoded by short motifs may not be easily detected by lower throughput studies. Such sequence-specific phenomena can be uncovered with a focused oligo library (OL) design that mitigates sequence-related variance, as exemplified herein.

## INTRODUCTION

Deconstructing genomes to their basic parts and then using those parts to construct de novo gene regulatory architectures is a central hallmark of synthetic biology. As a first step, a thorough breakdown of a genome to its basic regulatory and functional elements is required. Then, each element can be analyzed to decipher the properties and mechanisms that drive and attenuate its activity. Lastly, well-defined and well-characterized elements can, in theory, be used as building blocks for de novo systems. However, in practice, de novo genetic systems often fail to operate as designed, due to the complex interplay between different supposedly well-characterized elements.

A possible cause of such unexpected behavior is context. Here, context refers to the DNA sequences that connect the different elements of the de novo circuit, the flanking segments within the elements, and even parts of particular elements, any of which may encode unknown regulatory roles. Often, context effects are due to short-range, sequence-based interactions with nearby elements (Korbel et al., 2004). Such interactions might confer some secondary regulatory effect that is overlooked by standard analysis methods or is masked by a stronger regulatory effect in the native setting (Farley et al., 2015). Context effects can emerge from RNA secondary structure or from larger scale genomic properties that involve nearby transcriptionally active loci. For instance, the formation of secondary structure either near the ribosome binding sites or in configurations that sequester the ribosome binding site (RBS) via hybridization by an anti-Shine-Dalgarno (aSD) sequence has been suggested as strongly inhibiting or modulating the initiation of translation (de Smit and van Duin, 1990; Schwartz et al., 1981). In bacteria, context effects have also been explored with respect to coding regions. For example, bacterial codon usage 30 nt downstream of the start codon has been shown to be biased toward unstable secondary structure and is generally GC poor as a result (Gu et al., 2010; Kudla et al., 2009). Other intragenic regulatory phenomena that have been recently proposed in bacteria involve inactive  $\sigma^{54}$  promoters (i.e., promoters that do not have an associated upstream activating sequence that binds enhancer proteins) that, instead of triggering expression, function as binding sites for large DNA-binding proteins that repress expression either internally within genes or by competing with the binding of transcriptionally active RNA polymerase (RNAP) complexes (Bonocora et al., 2015). Alternatively, dynamical processes, such as transcriptional interference by an incoming RNA polymerase or transcriptional readthrough from an upstream locus, can also alter gene expression in a way that is not encoded in the individual parts (Epshtein et al., 2003; Hao et al., 2016). In summary, in light of such diverse effects, a more systematic understanding of context-related regulatory mechanisms, their effectors, and their sequence-related determinants is needed. Such understanding is important both in the context of natural cellular processes and for the design of reliable synthetic biology modalities.



Directed evolution screens have been suggested as a technique to avoid unwanted context effects in synthetic constructs (Yokobayashi et al., 2002). These do not fit every scenario and are often impractical for actual circuit designs. Synthetic oligonucleotide libraries (OLs), together with high-throughput focused screening methods, provide an alternative approach that enables direct investigation of context-related effects. Synthetic OLs have been used to examine regulatory elements systematically and have revealed the effects of element location and multiplicity (Kinney et al., 2010; Sharon et al., 2012). In this work, we applied OL technology to investigate secondary context-related phenomena in bacterial non-coding regulatory elements. An OL of closely related variants was designed and then embedded in a synthetic transcriptional readthrough regulatory architecture to explore the underlying sequence determinants of a downregulation effect. Following the experimental scheme introduced by Sharon et al. (2012), the library was first sorted using flow cytometry into bins, which are subsequently sequenced. The combined sequencing and fluorescence dataset facilitates the extraction of individual expression distributions for each variant. As a result, this process, which has been referred to as sort-seq or flow-seq (see Peterman and Levine, 2016 for review), generates a large distribution of regulatory elements and their associated behavior.

In this study, we used sort-seq to investigate a transcriptional insulation phenomenon, first observed in smaller scale within the context of the *glnK*  $\sigma^{54}$  promoter (glnKp). In the case of our genetic circuit design, RNA of a downstream gene cannot be transcribed by the  $\sigma^{54}$  promoter itself but rather via transcriptional readthrough, namely by a RNA polymerase arriving from an upstream locus. To understand the differential silencing of a downstream mCherry protein, we designed the sort-seq library to explore hundreds of mutant variants of the original *E. coli* glnKp context and search for similar insulator contexts in other annotated  $\sigma^{54}$  promoters from multiple bacterial species. Using this focused OL, we found that silenced variants are associated with short 3- to 5-nt CT-rich segments within the  $\sigma^{54}$  promoters. Furthermore, our analysis revealed that the strength of silencing is associated with the number and the location of the short CT motifs relative to the  $\sigma^{54}$  promoter transcriptional start site (TSS). Thus, our sort-seq analysis points to a bacterial insulation mechanism encoded not by a typical position-specific scoring matrix (PSSM)-like motif but rather through the presence of short-sequence segments. CT-rich sequences are potentially complementary to the SD sequence, and we therefore hypothesize an anti-SD to SD binding as the mechanism that drives insulation and provide evidence to support this hypothesis.

## RESULTS

### The $\sigma^{54}$ glnK Promoter Silences Expression from an Upstream Promoter

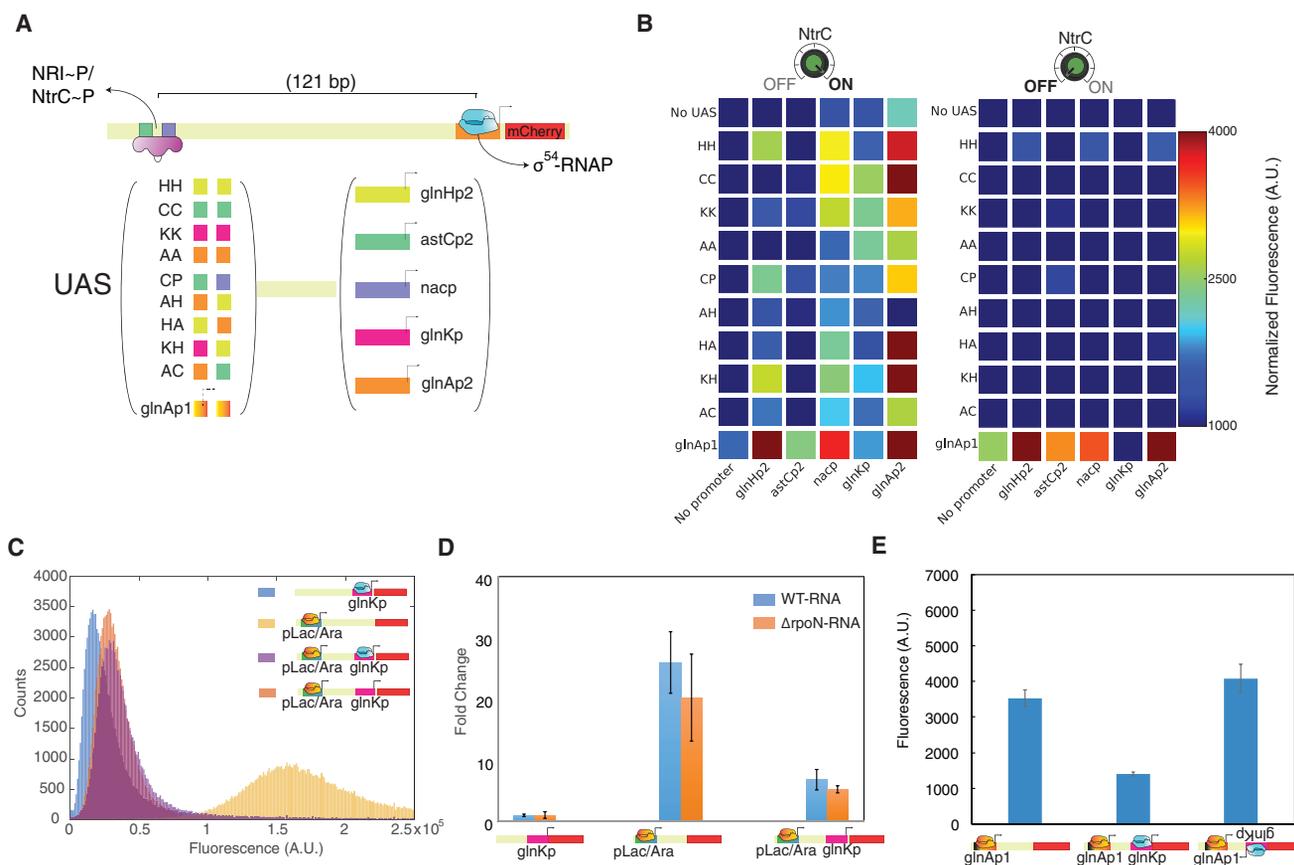
We engineered a set of synthetic circuits to test the components of bacterial enhancers in *E. coli*, initially in identical context. Bacterial enhancers typically consist of a poised  $\sigma^{54}$  ( $\sigma^{A/C}$  in gram positive) promoter, an upstream activating sequence (UAS) consisting of a tandem of binding sites for some activator protein located 100–200 bp away (e.g., NtrC, PspF, LuxO, etc.),

and an intervening sequence, facilitating DNA looping that often harbors additional transcription factor binding sites (Amit et al., 2011; Atkinson et al., 2002; Kiupakis and Reitzer, 2002). In our study of enhancer components, each synthetic circuit consisted of a UAS element and a  $\sigma^{54}$  promoter that were taken out of their natural contexts and placed in an identical context, namely with the same 70-bp loop sequence between the UAS and the TSS of the promoter and upstream of the same mCherry reporter gene (see Figures 1A, S1, and S2). We chose five *E. coli*  $\sigma^{54}$  promoters of varying known strengths (Atkinson et al., 2002; Feng et al., 1995; Kiupakis and Reitzer, 2002; Reitzer and Schneider, 2001; glnAp2, glnKp, glnHp, astCp, and nacp) and a no-promoter control. Ten UAS sequences were selected to cover a wide variety of binding affinities for NtrC and included four natural tandems, five chimeric tandems made from two halves of naturally occurring UASs, and one natural UAS, which is known to harbor a  $\sigma^{70}$  promoter overlapping the NtrC binding sites (glnAp1). Altogether, we synthesized 50 bacterial enhancers and 16 negative control circuits lacking either a UAS or a promoter. Finally, we compared mCherry levels between NtrC induction and non-induction states (see Supplemental Experimental Procedures for details of the positive-feedback synthetic enhancer circuit; Amit et al., 2011).

