Short CT-rich motifs can trigger context-specific silencing of gene expression in bacteria

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[1,3], 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 5000 variants together with a synthetic biology approach to study a generic context-dependent silencing phenomenon in *E. coli*. The observed silencing 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 synthetic constructs that encode a non-silenced RBS upstream of the silencing motif, that sequestration of the RBS and subsequent rapid messenger degradation is likely to be the mechanism driving the silencing effect. This sequestration is probably due to binding of the RBS to the upstream CU-rich motifs, which we call anti-Shine-Dalgarno (aSD) motifs. To provide further support for the importance of this mechanism in natural systems, we show bioinformatically that the genomes of mesophilic and psychrophilic bacteria are significantly depleted for the observed aSD motifs within 300 bp of putative Shine-Dalgarno motifs (GA-rich hexamers) as

compared with a random control in over 70% of the 591 genomes examined. In contrast, in genomes of thermophilic and hyperthermophilic bacteria there is no such depletion, which is consistent with a weak interaction between the short aSD CU-rich motif and the RBS that is thermodynamically less stable at higher ambient living temperatures. Our findings have important implications for understanding SNP/INDEL mutations in regulatory regions, as well as provide a mechanism for promoter/operon insulation in bacterial genomes.

Deconstructing genomes to their basic parts and then using those parts to construct *de novo* gene regulatory architectures are amongst the main goals of synthetic biology. First, a thorough breakdown of a genome to its basic regulatory and functional elements is required. Then, each element must be analyzed to decipher the properties and mechanisms that drive its activity. Lastly, these well-defined elements can be used as building blocks for *de novo* systems. However, *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 interaction with nearby elements.[1] Such interactions might endow some secondary regulatory benefit that is overlooked by standard analysis methods or is masked by a stronger regulatory effect in the native setting.[2] Despite being implicated in many gene-expression-related processes, context effects in bacteria have only 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 of context related constraints.[3–5] However, there has been little systematic study of the nature of context effects in bacterial non-coding or regulatory regions.

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.[6] 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. Accounting for context effects in the design of many synthetic circuits using this approach is therefore more often than not impractical.

Synthetic oligo libraries (OLs) together with high-throughput screening methods provide an alternative approach that enables direct studies of context-related effects. Synthetic OLs have been used to examine regulatory elements systematically, and have revealed the effects of element location and multiplicity.[7,8] This approach can be used to investigate secondary context-related phenomena in non-coding regulatory elements. In this work we study context effects in noncoding regions upstream of a Shine-Dalgarno (SD) sequence in bacteria. We first use a synthetic biology circuit to identify a single regulatory

sequence in *E. coli* for which we observe strong silencing with no obvious regulatory mechanism. Using synthetic OLs inspired by the identified sequence, fluorescence-activated cell sorting (FACS) and high-throughput sequencing (following the method of Sharon *et al.*), also in *E. coli*, we identify a general context CT-rich motif that correlates with gene silencing. Using a free-energy model, we propose a temperature-dependent silencing mechanism based on degradation. Here, the CT-rich motif triggers the formation of mRNA secondary structure that interferes with the ribosome binding site's availability, lowers ribosome occupancy and then leads to rapid degradation of the mRNA. Finally, we analyze 672 genomes of bacteria that live in a wide range of temperatures. We find that the context motif identified by our experiments is depleted in the genomic sequences of psychrophiles and mesophiles, but not in the genomes of thermophiles and hyperthermophiles. This observation is consistent with the proposed silencing mechanism.

Results

The σ^{54} glnKp promoter silences expression from an upstream promoter

We engineered a set of synthetic circuits to test the components of bacterial enhancers, 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) made 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 which often harbors additional transcription factor binding sites.[9–11] In our study, 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[12–17] 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 .[18]

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[14,19,20] 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 glnAp2 expression 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 any detected expression should emerge solely from the σ^{70} promoter. The only UAS for which significant expression was observed was glnAp1 (p-value<0.05, t-test after correction for multiple hypothesis). This is consistent with the lack of σ^{54} activity. However, glnAp1 showed a detectible fluorescence response for four of the five promoters only. 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 promoter instead of the glnAp1 UAS. In Fig. 1C, we show that the circuit with both the pLac/Ara and the glnKp promoters (purple) generates about a factor of 10 less fluorescence than the control lacking the glnKp promoter (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.

