Short CT-rich motifs encoded within σ^{54} promoters insulate downstream genes from transcriptional read-through

Authors: Lior Levy[1†], Leon Anavy[2†], Oz Solomon[1,2], Roni Cohen[1], Shilo Ohayon[1], Orna Atar[1], Sarah Goldberg[1], Zohar Yakhini[2,3], Roee Amit[1,4*]

Affiliation:

[1]Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa, Israel 32000.

[2]Department of Computer Science, Technion - Israel Institute of Technology, Haifa, Israel 32000.

[3]School of Computer Science, Interdisciplinary Center, Herzeliya, Israel.

[4]Russell Berrie Nanotechnology Institute, Technion - Israel Institute of Technology, Haifa 32000.

*Correspondence to: roeeamit@technion.ac.il

†These authors contributed equally.

Abstract: We use an oligonucleotide library of over 10000 variants together with a synthetic biology approach to identify an insulator sequence encoded within a subset of σ^{54} promoters. Insulation manifests itself as silencing of expression of a downstream gene during transcriptional read through. Insulation 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 using modeling, mutations to the CT-rich motif, and gene expression measurements on multiple sequence variants, that insulation is likely caused by binding of the RBS region to the upstream CT-rich motifs. We show that the strength of the insulation effect depends on the location and number of CU-rich motifs encoded within the promoters. Finally, we show that in *E.coli* these insulator sequences are preferentially encoded within σ^{54} promoters as compared to other promoter types, indicating that there is an important regulatory role for these sequences in natural contexts. Our findings have important implications for understanding SNP/INDEL mutations in regulatory regions and add to existing guidelines for designing synthetic promoters in bacterial systems.

Deconstructing genomes to their basic parts and then using those parts to construct *de novo* gene regulatory architectures are amongst the hallmarks 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 an unknown regulatory role. Often, context effects are due to short-range sequence-based interactions with nearby elements.[1] Such interactions might endow some secondary regulatory effect that is overlooked by standard analysis methods or is masked by a stronger regulatory effect in the native setting.[2] “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 RBS via hybridization by an anti-Shine Dalgarno (aSD) sequence have been shown to strongly inhibit or modulate the initiation of translation.[3–6] In bacteria, context effects have 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 towards unstable secondary structure and is generally GC-poor as a result.[7–9] Other intragenic regulatory phenomena that have been recently reported in bacteria involve inactive σ^{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 RNAP complexes.[10,11] Alternatively, dynamical processes such as transcriptional interference by an incoming RNA polymerase, or transcriptional read-through from an upstream locus can also alter gene expression in a way that is not encoded in the individual parts.[12–14] Consequently, it has become evident that due to such diverse effects, a more systematic understanding of context-related effects and their origin is needed for the design of reliable synthetic biology modalities, as well as for better understanding natural regulatory mechanisms.

One approach to avoiding unwanted context effects in non-coding regions is to employ directed evolution. It has been suggested that directed evolution screens should be applied to any synthetic biological circuit to generate the best-performing DNA sequence for a given application.[11] However, this approach avoids unwanted context effects without identifying either the problematic contexts or the regulatory mechanisms that they encode. Directed evolution screens also do not fit every scenario and are thus often impractical in the design of many synthetic biological circuits.

Synthetic oligo 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.[15,16] This approach can be used to investigate secondary context-related phenomena in non-coding regulatory elements. In this work we use a synthetic biology approach to identify a transcriptional insulation phenomenon, encoded by a putative aSD sequence within the context of the *glnK* σ^{54} promoter (glnKp). This insulation function is not "turned" on by the σ^{54} promoter itself, but rather via transcriptional read-through. Namely, by an RNA-polymerase arriving from an upstream locus. To further study this phenomenon, we use an OL to both generate hundreds of mutated variants of the original glnKp context, and search for additional insulator context in other annotated σ^{54} promoters from multiple bacterial species. Using the OL we annotated the insulator sequence, and found that it is encoded in a consensus aSD sequence that is typically located up to 25 nucleotides upstream from the RBS. Finally, we analyzed the sequence of the annotated *E. coli* promoters to show that our insulator sequences are enriched in σ^{54} promoters, implying that an insulator's presence was likely conserved for a particular regulatory role.

Results

The σ^{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 σ^{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, possibly facilitating DNA looping which often harbors additional transcription factor binding sites.[15–17] In our study of enhancer components, each synthetic circuit consisted of a UAS element and a σ^{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 Fig. 1A, Supp. Note 1, and Supp. Fig. 1-2). We chose five *E. coli* σ^{54} promoters of varying known[18–23] strengths (glnHp, astCp, glnAp2, glnKp, 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 σ^{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, the NtrC activator expression was optionally induced by anhydrous tetracyline using a separate positive-feedback synthetic enhancer circuit .[24]

We plot the mean fluorescence expression-level data in steady state together with their variation for the synthetic enhancers as a heat map in Fig. 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-σ^{54}-promoter control. The promoters which were previously reported to be "weak" (glnHp and astCp) and naturally bound by either IHF or ArgR[20,25,26] were indeed found to generate lower levels of expression as compared with glnAp2, nacp, and glnKp (p-value<0.05,10^{-3} respectively for paired t-test). 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 σ^{70} promoter induces expression in the no-promoter control, as expected.

To characterize the activity of the σ^{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 Fig. 1B. Without NtrC, σ^{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 correction for multiple testing). This is due to its dual role as a σ^{70} promoter in addition to being a σ^{54} UAS. However, glnAp1 showed a detectible fluorescence response for only four of the five promoters. The σ^{54} promoter glnKp manifested a different behavior. Namely, the glnAp1 UAS did not generate detectible expression as compared with each of the other promoters (t-tests, p-value<0.01 for all promoters). Thus, there seems to be an inhibitory mechanism embedded within the σ^{54} promoter glnKp.

We initially reasoned that the inhibitory phenomenon might be explained by unusually tight binding of the σ^{54}-RNAP complex to the glnKp core region, leading to the formation of a physical "road-block", which interferes with any upstream transcribing RNAP holoenzymes. To check this hypothesis, we constructed another gene circuit in which a pLac/Ara (σ^{70}) promoter was placed upstream of the σ^{54} glnKp instead of the glnAp1 UAS. In Fig. 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 the glnKp (yellow). However, when the circuit was placed in a Δ*rpoN* knockout strain (*rpoN* encodes the σ^{54} RNAP subunit), the same reduction in fluorescence was observed (orange). Moreover, in Fig. 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 (Fig. 1E). Consequently, in the context of our construct, the glnKp sequence not only encodes a σ^{54} promoter, but also some inhibitory function that is active when this sequence is placed downstream from an active σ^{70} promoter and upstream to the mCherry start codon.

glnK promoter encodes an insulator sequence

We hypothesized that the silencing phenomenon occurs at the post-transcriptional level. To provide initial support for this assertion, we plot in Fig. 2A two RNA 2D structure models [27] for the constructs with the no σ^{54} promoter (top) and with the glnKp (bottom). The structure models suggest that while the RBS for the no σ^{54} promoter remains single stranded, the one for the glnKp is sequestered in a double-stranded hairpin structure. Previous, studies[30–35] have shown that a folded RNA state frequently triggers increased degradation of the untranslated RNA by RnaseE. Since we also observed reduced RNA levels in the silenced strain (Fig. 1D), we wanted to rule out that the possibility of the silencing effect being a degradation artefact of our original circuit design of two closely positioned promoters.