We report the mean fluorescence expression level data in steady state together with their variation for the synthetic enhancers as Table S3 and Figure 1B. The left panel depicts mean mCherry expression levels with NtrC induced to high titers within the cells. The plot shows that all synthetic enhancer circuits are capable of generating fluorescence expression as compared with a no  $\sigma^{54}$  promoter control. The promoters that were previously reported to be “weak” (glnHp and astCp) and naturally bound by either integration host factor (IHF) or ArgR (Claverie-Martin and Magasanik, 1991; Hoover et al., 1990; Kiupakis and Reitzer, 2002) were indeed found to generate lower levels of expression as compared with glnAp2, nacp, and glnKp (p value < 0.05,  $10^{-3}$  paired t test, for glnHp2 and astCp2, respectively). Variability of expression driven by glnAp2 is significantly higher than that of nacp and glnKp (p value < 0.01; F test for variance equality). Finally, the glnAp1 UAS that contains an overlapping  $\sigma^{70}$  promoter induces expression in the no-promoter control, as expected.

To characterize the activity of the  $\sigma^{70}$  promoter in glnAp1 (the natural UAS for glnAp2), we plot the expression level data of the synthetic enhancers with NtrC uninduced in the right panel of Figure 1B. Without NtrC,  $\sigma^{54}$  promoter expression should be silent, and indeed, the only UAS for which we observed significant expression was glnAp1 (p value < 0.05; t test after correcting for multiple testing). This is due to its dual role as a  $\sigma^{70}$  promoter in addition to being a  $\sigma^{54}$  UAS. However, glnAp1 showed a detectible fluorescence response for only four of the five promoters. The  $\sigma^{54}$  promoter glnKp manifested a different behavior. Namely, the glnAp1 UAS did not generate detectible expression with the *glnK* promoter, as compared with each of the other promoters (t tests, using the distribution of technical replicates; p value < 0.01 for all promoters). Thus, there seems to be an inhibitory mechanism embedded within the  $\sigma^{54}$  promoter glnKp.

We initially reasoned that the inhibitory phenomenon might be explained by unusually tight binding of the  $\sigma^{54}$ -RNAP complex to the glnKp core region, leading to the formation of a physical



**Figure 1. The glnKp  $\sigma^{54}$  Promoter Can Downregulate Another Promoter Positioned Upstream**

(A) Synthetic enhancer design showing the different UAS and  $\sigma^{54}$  promoter combinations used in the experiment.

(B) (Left) mCherry expression with enhancer switched to “on” (NtrC induced), showing varying response for each promoter. Note that, for the dual UAS- $\sigma^{70}$  promoter glnAp1, there is expression with the “no promoter” control. (Right) mCherry expression for enhancers switched to “off” (NtrC not induced) is shown, showing “on” behavior for all enhancers containing the dual UAS- $\sigma^{70}$  promoter, except for the enhancer with the glnKp. Error level for the mean expression values is provided in Table S3.

(C) Flow cytometry data comparing mCherry fluorescence for the glnKp strain in the *E. coli* TOP10 strain (purple) and in the  $\sigma^{54}$  knockout strain (TOP10: $\Delta rpoN$ , orange).

(D) qPCR data showing a reduction in mRNA level in the silenced strain (right) as compared with non-silenced strains (middle) and the no- $\sigma^{70}$  control (left).

(E) Plate reader data showing rescue of mCherry fluorescence when the orientation of the glnKp is flipped relative to the upstream  $\sigma^{70}$  promoter.

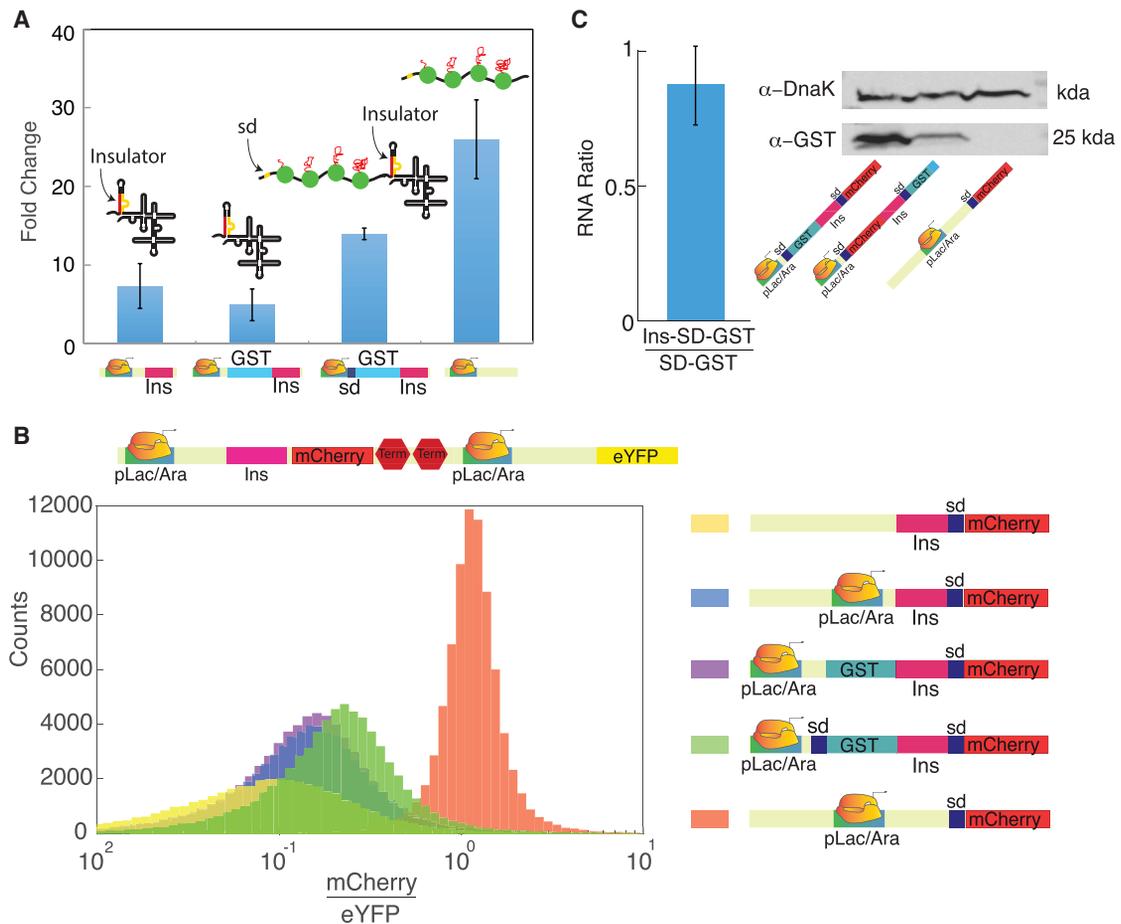
In (B) and (E), the value of 1,000 a.u. corresponds to the mean background fluorescence level observed on the Tecan plate reader with our constructs. Data are represented as mean  $\pm$  SD. Associated with Figure S1 and Tables S1–S3.

“road block,” which interferes with any upstream transcribing RNAP holoenzymes. To check this hypothesis, we constructed another gene circuit in which a pLac/Ara ( $\sigma^{70}$ ) promoter was placed upstream of the  $\sigma^{54}$  glnKp instead of the glnAp1 UAS. In Figure 1C, we show that the circuit with both the pLac/Ara promoter and the glnKp (purple) generates about a factor of ten less fluorescence than the control lacking glnKp (yellow). However, when the circuit was placed in a  $\Delta rpoN$  knockout strain (*rpoN* encodes the  $\sigma^{54}$  RNAP subunit), the same reduction in fluorescence was observed (orange). Moreover, in Figure 1D (center and right bars), we show that the reduction was observed not only at the protein level but also at the mRNA level, albeit to a lesser extent. The effect was observed only for glnKp oriented in the 5′-to-3′ direction relative to pLac/Ara, as flipping the orientation of the 50-bp glnKp sequence abolished the inhibitory effect (Figure 1E). Consequently, in the context of our construct,

the glnKp sequence not only encodes a  $\sigma^{54}$  promoter but also some function that leads to silencing and that is active when this sequence is placed downstream of an active  $\sigma^{70}$  promoter and upstream to the mCherry start codon. This silencing occurs both with and without rpoN.

### glnK Promoter Encodes an Insulator Sequence

In order to study a closer to nature configuration involving the glnK promoter, we encoded the full *glutathione S-transferase* (*GST*) gene upstream of the glnKp (under the control of pLac/Ara), with and without its own SD motif. By adding a gene upstream of the glnKp, we engineered a system that closely resembles a typical genomic architecture of one operon following another, thus allowing for transcriptional readthrough of the downstream gene from the upstream promoter. We reasoned that a translated gene placed upstream of an aSD sequence



**Figure 2. GlnKp Downregulation Is Generated by an aSD:SD Interaction**

(A) qPCR measurements showing the rescue of mCherry mRNA levels by a translated GST placed upstream of glnKp (third bar from left), as compared with a non-translated GST (second bar from left). The no-glnKp control is shown for comparison (fourth bar from left).

(B) Flow cytometry data showing strong insulation despite adding the *gst* gene. (Top) Circuit diagram shows added eYFP module without an insulating component for mCherry to eYFP expression level ratio measurements (e.l.r.). (Bottom) Data for the following constructs are shown: no  $\sigma^{70}$  promoter control (yellow); glnKp construct (blue); glnKp construct with a *gst* gene encoded upstream with (green) and without an RBS (purple), respectively; and the no- $\sigma^{54}$  promoter control (red).