Oligo library (OL) analysis

To further explore the silencing phenomenon induced by glnKp, and to check for its prevalence in other bacterial genomes, we constructed an OL of 12758 150-bp variants (Fig. 2A). 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.[8] The OL was designed to test both known σ^{54} promoters from various organisms and putative σ^{54} promoters from varying genomic regions in *E. coli* and *V. cholera* for the silencing effect. In addition, 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 2A). 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[8], we obtained the expression distribution for each sequence variant (Fig. 2B). Figure 2C shows the expression profiles for 5167 variants with sufficient total number of sequence counts (n>16, see Materials and Methods for details), revealing a wide range of expression levels. While a significant percentage of the variants showed low mean expression levels, a non-negligible set of variants produced high expression levels (Fig. 2B top and bottom show representative examples). Checking the distribution of mean expression levels (see Materials and Methods) within each class of variants (Fig. 2D and Supp. Note 2), we observed that the "no-promoter" class had a slight enrichment towards non-silencing expression values (p-value<0.05, mHG test[21]) and that the glnKp and σ^{70} classes both showed enrichment for the silencing and non-silencing regimes, respectively (p-value<0.001, mHG test). The core σ^{54} promoter and σ^{54}-like variants showed no enrichment, but rather displayed a wide distribution over the entire range of expression levels. Together, the library seems to indicate that the underlying mechanism leading to the large proportion of silencers detected is not associated with a primary regulatory function (e.g.

promoter type, transcription factor binding site, etc.) but rather is independently encoded separately within the 50 bp variable sequences themselves.

In order to determine the mechanism, which underlies the silencing, we performed a DRIMust motif search on our variant library sorted by the mean mCherry to eYFP expression ratio values. DRIMust is a tool designed to identify enriched sequence motifs in a ranked list of sequences.[21–23] Our analysis revealed that a CT-rich consensus motif is enriched in the silenced variants (p-value $< 10^{-54}$, mHG). We plot the results in Fig. 3A. The consensus motif is derived from a list of ten 5 bp CT-rich features, each enriched in the top of the ranked list. The middle panel of Fig 3A shows the position of each of these features, marked by a brown line in the corresponding variant. In the right panel, we show the running average of the number of motif occurrences over 50 variants. 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 0) is strongly associated with a variant exhibiting low mCherry to eYFP fluorescence ratio. To provide further support for this observation, we closely examined the 268 variants of glnKp (Fig. 3B and Supp. Note 2). Here, we see that mutations in the core σ^{54} promoter region yield a negligible change in the silencing effect. However, mutations in the CT-rich segments of the flanking region (labeled in blue) alleviate the effect, with the region closer to the SD motif leading to the greatest difference. Finally, to assess the effect of CT-rich motif position on the silencing levels we measured the silencing magnitude at each position. To do this, we compared the mean mCherry to eYFP fluorescence ratio between two groups of variants – those that have the CT-rich motif at the examined position and those that do not. Fig 3C shows the ratio between the two groups at each position revealing a strong silencing effect for motifs located within 18 bp upstream of the SD motif. As a result, we conclude that the context-dependent silencing phenomenon that we observed is likely induced by a 5 bp pyrimidine-rich segment, provided that the latter is positioned within 20 bp of the Shine-Dalgarno motif.

Degradation model and supporting experiments

Given that the RBS or Shine-Dalgarno (SD) sequence is typically encoded by a purine-rich motif (GGAGAA in our case), we hypothesized that the CT-motif encodes a weak anti-Shine-Dalgarno sequence (aSD) that is capable of binding the RBS sequence with as few as 3 complementary nucleotides forming a hairpin. This secondary structure, in turn, blocks the small sub-unit of the ribosome from assembling, which prevents translation. Consequently, the mRNA molecule for the most part will be devoid of ribosomes, triggering

the collapse of the molecule into a "branched" phase[24,25] (Fig. 4A, bottom), as compared with the translationally-active "pearled" phase (Fig. 4A, top). The branched form is susceptible to degradation[26] as follows. First, by facilitating the pyrophosphotation of the 5'end by RppH[27], normally inhibited by the ribosome.[28] This newly formed monophosphorylated 5'-terminus triggers the activity of the endonuclease RnaseE, which is known to bind this form of RNA with significantly higher affinity as compared with the tri-phosphorylated 5'-terminus.[29] Second, the collapsed branched structure is a natural substrate for RnaseIII, which cleaves either extended double stranded secondary features[30], or extended secondary structures containing internal loops.[31] Consequently, the mechanism underlying silencing should be a combination of ribosome binding inhibition and the resultant increased degradation rate.