In order to study a more natural-mimicking configuration involving the glnKp, 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 read-through of the downstream gene from the upstream promoter. We reasoned that a translated gene placed upstream of an aSD sequence would protect the entire mRNA from the pyrophosphotation of the 5’end by RppH.[31] This, in turn, would inhibit the RnaseE degradation pathway, leading to a partial rescue of the mCherry silencing effect. The quantitative PCR (qPCR) results are shown in

Fig. 2B. We show results for four strains: glnKp variant without *GST* (left bar), *GST*+glnKp+mCherry (second from the left), RBS+*GST*+glnKp+mCherry (second from the right), and a non-silenced strain (right). In Fig. 2B it can be seen that the mRNA level for the non-translated *GST* is identical to the one measured for the glnKp variant, and can thus be considered silenced. However, when the *GST* is translated, the mRNA levels rise considerably by a factor of ~3, representing approximately 50% recovery as compared with the non-silenced strain. Next, we measured the recovery in expression level (e.l.r) ratio of mCherry to eYFP. To do so, we added a non-insulated circuit expressing eYFP (Fig. 2C-top) to properly quantify the insulation effect as a function of e.l.r, and observe only a small recovery in e.l.r (Fig. 2C). While this recovery is consistent with the partial recovery in mRNA levels (x2 as compared with x3), the data shows that mCherry fluorescence for the construct containing a translated GST gene is still smaller by about x10 from the construct that does not encode the insulating sequence. Thus, the RBS sequestering CU-rich segment located ~20 bp upstream of the SD sequence (marked in red in Fig. 2A) can be said to encode an insulator sequence, which functions via a Shine-Dalgarno to anti-Shine Dalgarno (SD-aSD) interaction to silence expression. This sequence, encoded in the non-transcribed portion of the glnKp, insulates the downstream σ^{54}-regulated gene from translation when the template mRNA is erroneously transcribed via transcriptional read-through based events.

Oligo library (OL) analysis of glnKp mutants

To further explore and characterize the insulating sequence and mechanism encoded by glnKp, and to check for its prevalence in other bacterial genomes, we constructed an oligo library (OL) of 12758 150-bp variants (Fig. 3A). The OL was synthesized by Twist Bioscience (for technical characteristics see Supp. Note 2 and Supp. Fig. 3-7) and inserted into the synthetic enhancer backbone following the method introduced by Sharon et al.[36] The OL was designed to characterize 134 glnKp mutational variants. In addition, the OL was designed to screen both known σ^{54} promoters from various organisms and σ^{54}-like sequences from genomic regions in *E. coli* and *V. cholera* for the insulation based silencing effect, essentially searching for the phenomenon observed in glnKp, in other σ^{54} promoters. 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 an mCherry reporter gene, thus encoding a 5'UTR region with a variable 50 bp region positioned at +50 bp from the pLac/Ara TSS (Fig 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. By combining the OL with fluorescence-activated cell sorting and next-generation sequencing[36], we obtained the distribution of e.l.r for each sequence variant (Fig. 3B). Figure 3C shows the e.l.r profiles for 10438 variants with sufficient total number of sequence counts (n>10, see Materials and Methods for details), revealing a broad range of distributions of expression levels. While 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 (Fig. 3B top and bottom show representative examples), indicating that a combination of regulatory mechanisms may underlie this distribution of expression levels.

In order to localize the segment of the glnKp promoter responsible for insulation, we closely examined 123 mutant variants of this promoter (Fig. 3D and Supp. Note 2). The figure shows that the mutations in the core σ^{54} promoter region and in the distal CT-rich region (left - blue shaded region) did not correlate with increasing expression level ratios, 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 the 7 nucleotide CT-rich region (centered at -4) into a G or an A yielded the largest effect suggesting that the secondary structure depicted in Fig. 2A was likely abolished leading to the increased occupancy of ribosomes and to the subsequent increase in translation rate.

To provide further support for this structural interpretation of our mutational analysis of the glnKp, we combine RNA structure predictions with a model for increased degradation (see Supp. Note 3 for details). In brief, first, we used RNAfold[27] to compute the probability for the RBS to be sequestered in a secondary structure for each variant in our library. We then constructed a degradation model taking into account the probability that the RBS is sequestered. We assumed different degradation rates for RBS sequestered and non-sequestered mRNA structures (see Fig. 3E - inset). Finally, we used realistic constant rates for other kinetic processes (e.g. transcription rate, translation rate, etc.[37]). In Fig. 3E we plot the mean e.l.r as a function of the RNAfold computed probability for RBS to be sequestered in a secondary structure as a result of SD-aSD interaction. We found that it correlates strongly with the likelihood of the RBS to be non-sequestered. In addition, we also computed the predicted expression level ([P] - right y-axis) from our two-rate degradation model as a function of the SD-aSD secondary structure probability. We found (green-line) that our model describes the trends in the median experimental data (red-line) well, providing further support for the regulatory role that the secondary structure plays in generating the insulating effect.

Oligo library (OL) analysis of broader insulation

In order to check whether this insulating phenomenon is isolated to the glnKp, or is a more wide-spread phenomenon within σ^{54} promoters in general, we analyzed the mean e.l.r values for four groups of variants in our library: annotated σ^{54} promoters from various organisms obtained from the list compiled by[49] and EcoCyc database, annotated σ^{70} promoters obtained from EcoCyc, σ^{54}-like sequences, and non-σ^{54}-promoters who were scored >0.9 and <0.5 by our consensus σ^{54}-promoter calculator respectively (see Materials and Methods). Checking the distribution of mean e.l.r within defined classes of variants (Fig. 4A and Supp. Note 2), we observed that the "not σ^{54} promoter" class and the annotated σ^{70} class show higher expression levels compared to the annotated σ^{54} promoters and σ^{54}-like variants. A non-parametric Mann-Whitney U-test shows that the mean e.l.r distribution of each one of the two former classes is significantly higher than those of the latter classes (p-value < 10^{-4} compared to annotated σ^{54} promoters and p-value < 0.01 compared to σ^{54}-like variants).

In order to study the possible sequence determinants of the observed expression differences, we performed a DRIMust k-mer search on our variant library sorted by mean e.l.r. values. DRIMust is a tool designed to identify enriched sequence motifs in a ranked list of sequences.[38–40] Our analysis revealed that a CT-rich consensus motif is enriched in the silenced variants (p-value < 10^{-54}, mHG test). We plot the results in Fig. 4B. 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 Fig. 4B-top-right). We call these enriched 5-mers E5mers. The middle panel of Fig 4B shows the position of each E5mer, marked by a brown line in the corresponding variant. In the right panel, we show the running average of the number of E5mer occurrences over 20 variants showing the correlation between the presence of an E5mer and the mean e.l.r value. Together, the plots show that a high concentration of CT-rich motifs close to the purine-rich sequence that encodes the Shine-Dalgarno motif (SD – 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 σ^{54} promoters, we grouped the sequences into four groups (Fig.