(C) Insulation of a GST context. (Left) RT-PCR ratio of insulated and non-insulated GST RNA showing no effect is shown. (Right) Western blot with  $\alpha$ -GST comparing non-insulated (left band), insulated (middle band), and no-GST control (right band) is shown.

Data are represented as mean  $\pm$  SD. Associated with Figure S2.

would protect the entire mRNA from the pyrophosphatation of the 5' end by RppH (Deana et al., 2008). This, in turn, would inhibit the RnaseE degradation pathway (Mackie, 1998), leading to a partial decrease of the mCherry silencing effect. Previous studies (Calin-Jageman and Nicholson, 2003; Richards et al., 2012; Robertson, 1982) have shown that a folded RNA state frequently triggers increased degradation of the untranslated RNA. Because we also observed reduced RNA levels in the silenced strain (Figure 1D), we wanted to rule out the possibility of the silencing effect being a degradation artifact of our original circuit design of two closely positioned promoters. The qPCR results of the additional GST strains are shown in Figure 2A. We plot data for four strains: glnKp variant without the GST gene (left bar); GST+glnKp+mCherry (second from the left); RBS+GST+glnKp+mCherry (second from the right); and a non-

silenced strain (right). In Figure 2A, it can be seen that the mCherry mRNA level for the strain with a non-translated GST gene encoded upstream is identical to the one measured for the glnKp variant and can thus be considered silenced. However, when the GST is translated, the mCherry mRNA levels rise considerably by a factor of  $\sim 3$ , representing approximately 50% recovery as compared with the non-silenced strain.

Next, we measured the recovery in expression level ratio of mCherry on the same strains. To get a quantitative assessment of the relative increase between the different strains, we added a non-insulated circuit expressing eYFP on the same plasmid (Figure 2B, top), which allowed us to measure an mCherry to eYFP expression level ratio (e.l.r.) that is less prone to expression level noise. In Figure 2B, bottom, we plot the flow cytometry distributions for the following strains: the glnKp

variant without an active upstream promoter (yellow); *glnKp* variant without *GST* (blue); *GST+glnKp+mCherry* (purple); *RBS+GST+glnKp+mCherry* (green); and the variant with a non-silencing 5' UTR (red; note that, in the legend, the *glnKp* promoter is labeled in pink as “insulator”). A close examination of the data shows that an insignificant increase in e.l.r. (Figure 2C) is measured for the translationally active *GST* strain as compared with the e.l.r. levels of the non-silenced strain. Whereas this e.l.r. level is consistent with the partial recovery in mRNA levels shown in Figure 2A (two-fold as compared with three-fold), it remains smaller by about ten-fold from the construct that does not encode the putative silencing 5' UTR sequence.

Given the recovery in mRNA levels and the lack thereof in e.l.r., we hypothesized that the silencing phenomenon we observed occurs mainly at the post-transcriptional level. A close examination (Figure S2) of the sequence encoded in the flanking region of *glnKp* reveals that there is a CU-rich segment located ~20 bp upstream of the SD sequence. This sequence can potentially form a hairpin structure with the sequence. RBS sequestration via hairpin formation has been previously implicated in gene silencing (de Smit and van Duin, 1990; Schwartz et al., 1981). A hairpin presumably inhibits the formation of an elongating 70S ribosomal subunit in bacteria, thus leading to the silencing of the downstream gene. In our case, this aSD sequence is encoded within the *glnK* promoter and cannot be transcribed by the  $\sigma^{54}$  promoter itself. Instead, it can only become active and insulate the downstream  $\sigma^{54}$ -regulated gene from translation when the template mRNA is transcribed via transcriptional read-through-based events. A visual representation of possible secondary structures (Hofacker et al., 1994) for the 5' UTR region of the constructs with the no  $\sigma^{54}$  promoter (top) and with the *glnKp* context (bottom) is shown in Figure S2. The structure models suggest that, whereas the RBS for the construct without  $\sigma^{54}$  promoter remains single stranded, the one for the *glnKp* is sequestered in a double-stranded hairpin structure. To provide additional support for this assertion, we switched the 5'–3' order of the *mCherry* and *GST* genes in our circuit (Figure 2C). In the new construct, only the *GST* gene was subject to insulation, whereas *mCherry* was not subject to any putative inhibition encoded in its RNA sequence. We then measured the level of *GST* RNA, *mCherry* expression, and *GST* protein levels. The plot shows that, whereas the *GST* RNA level remained approximately the same for both the insulated and non-insulated configurations (Figure 2C, left), the amount of *GST* protein was sharply reduced when the gene was placed downstream of the insulator element within the *glnK* promoter (Figure 2C, right), in a manner similar to the reduction in *mCherry* fluorescence (Figure 2B). Consequently, insulation functions independently of downstream genetic context, consistent with it being encoded as an independent regulatory element at the RNA level.

### OL Analysis of *glnKp* Mutants

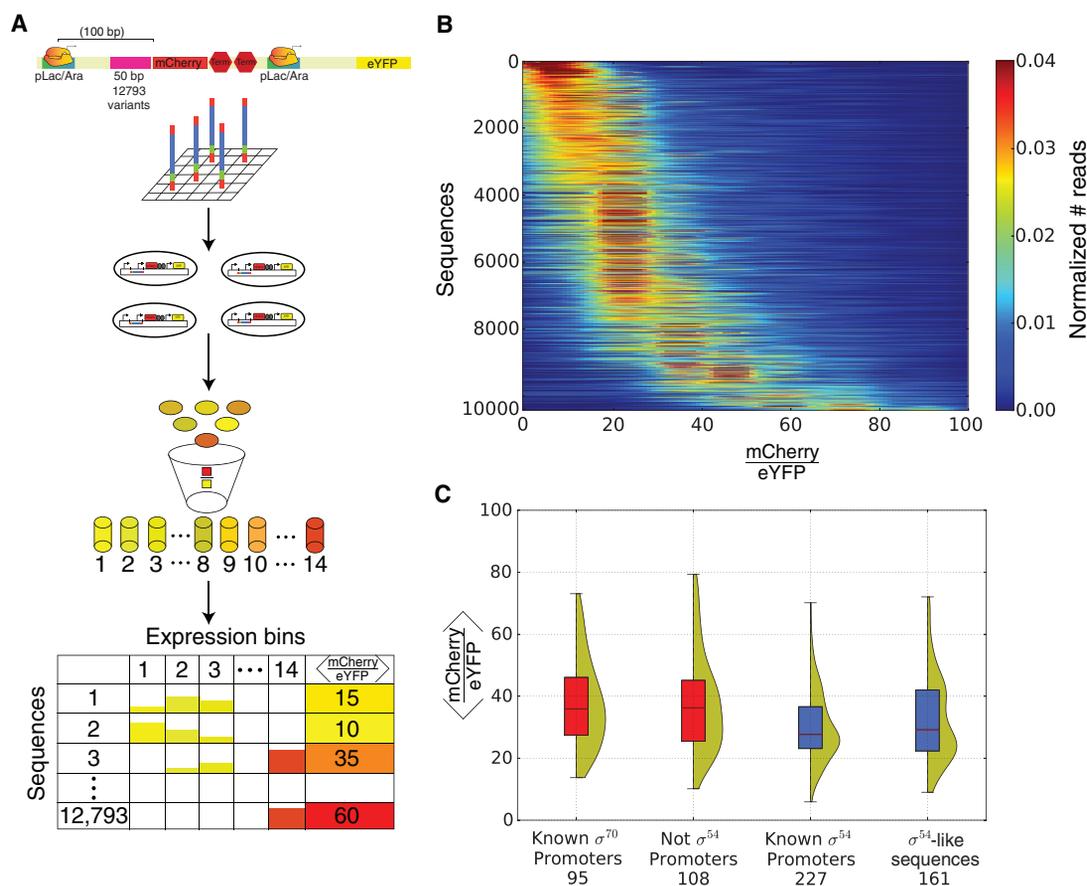
To further examine our hypotheses, to explore and characterize the insulating sequence and mechanism encoded by *glnKp*, and to check for its prevalence in other bacterial genomes, we constructed an OL of 12,758 150-bp variants (Figure 3A). The OL was synthesized by Twist Bioscience (for technical characteris-

tics, see Figures S3 and S4) and inserted into the synthetic enhancer backbone following the method introduced by Sharon et al. (2012). The OL was designed to screen both known  $\sigma^{54}$  promoters from various organisms and  $\sigma^{54}$ -like sequences from genomic regions in *E. coli* and *V. cholera* and examine the silencing effect, essentially searching for similarities to the phenomenon observed in *glnKp* and for their potential sequence determinants. In addition, the OL was designed to characterize 134 *glnKp* mutational variants. Finally, the OL was designed to conduct a broader study of the contextual regulatory effects induced by a downstream genomic sequence, in either a sense or anti-sense orientation, on an active upstream promoter positioned nearby. Each variant consisted of a pLac/Ara promoter, followed by a variable sequence, an identical RBS, and a *mCherry* reporter gene, thus encoding a 5' UTR region with a variable 50-bp region positioned at +50 bp from the pLac/Ara TSS (Figure 3A). Similar to the experiment shown in Figure 2, each plasmid also contained an eYFP control gene to eliminate effects related to copy number differences and to enable proper normalization of expression values. We combined OL with fluorescence-activated cell sorting and next-generation sequencing (Kinney et al., 2010; Sharon et al., 2012), yielding the distribution of e.l.r. for each sequence variant. Figure 3B shows the e.l.r. distribution profiles for 10,438 variants with sufficient total number of sequence counts ( $n > 10$ ; see Supplemental Experimental Procedures for details), revealing a broad range of distributions of expression levels. Whereas a significant percentage of the variants showed low mean e.l.r., similar to what was observed for *glnKp*, a non-negligible set of variants produced high and intermediate expression levels (Figure S5, top and bottom, show representative examples), indicating that a combination of regulatory mechanisms encoded within these sequences may underlie this distribution of expression levels.