To provide experimental support for this hypothesis, we designed two additional constructs: we encoded the full *glutathione S-transferase* (*GST*) gene upstream of the glnKp promoter (under the control of pLac/Ara), with and without a SD motif. We then measured the mRNA levels for GST as compared with mCherry. We reasoned that a translated gene placed upstream of the aSD sequence would protect the entire mRNA from the pyrophosphotation of the 5'end by RppH. This, in turn, would inhibit the RnaseE degradation pathway, leading to a partial rescue of the mCherry silencing effect, as the RnaseIII mode would still be active. The quantitative PCR (qPCR) results are shown in Fig. 4B. 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. 4B 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 fluorescence ratio recovery. Here, we observe only a modest recovery in fluorescence (Fig. 4C). This is consistent with the partial recovery in mRNA levels, and the remaining susceptibility of this part of the molecule to RnaseIII.

To provide further support for the combined sequestration/degradation mechanism, we employed two models. In the first, we use the RBSDesigner[32] engine to compute the predicted translation rate of the variants in our library. This model bases its prediction on secondary structure assessment and the inferred availability of the RBS, and ignores potential changes in degradation rate. The data show (Supp. Fig. 8) no obvious correlation between the measured fluorescence ratio and the predicted results. In the second model, we combine RNA structure predictions with a model for increased degradation (see Supp. Note 3 for details). In

brief, we used RNAfold[33] to compute the probability for the RBS to be sequestered in a secondary structure (see Supp. Note 3) for each variant in our library, and found that the mean expression level correlated with the computed probability. We then constructed a degradation model (see Supp. Note 3) taking into account the probability that the RBS is sequestered, under the assumption that a sequestered and non-sequestered RBS correspond to the branched and pearled phases respectively. We assumed different degradation rates for each phase, and used realistic constant rates for other kinetic processes (e.g. transcription rate, translation rate, etc.[34]). In Fig. 4D we plot the fluorescence ratio as a function of the probability of the RBS to be non-sequestered (i.e. pearl probability), and found that our degradation-based model matches the data closely (red line and inset).

Analyzing 672 genomes for evidence of CT-rich silencing

The OL results imply that if the discovered aSDs are prevalent in the upstream vicinity of naturally-occurring SDs in bacterial genomes, then many of the genes downstream of the SDs should be silenced. This implies that the average ribosome occupancy and the propensity of CU-mers to be double-stranded should be lower and higher, respectively, as compared with a randomized sequence. To check this, we analyzed both ribosome occupancy[35] and SHAPE-Seq[36] data sets obtained for *E. coli* and found both a clear anti-correlation with ribosome occupancy (Supp. Fig. 9) and an enrichment of double-stranded RNA hits ($p<0.016$, Wilcoxon – Supp. Fig. 10) as compared with a random sequence. We then reasoned that such contextual effects would be evolutionarily undesirable. Therefore, we predicted that sequences similar to our proposed aSD would be depleted upstream of SD loci, as compared with random sequences of the same length. To test our prediction, we analyzed 672 bacterial genomes representing four environmental temperature ranges: 13 psychrophiles, 578 mesophiles, 71 thermophiles, and 10 hyperthermophiles (their descriptions were taken from EcoCyc. See Supplementary Data 2. For additional details see Supp. Note 4 and Supp. Fig. 11). In Fig. 5A-D we present the distributions obtained for four representative strains, one from each environmental temperature range. We plot the distribution obtained from aSD SD pairs (aSD:SD) in orange bars, and compare it to random-sequence SD pairs (random:SD) for control in blue. The data show that for *E. coli* and for the psychrophilic *S. psychrophila* the CT-rich distribution is centered on lower percentage values as compared with the random control, indicating that aSD:SD occurrences within less than 300 bp are significantly less frequent than expected by the random model. In contrast, both thermophilic

and hyper-thermophilic species (Fig. 5C-D) show an opposite effect, namely a slightly higher prevalence of proximal occurrences of aSD:SD as compared with the random model.