4C): containing no E5mers, E5mers located at a distal position from the σ^{54} promoter 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 Fig. 4D the average number of E5mers per position (with the σ^{54} promoter TSS defined as 0) divided over three regimes: strong insulation (mean e.l.r<20), moderate insulation (20<mean e.l.r <30), and no/weak insulation (mean e.l.r>30). We chose the 25bp as the threshold due to the presence of a conserved sequence in the σ^{54} core promoter region (Fig. 4B-center-top) that is not an enriched region. The data shows (Fig. 4C-D) that there is a clear correlation between the number and position of the E5mers to the mean e.l.r . In particular, an ANOVA (See Supplementary Tables 4 and 5) test 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 not cooperative, but are instead additive. The pattern of insulation with E5mers at distal and proximal locations was detected in 18 annotated and σ^{54}-like promoters from various organisms (see list in SI), and is consistent with the observations for the glnKp, which also manifested this pattern. Overall, we found 60 strongly insulating σ^{54} promoters and σ^{54}-like sequences (e.l.r <20, Fig. 4D-bottom) out of the 388 tested in our OL.

To provide additional support that insulation is an important regulatory feature of a sub-set of promoters in the native *E. coli* genomic context we investigated the statistical difference between proximal occurrences of anti-Shine Dalgarno and Shine Dalgarno (aSD:SD) sequences around σ^{54} promoters to such occurrences around all other promoters (i.e. σ^{70}–like). We analyzed N=8174 promoters [41], B=91 of which are annotated as σ^{54} promoters. In a total of n=1383 of the promoters we found a potential near-perfect proximal aSD-SD pair (See Supplementary Figure 9), b=25 of them in σ^{54} promoters. Under a Hypergeometric model this yields an enrichment at p-value<0.008 (See Sup. Fig. 9). Thus, both our library and this bioinformatics analysis support the interpretation that the aSD:SD insulating mechanism is biologically important and is frequently utilized within the context of σ^{54} promoters, including in their genomic context.

Discussion

We used a synthetic oligonucleotide-library (OL) approach to uncover and study a context-dependent phenomenon of translational insulation in σ^{54} promoters. To do that, we design and constructed a library of over 10000 transcriptional read-through circuits, where the σ^{54}-like variants served as the variable sequence. The insulation phenomenon was characterized for a subset of annotated σ^{54} promoters, and additionally for σ^{54}-like sequences in *E.coli* and *V.cholera*, which may be either unknown promoters or σ^{54} intergenic binding sites. We also carried out massive parallel mutational analysis using the OL on the *E.coli* glnKp promoter to determine the precise sequence element responsible for the insulation phenomenon in glnKp.

Insulation was found to be dependent on the presence of a short (3-7 nucleotide) CT-rich sequence, which is positioned proximally upstream of a SD motif. The insulating CT-rich motif cannot be transcribed by the σ^{54} promoter to which it is attached, as it is typically centered up to 20 bps upstream from the σ^{54} promoter's TSS. Rather, this sequence is transcribed only in cases of transcriptional read-through, when the polymerase originates from another locus upstream. We provide evidence that in the context of our synthetic circuit, these short CT-rich sequences encode a CU-rich anti-Shine-Dalgarno (aSD) sequence that can bind the RBS, and likely trigger a collapse of the RNA molecule into a "branched phase" [42] due to the lack of translocating ribosomes on the mRNA molecule. In addition, literature shows that branched-phase mRNA molecules are degraded at a higher rate as compared to ribosome-bound mRNA molecules.[32] Previous studies have also shown that aSD-SD interactions or secondary structures involving the RBS can regulate gene expression in a variety of ways. This includes riboswitches up-regulation via RNA binding protein interactions with RNA[43,44], modulating expression levels with partially stable structures[3,4], and inducing translation via S1-interaction (non-SD initiation[45]). Other studies (see [28] 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[29], or via an accessible Shine-Dalgarno (SD) sequence. Therefore, if the SD sequence is also sequestered, then the likelihood of AUG inaccessibility to translation initiation will be high. Finally, recent work argued for the presence of a particular group of sequences upstream of the RBS, which play a role in recruiting the small subunit of the ribosome possibly by destabilizing secondary structures.[6] The insulation phenomenon and its prevalence in annotated σ^{54} promoters reported here is another regulatory manifestation of the effects of RNA secondary structure, and in particular the aSD-SD interaction in bacteria.

Given the potency of this regulatory effect, why is there an enrichment in the sub-class of σ^{54} promoters? Unlike most bacterial sigma factors that are members of the of σ^{70} family and encode a niche response, σ^{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, σ^{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 σ^{54} -regulated genes are not produced when there is accidental transcriptional read-through 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.[46] has shown that σ^{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 σ^{54} promoters that have also been associated with fruiting body development.[47] Thus, it is possible that the insulation mechanisms encoded within some σ^{54} promoters may be attributed to preventing metabolic and developmental consequences, which only in rare circumstances are needed for survival.

Finally, our study has consequences for synthetic biology design approaches. In synthetic biology applications, well-characterized parts, which can originate from different organisms, are often tailored together. Our results imply that careful attention also needs to be paid to the flanking or bridging sequences, which are used to stitch the parts together, as they may encode unwanted regulatory effects. While 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. Thus, in order to avoid other such context-dependent regulatory phenomena, we believe that our combined OL-synthetic circuit design approach can be used as a reliable systematic methodology for gene circuit design, construction, and massively-parallel mutagenesis analysis. Such an approach can be utilized to identify, characterize, and filter out other context-related effects, facilitating a positive outcome without resorting to an iterative design/characterization process. Thus, one can view our work as an example for a study that on the one hand advances our understanding of biology and regulation mechanisms in bacteria, while on the other provides additional set of guidelines to be used in synthetic biology designs.

Materials and Methods

Synthetic enhancer construction

Synthetic enhancer cassettes were ordered as dsDNA minigenes from Gen9, Inc. each minigene was ~500 bp long, and contained the following parts: NdeI restriction site, variable UAS, variable σ^{54} promoter, and KpnI restriction site at the 3' end. The UAS and σ^{54} promoter were separated by a looping segment of 70 bp. For sequence details see Supplementary Note 1 and Supplementary Tables 1 and 2.

Insertion of minigene cassettes into the plasmid was done by double digestion of both cassettes and plasmids with NdeI and KpnI, followed by ligation to a backbone plasmid containing an NtrC switch with TetR binding sites[24] and transformation into 3.300LG *E. coli* cells containing an auxiliary plasmid overexpressing TetR. Cloning was verified by Sanger sequencing.

Synthetic enhancer fluorescence measurement

Starters of strains containing the enhancer plasmids were growth in LB medium with regular antibiotics overnight (16 hrs). The next morning, the cultures were diluted 1:100 into fresh LB and antibiotics and grown to OD600 of 0.6. Cells were then pelleted and medium exchanged for BA with antibiotics. Fluorescence was measured after an additional 2 hrs of growth in BA. Measurements of mCherry and eYFP fluorescence were performed on a FACS Aria IIIu (without sorting).

TOP10:Δ*rpoN* strain construction

An *E. coli* TOP10:Δ*rpoN* strain was created in our lab following the protocol described in [48], using Addgene plasmids pCas (#62225) and pTarget:*rpoN* (based on Addgene plasmid #62226, with N20 target sequence 5'CCGTCCTTAAGCGGATCCAA3'), and a linear repair oligo constructed using overlap PCR containing the genomic sequences immediately upstream and downstream of the *rpoN* gene. After curing both plasmids, the genomic deletion was sequence-verified using Sanger sequencing of the *rpoN* genomic region. The lack of *rpoN* transcripts was further verified using qPCR with primers targeting *rpoN*.