To examine the observed silencing and the hypothesized insulation in broader context beyond *glnKp*, we analyzed the mean e.l.r. values for four groups of variants in our library: annotated  $\sigma^{54}$  promoters from various organisms obtained from the list compiled by Barrios et al. (1999) and EcoCyc database; annotated  $\sigma^{70}$  promoters obtained from EcoCyc;  $\sigma^{54}$ -like sequences (score  $> 0.9$ —see Supplemental Experimental Procedures for details); and non- $\sigma^{54}$  promoters (score  $< 0.5$ —see Supplemental Experimental Procedures for details). Checking the distributions of mean e.l.r. within defined classes of variants (Figure 3C), we observed that the “not  $\sigma^{54}$  promoter” class and the annotated  $\sigma^{70}$  distributions are shifted toward higher expression levels as compared with the annotated  $\sigma^{54}$  promoters and  $\sigma^{54}$ -like distributions. A non-parametric Mann-Whitney U test shows that the mean e.l.r. computed for each of the two former classes is significantly higher than those of the latter classes ( $p$  value  $< 10^{-4}$  compared to annotated  $\sigma^{54}$  promoters and  $p$  value  $< 0.01$  compared to  $\sigma^{54}$ -like variants). This analysis suggests that there may be a conserved sequence determinant within the  $\sigma^{54}$ -variant classes that contributes to this shift down in expression levels.

### OL Analysis of Broader Insulation

To search for potential sequence determinants that may be associated with the differences in e.l.r. distributions, we performed a DRIMust k-mer search on the OL variants sorted by



**Figure 3. Oligo Library Analysis of the Insulation Phenomenon**

(A) Oligo library design and schematic description of the protocol. In brief, the synthesized OL (Twist Bioscience) was cloned into *E. coli* competent cells, which were then grown and sorted by FACS into 14 expression bins according to mCherry to eYFP fluorescence or e.l.r. DNA from the cells of each bin was barcoded and pooled into a single sequencing run to produce an e.l.r. profile for each variant. For details, see [Experimental Procedures](#).

(B) Library expression distribution. Heatmap of smoothed, normalized number of reads per expression bin obtained for 10,438 analyzed variants ordered according to increasing mean e.l.r. is shown.

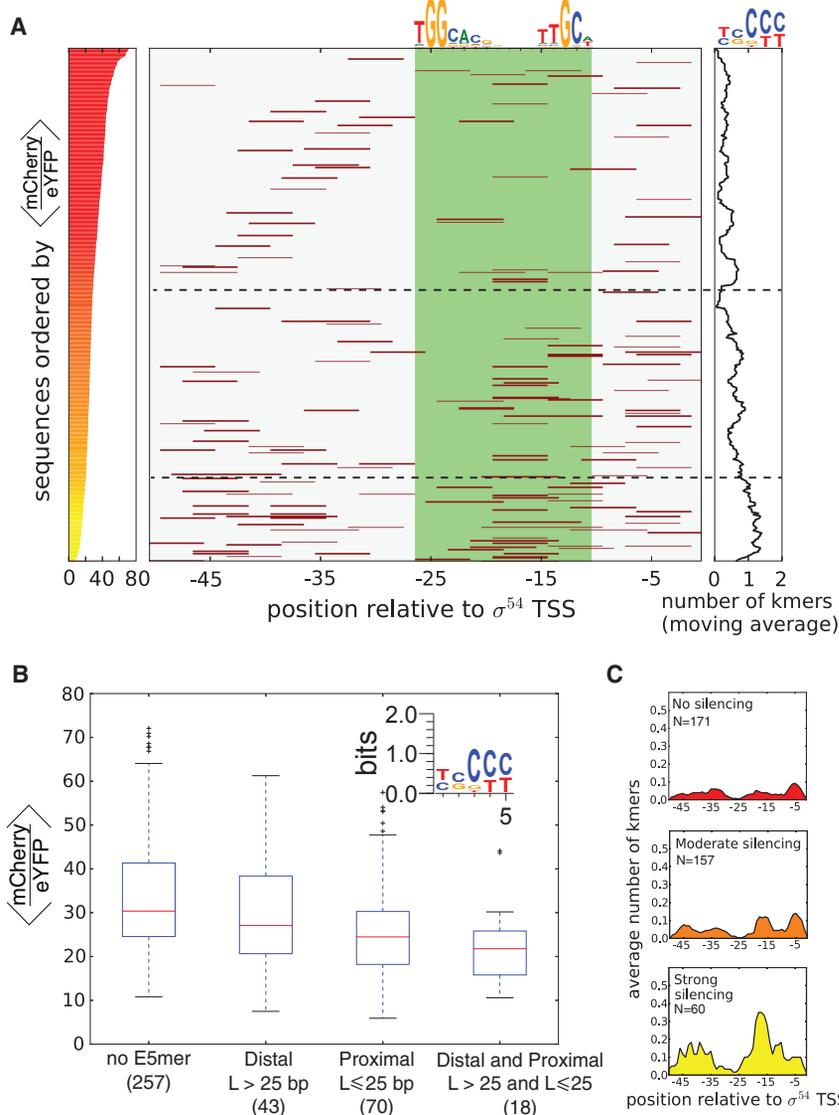
(C) Violin plots showing mean expression value distribution for the following variant groups in the library: known  $\sigma^{70}$  promoters (95 variants); no promoter (108 variants); known  $\sigma^{54}$  promoters (227 variants); and  $\sigma^{54}$ -like sequence (161 variants).  $\sigma^{54}$ -like sequence variants were selected based on similarity to the core promoter consensus sequence ([Barrios et al., 1999](#)).

Associated with [Figures S3–S5](#) and [Table S4](#).

mean e.l.r. values. DRIMust is a tool designed to identify enriched sequence motifs in a ranked list of sequences ([Eden et al., 2007](#); [Leibovich and Yakhini, 2012](#); [Leibovich et al., 2013](#)). Our analysis revealed that a single CT-rich consensus motif is enriched in the silenced variants (p value <  $10^{-54}$ ; minimum hypergeometric [mHG] test). We plot the results for 388 variants that are either annotated  $\sigma^{54}$  promoters (227 variants) or  $\sigma^{54}$ -like sequences (score > 0.9—161 variants) in [Figure 4A](#). The consensus motif is derived from a list of ten 5-bp CT-rich features, each enriched in the top of the ranked list (see [Figure 4A](#), top right). We call these enriched 5 mers “E5mers”. In the left panel, we plot the mean e.l.r. value for each variant, ordered by decreasing e.l.r. from top to bottom. The middle panel of [Figure 4A](#) shows the position of each E5mer, marked by a brown line in the corresponding variant on all 388 putative and annotated  $\sigma^{54}$  promoters. In the right panel, we show a 20-variant running average of the number of E5mer occurrences.

This analysis shows the correlation between the presence of an E5mer and the mean e.l.r. value in the  $\sigma^{54}$  promoter group. Together, the plots show that a high concentration of CT-rich motifs close to the purine-rich sequence that encodes the SD motif (positioned at +17) is strongly associated with variants leading to low mCherry to eYFP fluorescence ratio.

To further examine the dependence of the mean e.l.r. on the position of E5mers within the annotated and putative  $\sigma^{54}$  promoters, we grouped the 388 sequences into four classes ([Figure 4B](#)): containing no E5mers; E5mers located at a distal position from the  $\sigma^{54}$  promoter’s TSS (i.e., more than 25 bp upstream of the TSS); E5mers located at a TSS proximal position (i.e., less than 25 bp); and E5mers located at both the proximal and distal positions. In addition, we plot in [Figure 4C](#) the average number of E5mers per position (with the  $\sigma^{54}$  promoter TSS defined as 0) depicted over three regimes: strong insulation (mean e.l.r. < 20); moderate insulation (20 < mean



**Figure 4. Insulation Phenomenon Is Prevalent in Other  $\sigma^{54}$  Promoters**

(A) (Left) Heatmap ordering of the examined variants by mean e.l.r. value, with silenced variants at the top. (Middle) For each variant in the left panel, each enriched 5-mer (E5mer) appearance is marked by a brown line at its position within the variant sequence. (Green shade)  $\sigma^{54}$  core promoter region is shown. (Top)  $\sigma^{54}$  core promoter consensus sequence is shown (Barrios et al., 1999). (Right) Running average on the number of E5mers observed within a variant in the ordered heatmap is shown. (Top) A PSSM summarizing a multiple alignment of the E5mers found with DRIMust is shown.

(B) Box plot showing groups of  $\sigma^{54}$ -like and annotated  $\sigma^{54}$  promoters differentiated by the location of E5mers within the 50-bp variant sequence. Data are represented as boxes set at quartiles and whisker values set to  $\pm 1.5 \times$  box length.

(C) Plots depicting the average number of E5mers found per position on the 50-bp variants for putative and annotated  $\sigma^{54}$  promoters.

Variants are grouped by strong (bottom panel and variants below lower dashed line in B), moderate (middle panel and variants in between dashed lines in B), and no (top panel and variants above middle dashed line in B) silencing variants, respectively. Associated with Tables S5 and S6.