To gain a more quantitative understanding of this depletion effect, we compared the distributions of proximal occurrences of aSD:SD to proximal occurrences of random:SD pairs in all 672 bacteria from different environmental temperature ranges, and determined the significance of the observed depletion for each species. This comparison shows (Fig. 5E) that a significant percentage of psychrophilic and mesophilic indeed exhibit depletion ($p < 0.05$, Wilcoxon one tail) for most of the bacterial species (76.92% and 64.19% for psychrophilic and mesophilic, respectively. Supp. Fig. 15). However, in thermophiles and hyperthermophiles the fraction is significantly lower (35.21% and 20% for thermophilic and hyper-thermophilic, respectively. Supp. Fig. 15). Another way of viewing the significance of this result is to first compute the mean values for the aSD-SD and random-SD distributions for each organism (e.g. the distribution presented in Fig. 5A-D). Then, we plot the plot the distribution of these mean values for the mesophilic/psychrophilic group and separately for the thermophiles. It can be seen in Fig. 5F that for the bacteria that thrive at lower temperatures (top), the multi-species distribution of aSD:SD pairs is clearly shifted as compared with the proximal occurrences of random:SD controls, while the distributions of the mean values for the thermophilic bacteria overlap. Together, these analyses show that there is strong evolutionary pressure to deplete CT-rich sequences (like the aSD motifs) from the upstream vicinity of SD-like sequences in bacteria that thrive in the temperature range of -10-40°C, while much weaker pressure seems to be at work for thermophilic organisms.

Finally, we checked whether the depletion of the putative aSD occurs symmetrically around the SD. To do this we repeated our analysis, only this time we compared the distributions for proximal occurrences of SD:aSD (with the aSD located downstream to the RBS) to the distributions of SD:random pairs for every bacteria in our analysis (672 bacteria from above section). The results are presented in Figure 5G. The depletion effect nearly disappears in the SD:aSD configuration, indicating that the evolutionary pressure is an asymmetric phenomenon which only occurs when the putative aSD are located upstream of putative SD loci.

Discussion

We used a synthetic oligonucleotide-library approach to uncover a context-dependent phenomenon of silenced protein expression. The silencing phenomenon was found to depend on the presence of a short (3-5 nucleotide) CT-rich sequence motif upstream of a Shine-

Dalgarno motif. We speculate that these short CT-rich sequences encode a weakly-specific CU-rich sequence that can bind the RBS, and likely triggers a collapse of the RNA molecule into a branched phase due to the lack of translocating ribosomes on the mRNA molecule. Literature shows that branched-phase mRNA molecules are degraded at a higher rate as compared to ribosome-bound mRNA molecules.[28] Moreover, 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.[37] Therefore, the sequestration of the RBS and the resulting reduced translation initiation rate together with the increased degradation rate, lead to a sharply reduced protein product and thus to the observed silencing phenomenon.

Given the potency of this regulatory effect, we checked for evidence for the presence or absence of the defined aSD motif in the vicinity of the SD motif in 672 bacterial genomes. Our analysis shows that there is a strong depletion of these sequences upstream of the putative SD motifs in mesophilic and psychrophilic strains, but no such depletion was detected in thermophilic organisms. This suggests that the aSD-RBS interaction in mesophilic organisms consists of a weak base-pairing interaction, which is less stable at the higher ambient living temperatures of thermophilic organisms, thus reducing the deleterious potential of such sequences in the vicinity of thermophile ribosome binding sites. In addition, the phenomenon was not found to be symmetric around the SD motif, as the depletion was not detected immediately downstream of the SD (i.e. inside the coding region). This observation is consistent with past work[3–5,38] that showed that the codons in the N-terminus of genes select against secondary structures. In this sense, our analysis provides the complementary observation by showing that the 5'UTR region of mRNA should also be devoid of secondary structure in order to facilitate translation. Consequently, our findings have important implications for understanding codon usage and SNP/INDEL mutations in regulatory regions (i.e. any mutation into a C/T or away from it).