RNA extraction and reverse-transcription

Starters of *E. coli* TOP10 or TOP10:Δ*rpoN* containing the relevant constructs on plasmids were grown in LB medium with appropriate antibiotics overnight (16 hr). The next morning, the cultures were diluted 1:100 into fresh LB and antibiotics and grown to OD600 of 0.6. For each isolation, RNA was extracted from 1.5 ml of cell culture using standard protocols.

Briefly, cells were lysed using Max Bacterial Enhancement Reagent followed by Trizol treatment (both from Life Technologies). Phase separation was performed using chloroform. RNA was precipitated from the aqueous phase using isopropanol and ethanol washes, and then resuspended in Rnase-free water. RNA quality was assessed by running 500 ng on 1% agarose gel. After extraction, RNA was subjected to DNase (Ambion/Life Technologies) and then reverse-transcribed using MultiScribe Reverse Transcriptase and random primer mix (Applied Biosystems/Life Technologies). RNA was isolated from 3 individual colonies for each construct.

qPCR measurements

Primer pairs for mCherry, eYFP and GST genes, and normalizing gene *idnT*, were chosen using the Primer Express software, and BLASTed (NCBI) with respect to the *E. coli* K-12 substr. DH10B (taxid:316385) genome (which is similar to TOP10) to avoid off-target amplicons. qPCR was carried out on a QuantStudio 12K Flex (Applied Biosystems/Life Technologies) machine using SYBR-Green. 3 technical replicates were measured for each of the 3 biological replicates. A C_T threshold of 0.2 Was chosen for all genes.

σ^{54} consensus binding site scoring

The consensus probability matrix for σ^{54} binding (Appendix 1) was based on the compilation of 186 σ^{54} promoters (Table 3 in Barrios et al[49]). The genomes of *E.coli* and *V.cholera* were scanned using a Matlab script that assigns a σ^{54} probability score to all possible 16 bp-long sequences, based on similarity to the consensus site. In brief, each base in the 16 bp sequence is given a value of 0-1 according to SI-Table 2. The values of all 16 bases are summed, and the total is normalized by first subtracting the lowest possible total (1.679) and then dividing by the difference between the highest possible total (11.2) minus the lowest possible total (1.679), resulting in a final score in the range [0,1] (shown in the equation below). Genomic sequences with scores in the range [0.765, 1] were chosen as candidates for σ^{54} binding. Genomic sequences with scores in the range [0, 0.5] were chosen as candidates that were highly unlikely to bind σ^{54} and were taken as the "no promoter" group.

$$SCORE = \frac{\sum(matrix\,values) - 1.679}{11.2 - 1.679}$$

High-throughput oligo library (OL) expression assay

OL design

Each variant included a unique 50 bp sequence, placed 120 bp downstream from 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 σ^{54} and putative σ^{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-promoter set designed to form a non-coding positive control (130 variants), a set of 125 natural *E. coli* σ^{70} promoters (devoid of any annotated TF binding sites), a set of 228 annotated core σ^{54} promoters from multiple strains with their flanking sequences [49], a set of 134 mutant variants for the glnKp sequence in both the core elements and flanking sequences, and 5715 variants with σ^{54}-like core regions mined from the *E. coli* and *V. cholera* genomes with a match score > 0.765 as compared with the σ^{54} consensus sequence (score =1, see Supplementary Note 3). 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 contained 12758 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 Fig. 1).

OL technical assessment

See Supplementary Note 2.

OL cloning

Oligo library cloning was based on the cloning protocol developed by the Segal group[36] (see Supplementary Note 2 for additional details). Briefly, the 12758-variant ssDNA library from Twist BioScience was amplified in a 96-well plate using PCR, purified, and merged into one tube. Following purification, 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. cloni® 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 silencing assay for the transformed OL was developed based on Sharon *et al.*[36] and was carried out as follows:

Culture growth. Library-containing bacteria were grown with fresh LB and antibiotic (Kan). Cells were grown to mid-log phase (O.D600 of ~0.6) as measured by a spectrophotometer (Novaspec III, Amersham Biosciences) followed by resuspension with BA buffer and the appropriate antibiotic (Kan). Culture was grown in BA for 3 hours prior to sorting by FACSAria cell sorter (Becton-Dickinson).

FACS sorting. Sorting was done at a flow rate of ~20,000 cells per sec. Cells were sorted into 14 bins (500,000 cells per bin) according to the mCherry to eYFP ratio, in two groups: (i) bins 1-8: high resolution on low ratio bins (30% scale), (ii) 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 (TritonX100 0.1% in 1XTE: 15 μl, culture: 5 μl, 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, 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 10438 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 10432 variants (See Supplementary Table 4). The complete Python pipeline is available on Github.

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.

Position-dependent E5mer effect

To test the effect of E5mer position on the mean e.l.r we fitted the following linear model on the sequences included in Figure 4B:

$$< e.l.r >= \beta_0 + \beta_1 I_{proximal} + \beta_2 I_{distal} + \beta_3 I_{proximal} I_{distal}$$

where $I_{proximal}$ and I_{distal} are indicators for the presence of proximal ($position \geq -25$ relative to the sigma54 TSS) and distal ($position < -25$ relative to the sigma54 TSS) E5mers respectively. We then performed a Two-way ANOVA test on the fitted model. The fitted model and the ANOVA table are presented in Supplementary Note 4.2.

Model for translation level with a partially sequestered RBS

See Supplementary Note 3.

Supplemental Information

Supplementary Figures 1-9

Supplementary Tables 1-5

Supplementary Note 1: The NtrC switch.

Supplementary Note 2: Oligo library and technical assessment.

Supplementary Note 3: Degradation model.

Supplementary Note 4: Additional oligo library and bioinformatics.

Supplementary References.

Supplementary Data 1

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. The authors would like to acknowledge the Technion's LS&E staff (Tal Katz-Ezov, Efrat Barak, Anastasia Diviatis) for help with sequencing and FACS, and Adina Weinberger, Yitzhak Pilpel, Roy Kishony, Michal Brunwasser-Meirom, Noa Katz, Beate Kaufmann, Yaroslav Pollak, and Inbal Vaknin for useful discussions.

Author contribution

LL designed and carried out the experiments for both the initial σ^{54} and OL experiment. LA carried out data analysis for the OL library results and modeled the data. OS carried out the bioinformatics genomic context analysis. RC, SO, and SG designed and carried out the Δ*rpoN* and GST experiments. OA assisted with some of the experiments. RA, ZY, SG, LL, LA, and OS wrote the manuscript.

Figure Captions:

Figure 1: The glnKp σ^{54} promoter can down-regulate another promoter positioned upstream.