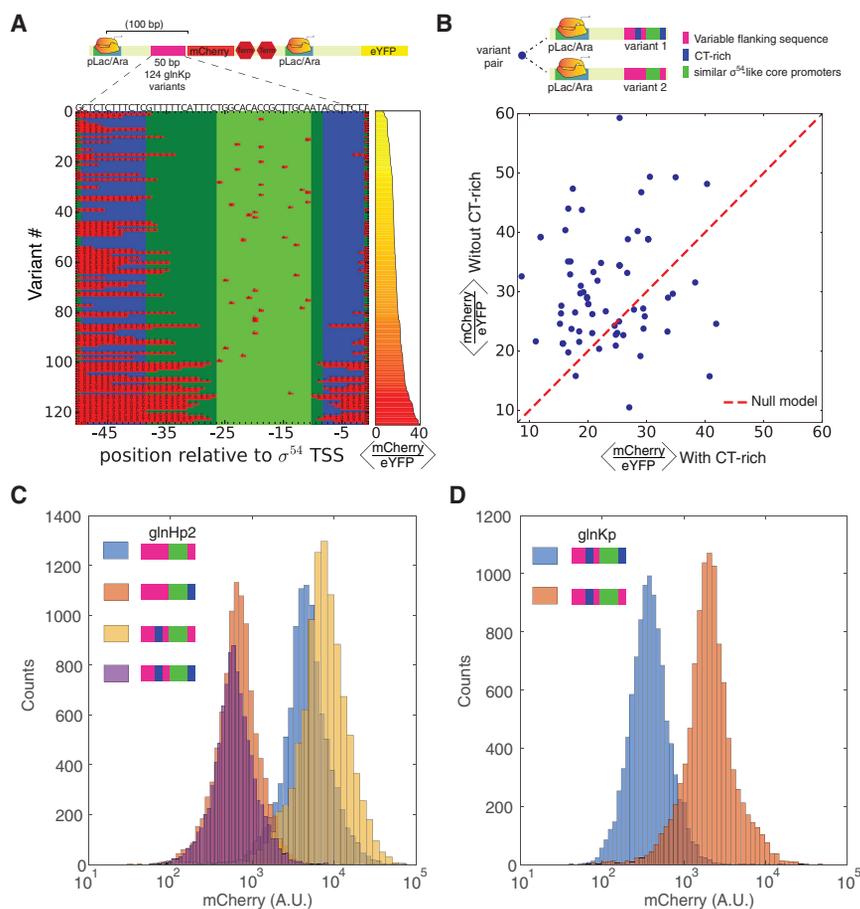
and  $\sigma^{54}$ -like sequences (e.l.r. < 20; Figure 4C, bottom) out of the 388 tested in our OL.

### Deletion of CT-Rich Motifs Relieves the Insulation Effect

To provide further evidence that the CT-rich segments identified in our library underlie the insulation phenomenon, we analyzed 123 mutant variants of the *glnK* promoter that yielded  $n > 10$  sequence counts (Figure 5A). The figure shows the original *glnK* sequence context at the top and the set of

e.l.r. < 30); and no/weak insulation (mean e.l.r. > 30). We chose 25 bp as the threshold separation between proximal and distal positions due to the presence of a conserved sequence in the  $\sigma^{54}$  core promoter region (Figure 4A, center top) that is not an enriched region. The data show (Figures 4B and 4C) that there is a clear correlation between the number and position of the E5mers to the mean e.l.r. In particular, an ANOVA test (see Tables S5 and S6) on the mean e.l.r. and the position of the E5mers shows that both proximal and distal E5mers affect the mean e.l.r., with the proximal effect being much more significant than the distal effect ( $p$  value <  $10^{-6}$  and 0.05, respectively). The test also shows that these effects are additive. The pattern of insulation with E5mers at distal and proximal locations was detected in 18 annotated and  $\sigma^{54}$ -like promoters from various organisms (see Data S1) and is consistent with the observations for the *glnK*p, which also manifested this pattern. Overall, we found 60 strongly insulating  $\sigma^{54}$  promoters

mutations to the original context for each mutant plotted as individual lines of text over a red-shaded box below. The lines of text overlay a colored map corresponding to previously identified regions in the *glnK*p context. These regions include the CT-rich flanking regions (blue), core sequence of  $\sigma^{54}$  promoter (light green), and unclassified flanking region (green). We arranged the mutant *glnK*p variants in order of increasing e.l.r. value from top to bottom with the most silenced variant at the top line (see heatmap gradient on the right, corresponding to the e.l.r. value measured for each *glnK*p mutant variant). The figure shows that the mutations in the core sequence region and in the distal CT-rich region (left, blue-shaded region) did not correlate with increasing e.l.r. but are rather evenly distributed throughout the mean e.l.r. range. However, increased amount of mutations in the proximal CT-rich segments of the flanking region (right, blue shade) and in positions immediately upstream of the TSS correlate strongly with elevated mean e.l.r. In particular, mutations in



**Figure 5. Analysis of CT-Rich Deletions in Insulated Variants**

(A) Analysis of the glnKp mutation subset of the library. (Top) Circuit schematic shows the location of the variant sequence. (Bottom) glnKp mutated variants with glnKp sequence are shown on top. Flanking regions, core  $\sigma^{54}$  promoter, the CT-rich regions, and mutations are denoted by dark green, light green, blue, and red boxes, respectively. Right panel denotes the mean e.l.r. value using a yellow-to-red scale.

(B) (Top) schematic showing the variant pair contents by region: (pink) variable flanking sequence; (blue) E5mers; and (green) core  $\sigma^{54}$  promoter. (Scatterplot) Analysis of expression levels of  $\sigma^{54}$ -promoter-like sequence pairs with and without a CT-rich region. Each point represents a pair of nearly identical sequences (see [Experimental Procedures](#)), one containing a CT-rich region in the flanking region and the other does not. In most pairs, the CT-rich-containing sequence (x axis) presents lower e.l.r. values (yellow points) than the one lacking a CT-rich region (y axis), indicating that the CT-rich region is related to the silencing phenomenon.

(C) Flow cytometry measurements of the *glnH* promoter mutated variants: (blue) native; (orange) proximal flanking region replaced by an E5mer segment; (yellow) distal flanking region replaced by an E5mer; and (magenta) both distal and proximal positions replaced by an E5mer segment. (D) Flow cytometry measurements of a *glnK* promoter variant with only the proximal E5mer mutated (red) as compared with the native sequence (blue).

the 7-nt CT-rich region (centered at  $-4$ ) into a G or an A yielded the largest increase, suggesting that the insulating effect was likely abolished.

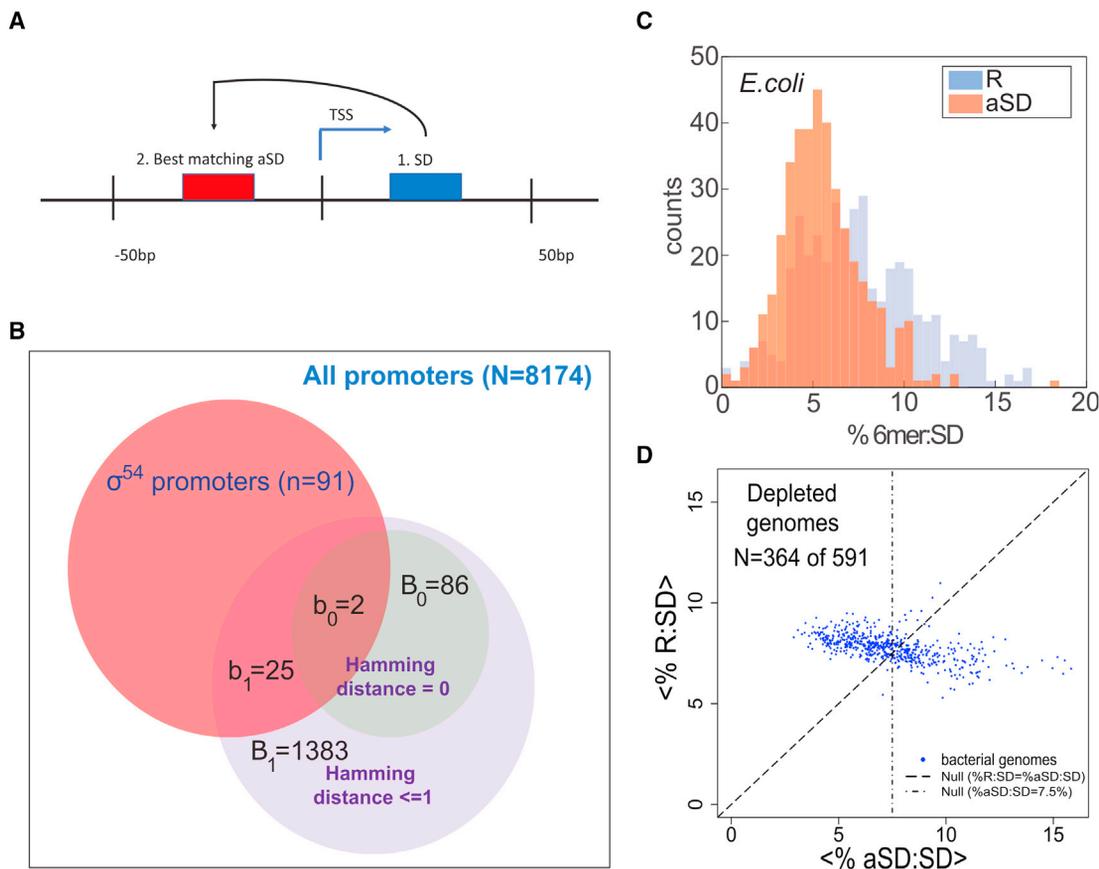
We next investigated pairs of variants with high sequence similarity in the putative  $\sigma^{54}$  core promoter region to further compare the effect of E5mers' presence on insulation. Variants were considered to be similar if at least 11 of the 16 core promoter bases were identical. Each pair consisted of one sequence variant whose flanking region contained the E5mer identified in [Figure 4A](#), whereas the other did not. Furthermore, the remaining 29 bp of unclassified flanking region were identical in at least 21 of the 29 bases (see [Figure 5B](#), top, for a schematic description of the variant pairs). Thus, this analysis compares  $\sigma^{54}$ -promoter-like variant that contains an E5mer to a close one that lacks it. In the plot ([Figure 5B](#), bottom), each variant pair is represented by a circle whose x coordinate and y coordinate correspond to the mean e.l.r. value measured for the variant that contains and lacks the E5mer, respectively. To check whether there is a bias toward higher e.l.r. values for variants that lack the E5mer, we add to the scatterplot a dashed line representing the null assumption that  $x = y$ . The plot shows, in comparison to the dashed line, the scatterplot is biased toward higher e.l.r. values, indicating that variants that have higher e.l.r. values (as compared to their paired variant) are typically those lacking the E5mer. Overall, we found 72 non-glnKp  $\sigma^{54}$ -core-promoter-like variant pairs in our library

with 52 pairs positioned above the  $x = y$  line and 20 below, yielding a p value of  $\leq 6.2 \times 10^{-7}$  (two-tailed t test versus an equal means null hypothesis).