We began this study by exploring the transcriptional variation encoded within σ^{54} promoters, which led to the observation of the glnKp silencing phenomenon. Interestingly, the analysis of our library showed that this silencing phenomenon is not isolated to glnKp, but can be found in ~20% of the annotated and putative σ^{54} promoters that we studied (Fig. 2D). Thus, the proportion of σ^{54} promoters encoding this effect is significantly larger than the genomic baseline of mesophiles, which averages about 5% of such motifs in the vicinity of putative RBS (Fig. 5). As a result, we speculate that this aSD-RBS silencing mechanism may be used to insulate the σ^{54} promoters containing the CT-rich motifs from upstream

transcriptional interference. This protection mechanism can play an important biological role in gene-dense microbial genomes by insulating operons that are only needed in particular stress or growth niches, as is the case for glnKp.

Finally, given the strength of the regulatory effect and the strong depletion observed in other bacterial genomes, why was this phenomenon not reported previously? While the effects of *cis* encoded anti-Shine-Dalgarno sequences are well-documented in the context of riboswitches and RNA binding protein interactions with RNA[39,40], the aSD sequences in those cases were encoded into longer motifs. In our case, the aSD motif we identified is only 3-5 nucleotides long, and would thus have been nearly impossible to detect in lower-throughput experiments, where the low number of variants limits the statistical significance needed for enrichment detection. As a result, we believe that our large-scale synthetic OL-based approach was the main catalyst for uncovering the regulatory role of the aSD-RBS interaction. Such an approach can be applied to other poorly understood context-related phenomena, where no obvious sequence signature emerges from preliminary studies. Thus, it is possible that context-related effects are phenomenon encoded by short sequence motifs, which can be uncovered in a focused library search such as the one used here.

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[18] 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 [41], 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.

High-throughput oligo library expression assay

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 library 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, 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, we designed our library to be 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 [42], 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 Materials and Methods). Finally, all variants were encoded so they would appear in both sense and anti-sense orientations with respect to the pLac/Ara driver promoter.

To generate the variant library, the following oligo library 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).

Oligo library technical assessment

See Supplementary Note 2.

Oligo library cloning

Oligo library cloning was based on the cloning protocol developed by the Segal group[8] (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.

Oligo library transcriptional-silencing assay

The oligo-library silencing assay for the transformed oligo-pool library was developed based on [8] 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 at least 8 reads in each set of bins (1-8, 9-15). 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 5167 variants. 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 (W.A.) 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.

Minimal hyper-geometric (mHG) enrichment tests

We sorted the sequences by their mean expression ratio values and calculated minimal hyper-geometric enrichment[21] scores for each of the variant groups (known σ^{70} promoters, known σ^{54} promoters, glnKp perturbations, etc.) for both the top and the bottom of the sorted list. The test was preformed using the xlmhg python library[43] version 1.1rc3.

Position-dependent motif effect

For each position, we calculated a score indicating its similarity to the consensus CT-rich motif by calculating the log likelihood of observing the 5 bp sequence originating from that position, given the probabilities from the motif position specific scoring matrix.[44]

$$Score(s,j) = \sum_{i=j}^{j+5} \log\big(M(i,s_i)\big),$$

where *M(i,x)* is the probability for *x* in the *i*th position of the motif.

We than averaged the position-specific scores over a 3-base running window. The averaged PSSM score was then transformed into a binary hit/miss score using a threshold of -11.5. This resulted in two groups of variants separated by their similarity to the CT-rich motif. The mean mCherry to eYFP ratio of the two groups is compared by calculating:

$$motif\ effect = \log_{10}\left(\frac{NoMotif}{Motif}\right).$$

Model for translation level with a partially sequestered RBS

See Supplementary Note 3.

Analysis of aSD-SD depletion in 672 bacterial genomes

See Supplementary Note 4.

See Supplementary Data 2 for details of the 672 strains.

Supplemental Information

Supplementary Figures 1-17

Supplementary Tables 1-3

Supplementary Note 1: The NtrC switch.

Supplementary Note 2: Oligo library and technical assessment.

Supplementary Note 3: Degradation model.

Supplementary Note 4: Bioinformatic analysis details

Supplementary References.