(A) Synthetic enhancer design showing the different UAS and σ^{54} promoter combinations used in the experiment. See Supplementary Note 1, and Supplementary Tables 1 and 2 for UAS and promoter details. (B) Left: mCherry expression with enhancer switched to "on'' (NtrC induced), showing varying response for each promoter. Note that for the dual UAS-σ^{70} promoter glnAp1 there is expression with the "no promoter'' control. Right: mCherry expression for enhancers switched to "off'' (NtrC not induced), showing "on'' behavior for all enhancers containing the dual UAS-σ^{70} promoter, except for the enhancer with the glnKp. (C) Flow cytometry data comparing mCherry fluorescence for the glnKp strain in the *E. coli* TOP10 strain (purple) and in the σ^{54} knock-out strain (TOP10:Δ*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-σ^{70} control (left). (E) plate-reader data showing rescue of mCherry fluorescence when the orientation of the glnKp is flipped relative to the upstream σ^{70} promoter.

Figure 2: GlnKp down-regulation is generated by an aSD-SD interaction.

(A) RNAfold generated secondary structures for the no σ^{54} promoter control (top: -46 to +24 – with 0 defined as the TSS), and the glnKp construct (bottom: -38 to +32 with 0 definedas the TSS) highlighting in green the RBS or Shine Dalgarno (SD) sequences for both structures. In the two case, SD sequence is either single stranded (Top), or sequestered in a hairpin structure (Bottom)._(B) qPCR measurements showing the rescue of mCherry mRNA levels by a translated *GST* gene placed upstream of glnKp (third bar from left), as compared with a non-translated *GST* gene (second bar from left). The no-glnKp control is shown for comparison (fourth bar from left). (C) Flow cytometry data showing strong insulation despite adding the *gst* gene. (Top) Circuit diagram showing added eYFP module without an insulating component for mCherry to EYFP expression level ratio measurements (e.l.r). (Bottom) Data for the following constructs: no σ^{70} promoter control (yellow), glnKp construct (blue), glnKp construct with a *gst* gene encoded upstream with (orange) and without an RBS (purple) respectively, and the no- σ^{54} promoter control (red).

Figure 3: Oligo-library analysis of the insulation phenomenon

(A) Oligo library design and schematic description of the protocol. In brief, the synthesized oligo library (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 expression level ratio (e.l.r). DNA from the cells of each bins was barcoded and pooled into a single sequencing run to produce an e.l.r profile for each variant. For details see Materials and Methods. (B) Single sequence variant expression profile: sample data showing the number of reads as a function of mean fluorescence ratio obtained for silencing (top) and non-silencing (bottom) variants, respectively. Straight lines correspond to a smoothing procedure done with a cubic-spline fit to the data. (C) Library expression distribution. Heat map of smoothed, normalized number of reads per expression bin obtained for 10438 analyzed variants ordered according to increasing mean e.l.r. (D) Analysis of the glnKp mutation subset of the library. Flanking regions, core σ^{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. (E) The library mean fluorescence ratio plotted in box-plot 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 (see Supplementary Note 3). Top: the translated pearled phase with low degradation rate. Bottom: the non-translated branched phase with high degradation rate.

Figure 4: Insulation phenomenon is prevalent in other σ^{54} promoters.

(A) Violin plots showing mean expression value distribution for the following variant groups in the library: known σ^{70} promoters (95 variants), no promoter (108 variants), known σ^{54} promoters (227 variants), and σ^{54}-like sequence(161 variants). σ^{54}-like sequence variants were selected based similarity to the core promoter consensus sequence[49] (B) Left: heat map 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 5mer (E5mer) appearance is marked by a brown line at its position within the variant sequence. (Green shade) σ^{54} core promoter region. Top: σ^{54} core promoter consensus sequence[49] Right: Running average on the number of E5mers observed within a variant in the ordered heat map. Top: a PSSM summarizing a multiple alignment of the E5mers found with DRIMust. (C) Box-plot showing groups of σ^{54} –like and annotated σ^{54} promoters differentiated by the location of 5Emers within the 50bp variant sequence. (D) Plots depicting the average number of E5mers found per position on the 50 bp variants for putative and annotated σ^{54} promoters. Variants are grouped by strong (bottom panel and variants below lower dashed line in panel B), moderate (middle panel and variants in between dashed lines in panel B), and no (top panel and variants above middle dashed line in panel B) silencing variants respectively.

References

1. Korbel, J. O., Jensen, L. J., von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22,** 911–917 (2004).
2. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350,** 325–328 (2015).
3. Schwartz, M., Roa, M. & Débarbouillé, M. Mutations that affect lamB gene expression at a posttranscriptional level. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 2937–2941 (1981).
4. De Smit, M. H. & Duin, J. V. in *Progress in Nucleic Acid Research and Molecular Biology* (ed. Moldave, W. E. C. and K.) **38,** 1–35 (Academic Press, 1990).
5. Ma, J., Campbell, A. & Karlin, S. Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures. *J. Bacteriol.* **184,** 5733–5745 (2002).
6. Campo, C. D., Bartholomäus, A., Fedyunin, I. & Ignatova, Z. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLOS Genet* **11,** e1005613 (2015).
7. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **9,** 675 (2013).
8. Gu, W., Zhou, T. & Wilke, C. O. A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. *PLOS Comput Biol* **6,** e1000664 (2010).
9. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–258 (2009).
10. Sneppen, K. *et al.* A Mathematical Model for Transcriptional Interference by RNA Polymerase Traffic in Escherichia coli. *J. Mol. Biol.* **346,** 399–409 (2005).

11. Yokobayashi, Y., Weiss, R. & Arnold, F. H. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci.* **99,** 16587–16591 (2002).
12. Hao, N. *et al.* Road rules for traffic on DNA—systematic analysis of transcriptional roadblocking in vivo. *Nucleic Acids Res.* **42,** 8861–8872 (2014).
13. Epshtein, V., Toulmé, F., Rahmouni, A. R., Borukhov, S. & Nudler, E. Transcription through the roadblocks: the role of RNA polymerase cooperation. *EMBO J.* **22,** 4719–4727 (2003).
14. Shearwin, K. E., Callen, B. P. & Egan, J. B. Transcriptional interference – a crash course. *Trends Genet.* **21,** 339–345 (2005).
15. Buck, M., Gallegos, M.-T., Studholme, D. J., Guo, Y. & Gralla, J. D. The Bacterial Enhancer-Dependent ς54(ςN) Transcription Factor. *J. Bacteriol.* **182,** 4129–4136 (2000).
16. Bush, M. & Dixon, R. The Role of Bacterial Enhancer Binding Proteins as Specialized Activators of σ54-Dependent Transcription. *Microbiol. Mol. Biol. Rev.* **76,** 497–529 (2012).
17. Murakami, K. S., Masuda, S. & Darst, S. A. Structural Basis of Transcription Initiation: RNA Polymerase Holoenzyme at 4 Å Resolution. *Science* **296,** 1280–1284 (2002).
18. Atkinson, M. R., Blauwkamp, T. A., Bondarenko, V., Studitsky, V. & Ninfa, A. J. 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 (2002).
19. Atkinson, M. R., Savageau, M. A., Myers, J. T. & Ninfa, A. J. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell* **113,** 597–607 (2003).
20. Claverie-Martin, F. & Magasanik, B. Role of integration host factor in the regulation of the glnHp2 promoter of Escherichia coli. *Proc. Natl. Acad. Sci.* **88,** 1631–1635 (1991).