To confirm that a proximal occurrence of the E5mer is the causative sequence element for insulation of a downstream gene, we designed four new constructs based on the *E. coli* non-insulated *glnH* and insulated *glnK* promoter contexts (see [Figure 1](#)). For glnHp2 variant ([Figure 5C](#); [Supplemental Experimental Procedures](#)), we replaced the natural context with three mutated constructs containing either proximal, distal, or both proximal and distal occurrences of the E5mer. For glnKp ([Figure 5D](#); [Supplemental Experimental Procedures](#)), we mutated the proximal occurrence of the E5mer to a randomized five nucleotide sequence. We then measured the mCherry expression level for all mutated constructs in comparison to the original promoters. The data show that, for the *glnH* promoter ([Figure 5C](#)), an E5mer insertion in a proximal position to the TSS is sufficient for insulation, whereas for the *glnK* promoter ([Figure 5D](#)), deletion of the proximal E5mer eliminates the insulation effect.

### CT-Rich Segments Are Depleted in Bacterial Genomes Upstream of Putative RBS

Next, we performed bioinformatics analysis to examine the occurrence of aSD:SD potential interactions around  $\sigma^{54}$



**Figure 6. Analysis of Prevalence of CT-Rich k-mers around *E. coli* Promoters and in Other Bacterial Genomes**

(A) A scheme for the analysis of the occurrences of aSD:SD around  $\sigma^{54}$  TSS positions. In step 1, we scan the 50-bp region downstream to TSS and locate the SD. In step 2, the best matched aSD, i.e., the hexamer that is the best Hamming reverse complement to the SD we found in step 1, is found in the 50 bp upstream the TSS.

(B) Venn diagram for promoters found with an aSD sequence that is either a perfect match or at most 1 bp away. Green and purple circles, promoters that possess either a perfect aSD match or one off by 1 bp; red circle, the space of all putative *E. coli*  $\sigma^{54}$  promoters; square, the space of all putative *E. coli* promoters.

(C) Distribution of percentage proximal occurrences (%6-mer:SD) within 300 bp separation (see Supplemental Experimental Procedures) of CT-rich to GA-rich (aSD:SD) pairs (orange) as compared with the percentage proximal occurrences of random to GA-rich (R:SD) hexamer pairs for *E. coli*.

(D) Scatterplot for 591 mesophile and psychrophile bacterial genomes, where each genome is represented by the mean value for the aSD:SD (x axis) and R:SD (y axis) percentage proximal occurrence distributions. Dashed line ( $x = y$ ) corresponds to the null model, assuming that mean aSD:SD should equal to mean R:SD. Vertical dashed-dot line ( $x = 7.5\%$  occurrences at 300 bp) corresponds to the null expected value.

Associated with Figures S6 and S7 and Table S7.

promoters in the native *E. coli* genomic context. To this end, we counted proximal occurrences of aSD:SD sequences around  $\sigma^{54}$  promoters and compared the numbers to such occurrences around all other promoters (i.e.,  $\sigma^{70}$ -like). We first identified  $n = 8,174$  putative promoters and their respective TSSs (Salgado et al., 2013),  $B = 91$  of which are annotated as  $\sigma^{54}$  promoters. We then identified the first putative SD (purine-rich) hexamer downstream of the TSS, and the best matching (Hamming) aSD (pyrimidine-rich) hexamer upstream of the TSS, both up to 50 bp (Figure 6A). In a total of  $B = 1,383$  of the  $n = 8,174$  promoters, we found a potential near-perfect (Hamming distance  $\leq 1$ ) proximal aSD:SD pair (see Figure 6B),  $B = 25$  of them in the  $n = 91$   $\sigma^{54}$  promoters. Under a hypergeometric model, this yields an enrichment at p value  $< 0.008$  for the occurrence of these sequences within  $\sigma^{54}$  promoters.

To put this enrichment of aSD:SD occurrences near  $\sigma^{54}$  promoters in an even broader bacterial context and also to check for potential promoter sequence bias that may affect this finding, we carried out whole-genome analysis on multiple bacterial genomes. We hypothesized that, if pyrimidine-rich hexamers encode a regulatory function, then the prevalence of aSD:SD will be reduced as compared to random:SD occurrences, where random is some random hexamer. In order to test this hypothesis, we developed a specialized algorithm to quantitatively assess the prevalence of aSD and random (R) motifs upstream of putative SD hexamers at the genomic level (see Supplemental Experimental Procedures for details and Figure S6 for detailed description of algorithm). In brief, we first constructed a list of all hexamers (4,096), and for each one of them, we calculated the free energy value ( $\Delta G$ ) for their hybridization with the 16S rRNA

of *E. coli* (as in Li et al., 2012). We ranked this list of hexamers from low to high free energy value, which corresponds to higher to lower probability for hybridization with the 16S rRNA. The top 20 hexamers were defined as the best 20 putative SD hexamers. Next, we computed a set of 20 pyrimidine-rich hexamers with the highest PSSM scores according to our aSD E5mer motif (see Figure 4A, top right, for logo). Finally, we computed a set of 20 randomized hexamers, which did not score highly for either motif (see Table S7 for all hexamers). To compare the occurrence frequency of proximal aSD:SD to that of R:SD (random:SD), we first identified all of the hits in the genome for a particular SD hexamer. An occurrence is then defined as the location of the first appearance of a particular aSD hexamer upstream of one of the hits. Next, we computed all occurrences in the genome for a particular aSD:SD and R:SD pair in the range  $10 < d < 10,000$ , where  $d$  is the distance in base pairs upstream of the putative purine-rich hexamer. Finally, we computed the percentage of proximal pairs (%aSD:SD—defined as the percentage of occurrences that fall in the range  $10 < d < 300$ ) for each pair of aSD:SD and R:SD hexamers. This process yielded a distribution of 400 percentage numbers for both randomized (R) and CT-rich (aSD) to GA-rich (SD) segments. As an example, we plot the distribution of percentage of proximal pairs obtained for *E. coli* in Figure 6C. The plot shows that, in *E. coli*, aSD:SD proximal occurrences (orange) are significantly depleted as compared to the R:SD (blue; Wilcoxon  $p$  value  $< 3.83 \times 10^{-12}$ ).

We applied our aSD:SD prevalence algorithm to 591 psychrophile and mesophile genomes (i.e., bacteria that thrive in the temperature range of  $-10^{\circ}\text{C}$ – $40^{\circ}\text{C}$ ), obtaining two percentage distributions (as in Figure 6C) for each genome. We then computed the mean for each distribution, allowing us to represent each genome by these two numbers. To quantitatively assess whether the depletion of aSD:SD is prevalent in mesophilic genomes, and not just specifically in *E. coli*, we represent each genome as circle in a scatterplot with the  $x$  and  $y$  coordinates set as the mean proximal R:SD and aSD:SD occurrences, respectively. We found that, whereas the value for the mean percentage R:SD for all genomes is scattered tightly around the null expected 7.5%, the aSD:SD genomic scatter varies widely over a considerably larger range. In addition, assuming no significant regulatory function for the aSD motifs, we should expect to find approximately half the genomes above the dashed line ( $x = y$ ) and half below. Instead, a disproportionately large number of genomes (364 of 591 or  $\sim 62\%$ ) are found with a larger mean R:SD percentage than aSD:SD. Moreover, 342 of 591 (or 57.87%) of the genomes are found with mean aSD:SD  $< 7.5\%$  (left of the vertical dashed-dotted line).

Finally, to provide further support for the above analysis and to eliminate potential biases that may be due to the particular sets of hexamers chosen, we repeated the analysis in Figures 6C and 6D using 50 different sets of randomized hexamers. We also carried out a random compared to random control (R:SD versus R:SD) by replacing the aSD set with another set of randomized hexamers. For each set of randomized hexamers, we computed the “depletion percentage” defined as the proportion of genomes for which mean aSD:SD  $<$  mean R:SD (e.g., 62% for the set shown in Figure 6D). We find (see Supplemental Experimental Procedures and Figure S7) that, in the aSD:SD versus

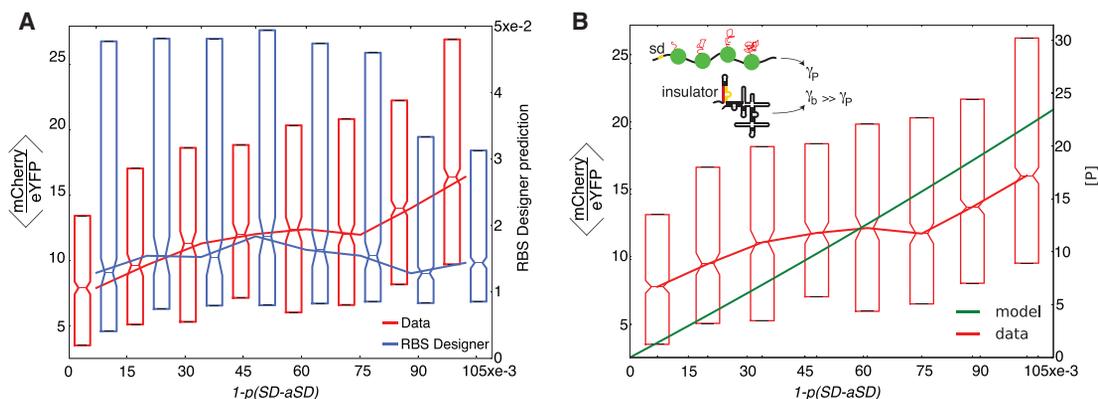
R:SD comparison, 43 of the 50 randomized sets yield a depletion percentage  $>50\%$ , whereas, for the R:SD versus R:SD comparison, only 24 of the 50 randomized sets yield a depletion percentage  $>50\%$ , matching the null model’s expectations. In summary, whereas merely a statistical finding that may also reflect other evolutionary constraints, the enrichment of aSD:SD pairs in the vicinity  $\sigma^{54}$  promoters as compared with other promoter types, together with the overall depletion of aSD:SD proximal occurrences in bacterial genomes, is consistent with a natural role for the insulating mechanism we observed in the synthetic systems, one that is mostly utilized within the context of  $\sigma^{54}$  promoters as defined by the annotation used herein.

## DISCUSSION

We used a synthetic OL-based approach, following protocols introduced in recent years (Kinney et al., 2010; Sharon et al., 2012), to uncover and study a context-dependent phenomenon of translational insulation in  $\sigma^{54}$  promoters.