Supplementary Data 1-2

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 the analysis for the OL library results and modeled the data. OS carried out the bioinformatic 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 downregulate 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: ratio of mCherry to eYFP 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: Expression ratio 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 promoter. (C) Flow cytometry data comparing mCherry to eYFP 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) platereader data showing rescue of mCherry fluorescence when the orientation of the glnKp promoter is flipped relative to the upstream σ^{70} promoter.

Figure 2: The silencing phenomenon is prevalent in the oligo library.

(A) Oligo library design and schematic for protocol. In brief, the synthesized oligo library (Twist Bioscience) was cloned into *E. coli* competent cells, and 10 million colonies were collected. The colonies were grown in LB to OD 0.6 and sorted by FACS into 14 expression bins according to cell mCherry to eYFP fluorescence ratio. A PCR reaction was carried out on each bin, during which a bin-specific barcode was added. The combined PCR reactions were sequenced using Illumina HiSeq platform. For details see Materials and Methods. (B) Single-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 5167 analyzed variants ordered according to increasing mean expression level ratio. The OL variants included the following classes: without σ^{54} promoter ("No promoter"), with σ^{70} promoter instead of σ^{54} ("σ^{70} promoters"), mutated glnK promoter ("glnKp") and other σ^{54} promoters and σ^{54}-like variants ("other"). (D) Mean expression level distribution across variant groups: violin plots showing mean

expression value distribution for each of the variant groups in the library. Stars correspond to degree of enrichment/depletion w.r.t to the "Others" group.

Figure 3: Silencing is localized to a CT-rich pentamer which encodes a weak anti-Shine-Dalgarno sequence.

(A) Left: heat map ordering of the examined variants by mean fluorescence ratio, with silenced variants at the top. We use a yellow-to-red color scale to present the variants according to increasing values of mean mCherry to eYFP fluorescence ratio. Middle: for each variant in the left panel, each CT-pentamer appearance is marked by a brown line at its position within the variant sequence. Right: Running average on the number of CT-rich pentamers that occur within a variant in the ordered heat map. (B) Analysis of the glnKp mutation subset of the library. Flanking regions, core σ^{54} promoter, the CT-pentamers, and mutations are denoted by dark green, light green, blue, and red boxes, respectively. Right panel denotes the mean fluorescence ratio using the yellow-to-orange scale of (A). (C) CT-motif effect on expression as a function of upstream distance from the RBS (positioned at 0) calculated for the library variants (glnK variants excluded).

Figure 4: CT-rich pentamer triggers mRNA degradation.

(A) Schematic for the degradation model. Top: the translated pearled phase with low degradation rate. Bottom: the non-translated branched phase with high degradation rate. (B) qPCR measurements showing the rescue of mCherry fluorescence 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 increased mCherry expression when *GST* is translated. (D) 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: the degradation model (see Supplementary Note 3) prediction for RNA concentration (green) as compared with the same model with equal degradation rates for the pearled and branched phases (blue).

Figure 5: Proximal occurrences of aSD:RBS pairs are depleted in cold- and moderate-living-temperature bacteria.

(A)-(D) Distributions of % proximal occurrences (where aSD is located upstream and in close proximity to SD) of aSD-SD pairs (orange) as compared with the distributions of

proximal occurrences of random-SD pairs for 4 bacteria representing the different environmental temperatures ranges. One-tail Wilcoxon p-values for the depletion or enrichment of proximal occurrences of aSD:SD are (A) mesophilic *E. coli*, $< 10^{-15}$ (depletion), (B) psychrophylic *S. psychrophila*, $< 10^{-9}$ (depletion), (C) thermophilic *Aciduliprofundurn sp.*, $< 10^{-15}$ (enrichment), and (D) hyper-thermophilic *T. ruber*, $< 10^{-15}$ (enrichment). (E) Bar graph showing the percentage of bacteria exhibiting aSD:SD depletion in the different environmental temperature ranges (hyper-thermophilic, thermophilic, mesophilic and psychrophilic). Depletion is defined as reduction in proximal occurrences of aSD:SD compared to proximal occurrences of random:RBS as determined by one-tail Wilcoxon test at p-value<0.05. (F) The mean value for the distribution of proximal occurrences of aSD:SD (orange) versus random:SD (blue) in each bacteria type. Top - mesophilic bacteria. Bottom – thermophilic bacteria. (G) Asymmetry in the observed depletion trend. Proximal occurrences of SD:aSD (where the aSD is located in close proximity and downstream to the SD motif, solid orange line) are compared to proximal occurrences of aSD:SD (dashed orange line), and plotted with the corresponding random models (blue lines).