21. Feng, J., Goss, T. J., Bender, R. A. & Ninfa, A. J. Repression of the Klebsiella aerogenes nac promoter. *J. Bacteriol.* **177,** 5535–5538 (1995).
22. Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41,** D605–D612 (2013).
23. Reitzer, L. & Schneider, B. L. Metabolic context and possible physiological themes of sigma(54)-dependent genes in Escherichia coli. *Microbiol. Mol. Biol. Rev. MMBR* **65,** 422–444, table of contents (2001).
24. Amit, R., Garcia, H. G., Phillips, R. & Fraser, S. E. Building Enhancers from the Ground Up: A Synthetic Biology Approach. *Cell* **146,** 105–118 (2011).
25. Hoover, T. R., Santero, E., Porter, S. & Kustu, S. The integration host factor stimulates interaction of RNA polymerase with NIFA, the transcriptional activator for nitrogen fixation operons. *Cell* **63,** 11–22 (1990).
26. Kiupakis, A. K. & Reitzer, L. ArgR-Independent Induction and ArgR-Dependent Superinduction of the astCADBE Operon in Escherichia coli. *J. Bacteriol.* **184,** 2940–2950 (2002).
27. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chem. Chem. Mon.* **125,** 167–188
28. Nakamoto, T. A unified view of the initiation of protein synthesis. *Biochem. Biophys. Res. Commun.* **341,** 675–678 (2006).
29. Scharff, L. B., Childs, L., Walther, D. & Bock, R. Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites. *PLOS Genet.* **7,** e1002155 (2011).
30. Hui, M. P., Foley, P. L. & Belasco, J. G. Messenger RNA Degradation in Bacterial Cells. *Annu. Rev. Genet.* **48,** 537–559 (2014).
31. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature* **451,** 355–358 (2008).

32. Richards, J., Luciano, D. J. & Belasco, J. G. Influence of translation on RppH-dependent mRNA degradation in Escherichia coli. *Mol. Microbiol.* **86,** 1063–1072 (2012).

33. Mackie, G. A. Ribonuclease E is a 5′-end-dependent endonuclease. *Nature* **395,** 720–724 (1998).

34. Robertson, H. D. Escherichia coli ribonuclease III cleavage sites. *Cell* **30,** 669–672 (1982).

35. Calin-Jageman, I. & Nicholson, A. W. Mutational Analysis of an RNA Internal Loop as a Reactivity Epitope for Escherichia coli Ribonuclease III Substrates. *Biochemistry (Mosc.)* **42,** 5025–5034 (2003).

36. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30,** 521–530 (2012).

37. Philips, R. M. & R. *Cell Biology by the Numbers*. (Garland Science, 2016).

38. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering Motifs in Ranked Lists of DNA Sequences. *PLOS Comput Biol* **3,** e39 (2007).

39. Leibovich, L. & Yakhini, Z. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res.* gks206 (2012). doi:10.1093/nar/gks206

40. Leibovich, L., Paz, I., Yakhini, Z. & Mandel-Gutfreund, Y. DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res.* **41,** W174–W179 (2013).

41. Salgado, H. *et al.* RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41,** D203–D213 (2013).

42. Schwab, D. & Bruinsma, R. F. Flory Theory of the Folding of Designed RNA Molecules. *J. Phys. Chem. B* **113,** 3880–3893 (2009).

43. Babitzke, P., Baker, C. S. & Romeo, T. Regulation of Translation Initiation by RNA Binding Proteins. *Annu. Rev. Microbiol.* **63,** 27–44 (2009).

44. Winkler, W. C. & Breaker, R. R. Regulation of Bacterial Gene Expression by Riboswitches. *Annu. Rev. Microbiol.* **59,** 487–517 (2005).

45. Komarova, A. V., Tchufistova, L. S., Supina, E. V. & Boni, I. V. Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *RNA* **8,** 1137–1147 (2002).

46. Francke, C. *et al.* Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12,** 385 (2011).

47. Jakobsen, J. S. *et al.* σ54 Enhancer Binding Proteins and Myxococcus xanthus Fruiting Body Development. *J. Bacteriol.* **186,** 4361–4368 (2004).

48. Jiang, Y. *et al.* Multigene Editing in the Escherichia coli Genome via the CRISPR-Cas9 System. *Appl. Environ. Microbiol.* **81,** 2506–2514 (2015).

49. Barrios, H., Valderrama, B. & Morett, E. Compilation and analysis of σ54-dependent promoter sequences. *Nucleic Acids Res.* **27,** 4305–4313 (1999).