Insulation was found to be dependent on the prevalence of short (3- to 5-nt) CT-rich sequences, which were distributed at various positions within the 50-bp variants encoding core  $\sigma^{54}$ , core  $\sigma^{54}$ -like, and flanking sequences. Strength of insulation was statistically associated with the number and proximity of CT-rich segments to the putative  $\sigma^{54}$  promoter TSS. Because these CT-rich segments cannot be transcribed by the  $\sigma^{54}$  promoter to which they belong, insulation seems likely to be associated with transcriptional readthrough, when the polymerase originates from another locus upstream. Together, the segments create a cumulative effect that possibly triggers a collapse of the RNA molecule into a “branched phase” via aSD:SD interaction (Schwab and Bruinsma, 2009) due to the lack of translocating ribosomes on the mRNA molecule, leading to rapid degradation of the insulated mRNA molecule (Richards et al., 2012). This hypothesized mechanism is supported by modeling evidence as well as by comparing variants that mostly differ by the presence of the putative aSD element. In this work, this comparison includes a high coverage mutational analysis for glnKp, a sparser analysis for other variants, and an insertion experiment that converted the non-insulating glnHp promoter into an insulating one. However, there are likely other “context”-related regulatory mechanisms encoded into these variants that may contribute to the wide distribution observed for the OL gene expression profile. Those remain obscure due to the unavoidable partial coverage provided by our library.

Previous studies have implicated aSD:SD interactions and other secondary structure formations involving the RBS in a variety of regulatory phenomena. These include riboswitches, upregulation via RNA-binding protein interactions with RNA (Babitzke et al., 2009; Winkler and Breaker, 2005), modulating expression levels with partially stable structures (de Smit and van Duin, 1990; Schwartz et al., 1981), and inducing translation via S1 interaction (non-SD initiation; Komarova et al., 2002). Other studies (see Nakamoto, 2006 for review and references therein) have suggested that translation initiation is inhibited when the AUG is sequestered in a double-stranded structure and that this can be avoided by either having a non-structured 5′ UTR region (Scharff et al., 2011) or via an accessible SD



**Figure 7. Comparison of RBS Designer Tool and Insulation Model for Predicting Gene Expression on Our Library**

(A) Binned mean fluorescence ratio of the library variants plotted in boxplot form according to the probability of the RBS to be unbound (red) together with the RBS Designer tool predicted mean expression levels for each variant sequence (blue). The comparison shows that the RBS Designer predicted RBS availability does not reflect the measured mean fluorescence ratio and as a result is a poor predictor for our data.

(B) The library mean fluorescence ratio plotted in boxplot form as a function of the probability of the RBS to be unbound (red). The prediction from the degradation model is plotted for comparison (green). (Inset) A scheme of the degradation model is shown (see [Supplemental Experimental Procedures](#)). (Top) The translated pearly phase with low degradation rate is shown. (Bottom) The non-translated branched phase with high degradation rate is shown. Data are represented as boxplots set with quartiles.

sequence. Therefore, if the SD sequence is also sequestered, then the likelihood of AUG inaccessibility to translation initiation will be high. The insulation phenomenon described in the present work and its relative prevalence in annotated  $\sigma^{54}$  promoters is another regulatory manifestation of the effects of RNA secondary structure and of the effects of aSD:SD interactions in bacteria. In particular, this is the first reported example in bacteria of a sequence-specific phenomenon whose role is to “ensure” against downstream gene expression if transcribed and is otherwise an integral part of a functional promoter.

Given the potency of this regulatory effect and its statistical depletion in most bacterial genomes, why is there an enrichment in the sub-class of  $\sigma^{54}$  promoters? Unlike most bacterial sigma factors that are members of the  $\sigma^{70}$  family and encode a niche response,  $\sigma^{54}$  promoters are unique. The polymerase is unable to initiate transcription by itself but rather absolutely requires the energy of ATP hydrolysis via the binding of an associated bacterial enhancer binding protein. As a result,  $\sigma^{54}$  promoters do not suffer from promoter leakage and are usually fully repressed when the bacterial enhancer binding protein is absent. The encoding of the aSD sequences in the non-transcribed portion of these promoters generates another level of security against errant transcriptional events, ensuring that  $\sigma^{54}$ -regulated genes are not produced when there is accidental transcriptional readthrough from an upstream promoter. Recent analysis has revealed a common functional theme across multiple bacterial species, which can provide an explanation for the additional measures used in these promoters against leaky or errant transcription. The analysis of [Francke et al. \(2011\)](#) has shown that  $\sigma^{54}$  promoters predominantly regulate genes that control the transport and biosynthesis of the molecules that constitute the bacterial exterior, thus affecting cell structure, developmental phase, and interaction potential with the environment. For instance, in *M. xanthus*, there are many  $\sigma^{54}$  promoters that

have been associated with fruiting body development ([Jakobsen et al., 2004](#)). Thus, it is possible that the presence of CT-rich insulators encoded within some  $\sigma^{54}$  promoters may be attributed to preventing metabolic and developmental consequences, which only in rare circumstances are needed for survival.

Whereas our results provide additional support to the observations that short CT-rich sequences in the vicinity of an RBS can affect expression, they are by no means the only context-dependent effect, which can be encoded in an accidental fashion, in the design of synthetic circuits. To avoid other such context-dependent regulatory phenomena, we believe that it can be useful to apply OL-based approaches (reviewed in [Peterman and Levine, 2016](#)) as exemplified here. This experimental approach constitutes a reliable systematic methodology in synthetic biology to support gene circuit design and construction. It is a high-throughput controlled mutagenesis that can serve as “debugging” analysis. Sort-seq approaches can be utilized to identify, characterize, and filter out unknown context-related effects, facilitating a desirable outcome without resorting to an iterative design/characterization process. Studies based on sort-seq can therefore lead to improved performance of design tools, such as discussed in comparing RBS designer to a refined model in [Figure 7](#) ([Na and Lee, 2010](#); see [Supplemental Experimental Procedures](#) for detailed discussion). In summary, our findings advance our understanding of translational regulation in bacteria. From a practical perspective, our findings provide potentially useful refinements of guidelines to be used in synthetic biology designs.

## EXPERIMENTAL PROCEDURES

### OL Design

Each variant included a unique 50-bp sequence, placed 120 bp downstream of the pLac/Ara promoter and adjacent to an mCherry RBS, thus encoding a variable 5' UTR region with an interchangeable 50-bp region positioned

at +50 from the TSS. The OL was designed to test both additional  $\sigma^{54}$  and putative  $\sigma^{54}$  promoters, from *E. coli* as well as other bacteria, for the silencing effect. In addition, we designed the OL to conduct a broader study of the contextual regulatory effects induced by a downstream promoter on an active upstream promoter positioned nearby in either a sense or anti-sense orientation. To do so, our library is composed of four sub-classes: a no  $\sigma^{54}$  promoter set taken from *E. coli* and *V. cholera* genomic regions with low binding score (score < 0.5), which was designed to form a non-coding positive control (130 variants); a set of 125 annotated *E. coli*  $\sigma^{70}$  promoters (devoid of any annotated transcription factor [TF] binding sites) taken from the EcoCyc database (Keseler et al., 2013); a set of 275 annotated core  $\sigma^{54}$  promoters from multiple strains with their flanking sequences taken from both EcoCyc (Keseler et al., 2013; 98 variants) and (Barrios et al., 1999; 177 variants); a set of 134 mutant variants for the glnKp sequence in both the core elements and flanking sequences; and 5,715 variants with  $\sigma^{54}$ -like core regions mined from the *E. coli* and *V. cholera* genomes with a match score >0.765 as compared with the  $\sigma^{54}$  consensus sequence (score = 1; see  $\sigma^{54}$  Consensus Binding Site Scoring in the Supplemental Experimental Procedures). Finally, all variants were encoded so they would appear in both sense and anti-sense orientations with respect to the pLac/Ara driver promoter. An OL of all variants was synthesized by Twist Bioscience. The library design contained 12,762 unique sequences, each of length 145–148 bp. Each oligo contained the following parts: 5' primer binding sequence; NdeI restriction site; specific 10-bp barcode; variable tested sequence; XmaI restriction site; and 3' primer binding sequence. The barcode and the promoter sequences were separated by a spacer segment of 23 bp (cassette design is shown in Figure 1). We received 12,758 sequences of the 12,762 ordered.

#### OL Cloning

Oligo library cloning was based on the cloning protocol developed by the Segal group. Briefly, the 12,758-variant single-stranded DNA (ssDNA) library from Twist BioScience was amplified in a 96-well plate using PCR, purified, and merged into one tube. Following purification, double-stranded DNA (dsDNA) was cut using XmaI and NdeI and dsDNA with the desired length was gel separated and cleaned. Resulting DNA fragments were ligated to the target plasmid using a 1:1 ratio. Ligated plasmids were transformed to *E. coli* cells (Lucigen) and plated on 28 large agar plates (with antibiotics) in order to conserve library complexity. Approximately ten million colonies were scraped and transferred to an Erlenmeyer for growth.

#### OL Transcriptional Silencing Assay

The oligo library (OL) silencing assay for the transformed OL was developed based on Sharon et al. (2012) and was carried out as follows.

#### Culture Growth

Library-containing bacteria were grown with fresh lysogeny broth (LB) and antibiotic (Kan). Cells were grown to mid-log phase (OD<sub>600</sub> of ~0.6) as measured by a spectrophotometer (Novaspec III; Amersham Biosciences) followed by resuspension with BioAssay (BA) buffer and the appropriate antibiotic (Kan). Culture was grown in BA for 3 hr prior to sorting by FACS Aria cell sorter (Becton Dickinson).

#### Fluorescence-Activated Cell Sorting

Sorting was done at a flow rate of ~20,000 cells per second. Cells were sorted into 16 bins (500,000 cells per bin) according to the mCherry to eYFP ratio in two groups: (1) bins 1–8: high resolution on low ratio bins (30% scale) and (2) bins 9–16: full-resolution bins (3% scale).