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. 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).
4. 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).
5. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–258 (2009).

6. Yokobayashi, Y., Weiss, R. & Arnold, F. H. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci.* **99,** 16587–16591 (2002).
7. Shalem, O. *et al.* Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. *PLOS Genet* **11,** e1005147 (2015).
8. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30,** 521–530 (2012).
9. Buck, M., Gallegos, M.-T., Studholme, D. J., Guo, Y. & Gralla, J. D. The Bacterial Enhancer-Dependent σ^{54}(σN) Transcription Factor. *J. Bacteriol.* **182,** 4129–4136 (2000).
10. Bush, M. & Dixon, R. The Role of Bacterial Enhancer Binding Proteins as Specialized Activators of σ^{54}-Dependent Transcription. *Microbiol. Mol. Biol. Rev.* **76,** 497–529 (2012).
11. 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).
12. 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).
13. 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).
14. 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).
15. Feng, J., Goss, T. J., Bender, R. A. & Ninfa, A. J. Repression of the Klebsiella aerogenes nac promoter. *J. Bacteriol.* **177,** 5535–5538 (1995).

16. Keseler, I. M. *et al*. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res*. **41,** D605–D612 (2013).

17. Reitzer, L. & Schneider, B. L. Metabolic Context and Possible Physiological Themes of σ^{54}-Dependent Genes in Escherichia coli. *Microbiol. Mol. Biol. Rev.* **65,** 422–444 (2001).

18. 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).

19. 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).

20. 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).

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

22. 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

23. 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).

24. Bundschuh, R. & Hwa, T. RNA Secondary Structure Formation: A Solvable Model of Heteropolymer Folding. *Phys. Rev. Lett.* **83,** 1479–1482 (1999).

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

26. Hui, M. P., Foley, P. L. & Belasco, J. G. Messenger RNA Degradation in Bacterial Cells. *Annu. Rev. Genet.* **48,** 537–559 (2014).

27. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature* **451,** 355–358 (2008).

28. 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).

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

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

31. 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).

32. Na, D. & Lee, D. RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinforma. Oxf. Engl.* **26,** 2633–2634 (2010).

33. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chem. Chem. Mon.* **125,** 167–188

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

35. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–541 (2012).

36. Poulsen, L. D., Kielpinski, L. J., Salama, S. R., Krogh, A. & Vinther, J. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA* **21,** 1042–1052 (2015).

37. 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).

38. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **342,** 475–479 (2013).

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

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

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

42. Barrios, H., Valderrama, B. & Morett, E. Compilation and analysis of σ^{54}-dependent promoter sequences. *Nucleic Acids Res.* **27,** 4305–4313 (1999).

43. Wagner, F. The XL-mHG Test For Enrichment: A Technical Report. *ArXiv150707905 Stat* (2015).

44. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.* **10,** 2997–3011 (1982).