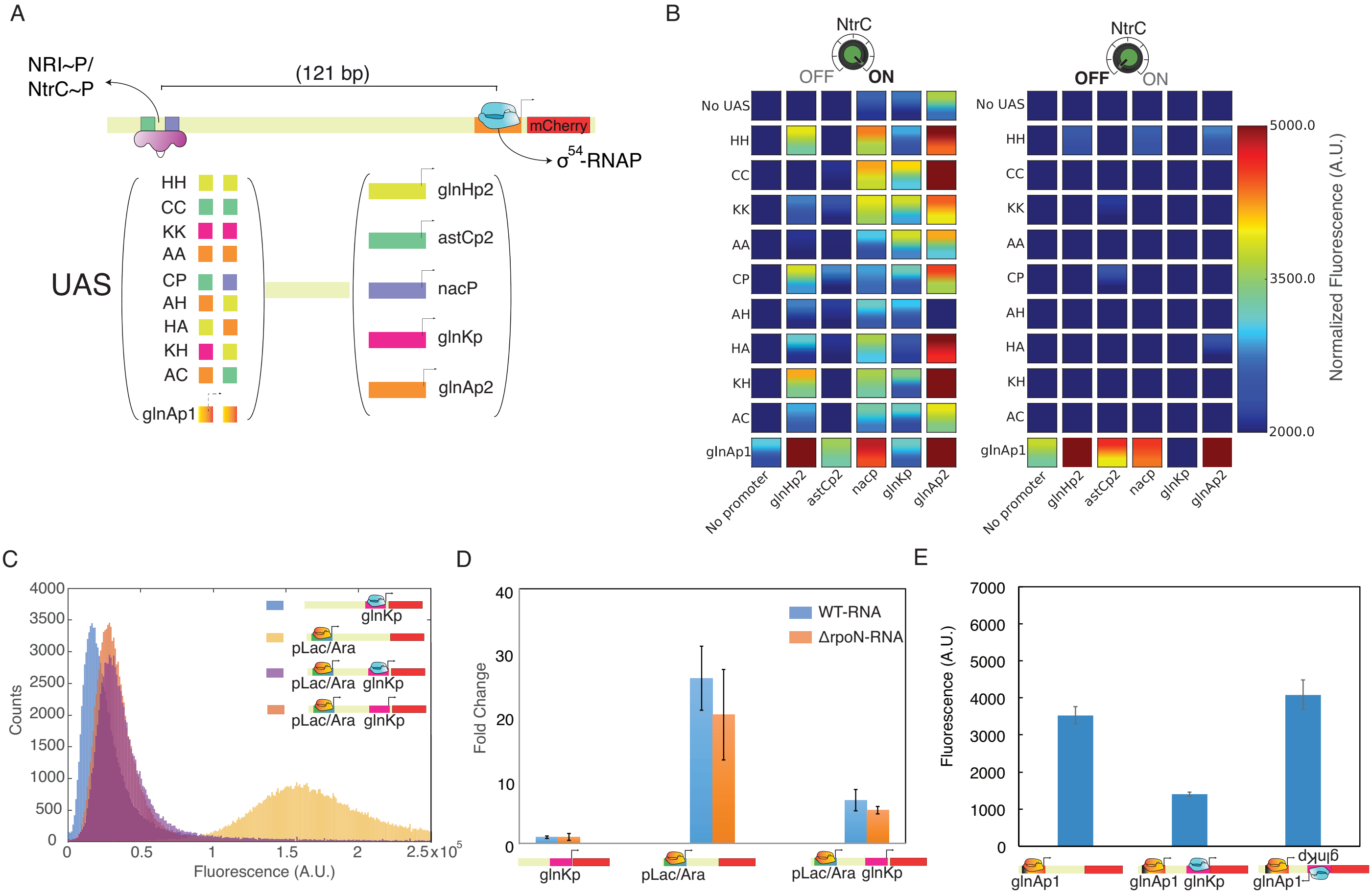

A
NRI~P/
NtrC~P
(121 bp)
mCherry
σ54-RNAP
UAS
HH
CC
KK
AA
CP
AH
HA
KH
AC
glnAp1
glnHp2
astCp2
nacP
glnKp
glnAp2
B
NtrC
OFF
ON
No UAS
No promoter
nacp
5000.0
3500.0
2000.0
Normalized Fluorescence (A.U.)
C
Counts
Fluorescence (A.U.)
0
0.5
1
1.5
2
2.5x10^5
4000
3500
3000
2500
2000
1500
1000
500
pLac/Ara
pLac/Ara glnKp
D
Fold Change
40
30
20
10
WT-RNA
ΔrpoN-RNA
E
Fluorescence (A.U.)
7000
6000
5000
4000
3000
2000
1000
glnAp1 glnKp

Figure 1

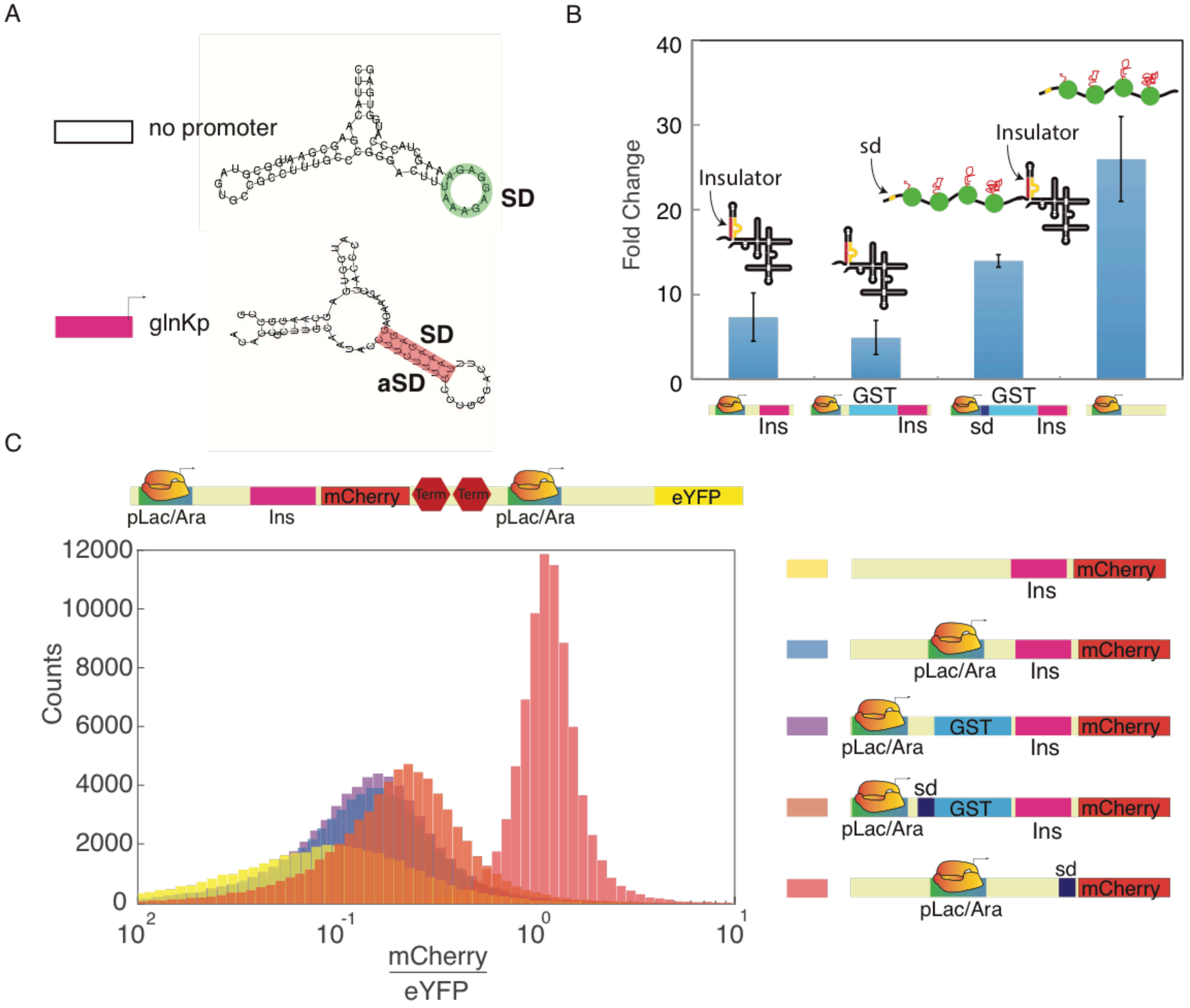
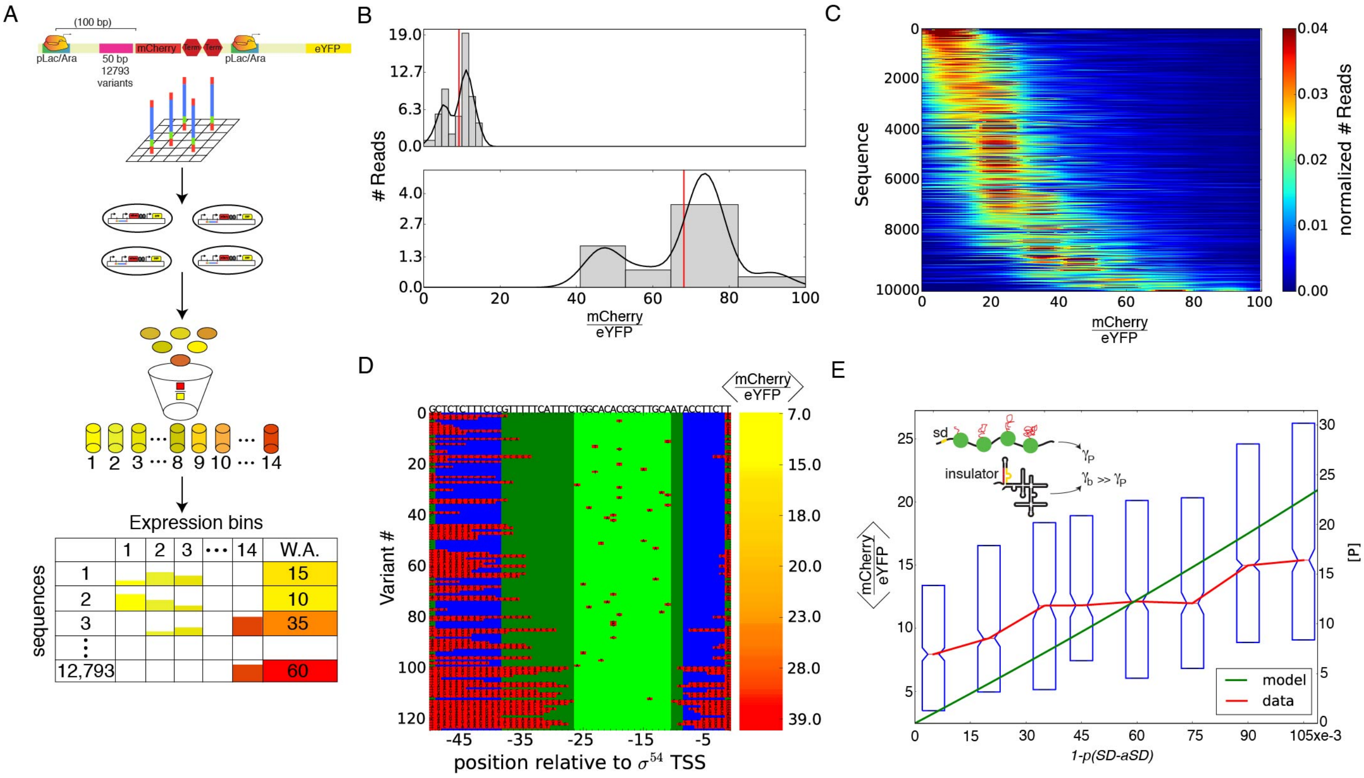