#### Sequencing Preparation

Sorted cells were grown overnight in 5 mL LB and appropriate antibiotic (Kan). In the next morning, cells were lysed (Triton X-100 0.1% in 1XTE: 15  $\mu$ L; culture: 5  $\mu$ L; 99°C for 5 min and 30°C for 5 min) and the DNA from each bin was subjected to PCR with a different 5' primer containing a specific bin barcode. PCR products were verified in an electrophoresis gel and cleaned using PCR Clean-Up kit (Promega). Equal amounts of DNA (10 ng) from each bin were joined to one 1.5-mL microcentrifuge tube for further analysis.

#### Sequencing

Sample was sequenced on an Illumina HiSeq 2500 Rapid Reagents V2 100 bp paired-end chip. 10% PhiX was added as a control. This resulted in ~140 million reads.

#### NGS Processing

From each read, the bin barcode and the sequence of the strain were extracted using a custom Python script consisting of the following steps: paired-end read merge; read orientation fix; and identification of the constant parts in the read and extracting the variables: bin barcode, sequence barcode, and the variable tested sequence. Finally, each read was mapped to the appropriate combinations of tested sequence and expression bin. This resulted in ~38 million uniquely mapped reads, each containing a perfect match variance sequence and expression bin barcode pair.

#### Inference of per Variant Expression Profile

We first removed all reads mapped to bin number 16 from the analysis to eliminate biases originating from out-of-range fluorescence measurements. Next, we filtered out sequences with low read counts, keeping only those with a total of at least 10 reads across the bins. This left us with a total of ~36 million reads distributed over 10,438 variants. We then generated a single profile by replacing bin 9 with bins 1–8 and redistributing the reads in bin 9 over bins 1–8 according to their relative bin widths. Next, for each sequence, we calculated the fraction of cells in each bin, based on the number of sequence reads from that bin that mapped to that variant (the reads of each bin were first normalized to match the fraction of the bin in the entire population). This procedure resulted in expression profiles over 14 bins for 10,432 variants (see Data S1).

#### Inference of per Variant Mean Expression Level

For each variant, we defined the mean expression ratio as the weighted average of the ratios at the geometric centers of the bin, where the weight of each bin is the fraction of the reads from that variant in that bin.

#### Synthetic Gene Circuit Construction and Associated Assays

See Supplemental Experimental Procedures.

#### OL Technical Assessment

See Supplemental Experimental Procedures.

#### Rate Equation Model for Insulation

See Supplemental Experimental Procedures.

#### Bioinformatic Analysis Methods

See Supplemental Experimental Procedures.

#### DATA AND SOFTWARE AVAILABILITY

The accession number for the sequencing data reported in this paper is NCBI: PRJNA407221.

#### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, seven tables, and two data files and can be found with this article online at <https://doi.org/10.1016/j.celrep.2017.09.063>.

#### AUTHOR CONTRIBUTIONS

L.L. designed and carried out the experiments for both the initial  $\sigma^{54}$  and OL experiment. L.A. designed and carried out data analysis for the OL library results and modeled the data. O.S. designed and performed the bioinformatics genomic context analysis. R.C., S.O., and S.G. designed and carried out the  $\Delta$ rhoN and GST experiments. M.B.-M. carried out the western blot experiment. O.A. assisted with some of the experiments. R.A. and Z.Y. supervised the study. R.A., Z.Y., S.G., L.L., L.A., and O.S. wrote the manuscript.

## ACKNOWLEDGMENTS

This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 664918—MRG-GRammar, the Israel Science Foundation through grant no. 1677/12, the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (grant no. 152/11), and Marie Curie Reintegration grant no. PCIG11-GA-2012-321675. L.A. is supported by the Adams Fellowships Program of the Israel 6 Academy of Sciences and Humanities. The authors would like to acknowledge the Technion's LS&E staff (Tal Katz-Ezov, Efrat Barak, and Anastasia Diviatis) for help with sequencing and FACS and Adina Weinberger, Yitzhak Pilpel, Roy Kishony, Shay Ben-Elazar, Noa Katz, Beate Kaufmann, Yaroslav Pollak, and Inbal Vaknin for useful discussions.

Received: March 2, 2017

Revised: August 30, 2017

Accepted: September 18, 2017

Published: October 17, 2017

## REFERENCES

- Amit, R., Garcia, H.G., Phillips, R., and Fraser, S.E. (2011). Building enhancers from the ground up: a synthetic biology approach. *Cell* 146, 105–118.
- Atkinson, M.R., Blauwkamp, T.A., Bondarenko, V., Studitsky, V., and Ninfa, A.J. (2002). Activation of the *glnA*, *glnK*, and *nac* promoters as *Escherichia coli* undergoes the transition from nitrogen excess growth to nitrogen starvation. *J. Bacteriol.* 184, 5358–5363.
- Babitzke, P., Baker, C.S., and Romeo, T. (2009). Regulation of translation initiation by RNA binding proteins. *Annu. Rev. Microbiol.* 63, 27–44.
- Barrios, H., Valderrama, B., and Morett, E. (1999). Compilation and analysis of  $\sigma(54)$ -dependent promoter sequences. *Nucleic Acids Res.* 27, 4305–4313.
- Bonocora, R.P., Smith, C., Lapierre, P., and Wade, J.T. (2015). Genome-scale mapping of *Escherichia coli*  $\sigma 54$  reveals widespread, conserved intragenic binding. *PLoS Genet.* 11, e1005552.
- Calin-Jageman, I., and Nicholson, A.W. (2003). Mutational analysis of an RNA internal loop as a reactivity epitope for *Escherichia coli* ribonuclease III substrates. *Biochemistry* 42, 5025–5034.
- Claverie-Martin, F., and Magasanik, B. (1991). Role of integration host factor in the regulation of the *glnHp2* promoter of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 88, 1631–1635.
- de Smit, M.H., and van Duijn, J. (1990). Control of prokaryotic translational initiation by mRNA secondary structure. *Prog. Nucleic Acid Res. Mol. Biol.* 38, 1–35.
- Deana, A., Celesnik, H., and Belasco, J.G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature* 457, 355–358.
- Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* 3, e39.
- Epshtein, V., Toulmé, F., Rahmouni, A.R., Borukhov, S., and Nudler, E. (2003). Transcription through the roadblocks: the role of RNA polymerase cooperation. *EMBO J.* 22, 4719–4727.
- Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., and Levine, M.S. (2015). Suboptimization of developmental enhancers. *Science* 350, 325–328.
- Feng, J., Goss, T.J., Bender, R.A., and Ninfa, A.J. (1995). Repression of the *Klebsiella aerogenes* *nac* promoter. *J. Bacteriol.* 177, 5535–5538.
- Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* 12, 385.
- Gu, W., Zhou, T., and Wilke, C.O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* 6, e1000664.
- Hao, N., Palmer, A.C., Ahlgren-Berg, A., Shearwin, K.E., and Dodd, I.B. (2016). The role of repressor kinetics in relief of transcriptional interference between convergent promoters. *Nucleic Acids Res.* 44, 6625–6638.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188.
- Hoover, T.R., Santero, E., Porter, S., and Kustu, S. (1990). The integration host factor stimulates interaction of RNA polymerase with NIFA, the transcriptional activator for nitrogen fixation operons. *Cell* 63, 11–22.
- Jakobsen, J.S., Jelsbak, L., Jelsbak, L., Welch, R.D., Cummings, C., Goldman, B., Stark, E., Slater, S., and Kaiser, D. (2004).  $\sigma 54$  enhancer binding proteins and *Myxococcus xanthus* fruiting body development. *J. Bacteriol.* 186, 4361–4368.
- Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., et al. (2013). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 41, D605–D612.
- Kinney, J.B., Murugan, A., Callan, C.G., Jr., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. USA* 107, 9158–9163.
- Kiupakis, A.K., and Reitzer, L. (2002). ArgR-independent induction and ArgR-dependent superinduction of the *astCADBE* operon in *Escherichia coli*. *J. Bacteriol.* 184, 2940–2950.
- Komarova, A.V., Tchufistova, L.S., Supina, E.V., and Boni, I.V. (2002). Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *RNA* 8, 1137–1147.
- Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* 22, 911–917.
- Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324, 255–258.
- Leibovich, L., and Yakhini, Z. (2012). Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res.* 40, 5832–5847.
- Leibovich, L., Paz, I., Yakhini, Z., and Mandel-Gutfreund, Y. (2013). DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res.* 41, W174–W179.
- Li, G.-W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 538–541.
- Mackie, G.A. (1998). Ribonuclease E is a 5'-end-dependent endonuclease. *Nature* 395, 720–723.
- Na, D., and Lee, D. (2010). RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics* 26, 2633–2634.
- Nakamoto, T. (2006). A unified view of the initiation of protein synthesis. *Biochem. Biophys. Res. Commun.* 341, 675–678.
- Peterman, N., and Levine, E. (2016). Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* 17, 206.
- Reitzer, L., and Schneider, B.L. (2001). Metabolic context and possible physiological themes of sigma(54)-dependent genes in *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* 65, 422–444.
- Richards, J., Luciano, D.J., and Belasco, J.G. (2012). Influence of translation on RppH-dependent mRNA degradation in *Escherichia coli*. *Mol. Microbiol.* 86, 1063–1072.
- Robertson, H.D. (1982). *Escherichia coli* ribonuclease III cleavage sites. *Cell* 30, 669–672.
- Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., et al. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 41, D203–D213.

- Scharff, L.B., Childs, L., Walther, D., and Bock, R. (2011). Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet.* 7, e1002155.
- Schwab, D., and Bruinsma, R.F. (2009). Flory theory of the folding of designed RNA molecules. *J. Phys. Chem. B* 113, 3880–3893.
- Schwartz, M., Roa, M., and Débarbouillé, M. (1981). Mutations that affect lamB gene expression at a posttranscriptional level. *Proc. Natl. Acad. Sci. USA* 78, 2937–2941.
- Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* 30, 521–530.
- Winkler, W.C., and Breaker, R.R. (2005). Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* 59, 487–517.
- Yokobayashi, Y., Weiss, R., and Arnold, F.H. (2002). Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci. USA* 99, 16587–16591.