A
no promoter
SD
glnKp
SD
aSD
B
Fold Change
40
30
20
10
0
Insulator
sd
Insulator
Ins
GST
Ins
GST
sd
Ins
C
pLac/Ara
Ins
mCherry
Term
Term
pLac/Ara
eYFP
Counts
12000
10000
8000
6000
4000
2000
0
10^{-2}
10^{-1}
10^{0}
10^{1}
mCherry / eYFP
Ins
mCherry
pLac/Ara
Ins
mCherry
pLac/Ara
GST
Ins
mCherry
pLac/Ara
sd
GST
Ins
mCherry
pLac/Ara
sd
mCherry

Figure 2



Expression bins

sequences	1	2	3	•••	14	W.A.
1						15
2						10
3						35
⋮						
12,793						60

Figure 3

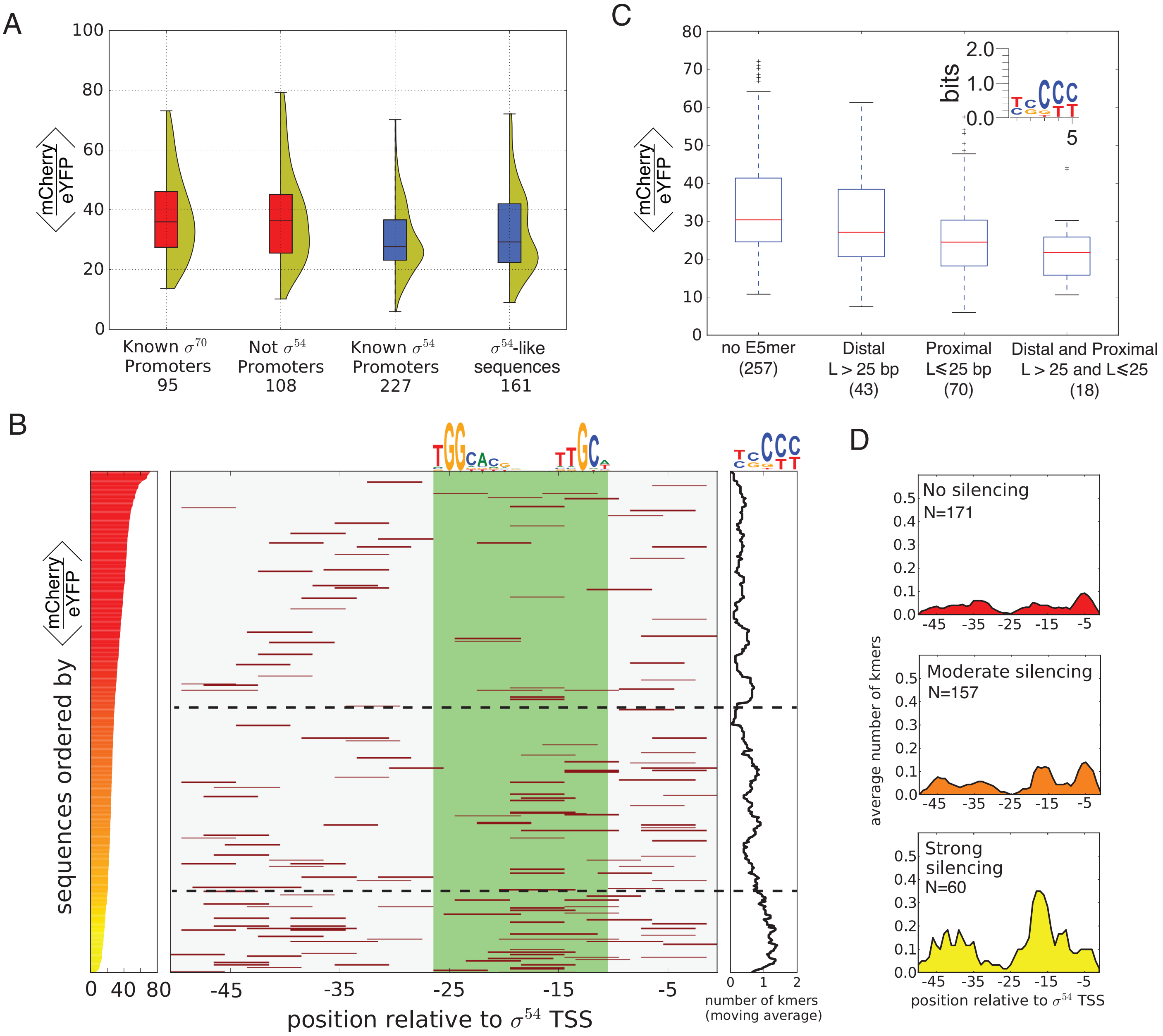

A
100
80
60
40
20
0
⟨mCherry/eYFP⟩
Known σ^{70} Promoters 95
Not σ^{54} Promoters 108
Known σ^{54} Promoters 227
σ^{54}-like sequences 161
C
80
70
60
50
40
30
20
10
0
⟨mCherry/eYFP⟩
bits
2.0
1.0
0.0
5
no E5mer (257)
Distal L > 25 bp (43)
Proximal L⩽25 bp (70)
Distal and Proximal L > 25 and L⩽25 (18)
B
sequences ordered by ⟨mCherry/eYFP⟩
0 40 80
-45
-35
-25
-15
-5
position relative to σ^{54} TSS
0 1 2
number of kmers (moving average)
D
average number of kmers
No silencing N=171
Moderate silencing N=157
Strong silencing N=60
0.5
0.4
0.3
0.2
0.1
0.0
-45 -35 -25 -15 -5
position relative to σ^{54} TSS

Figure 4