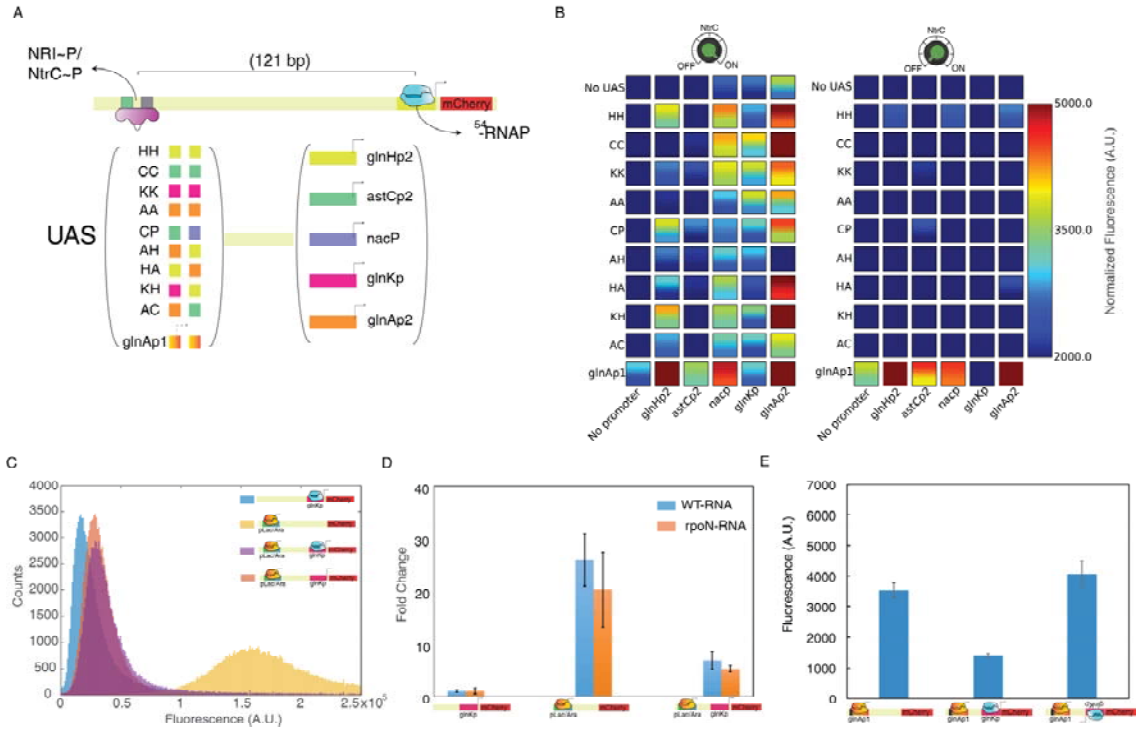

Figure 1

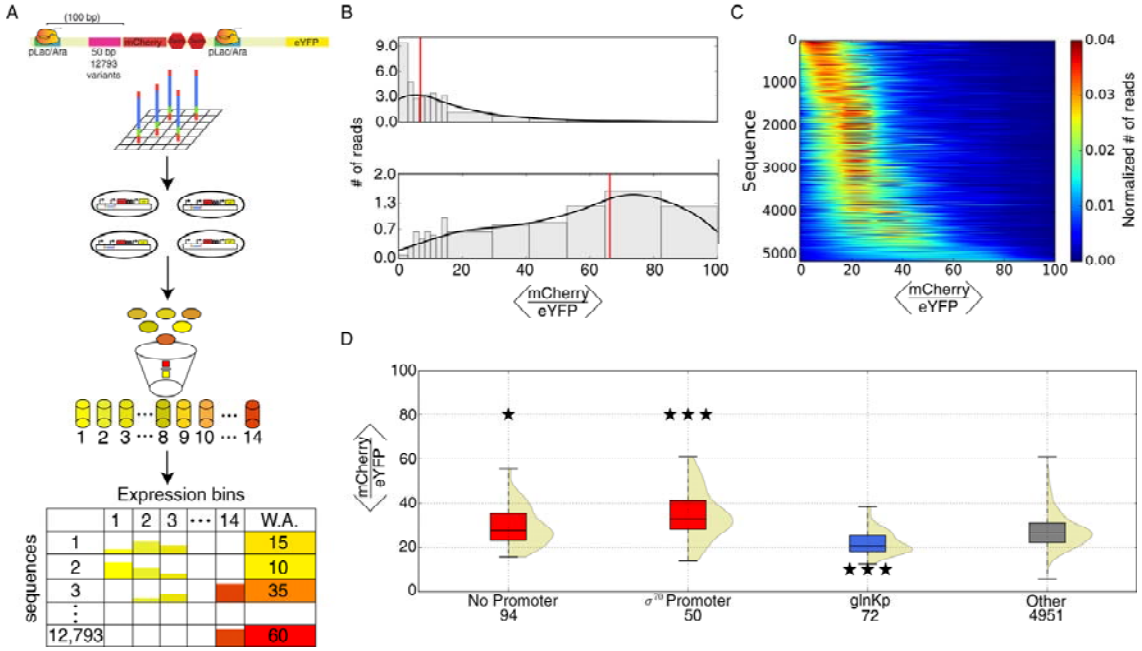

Figure 2

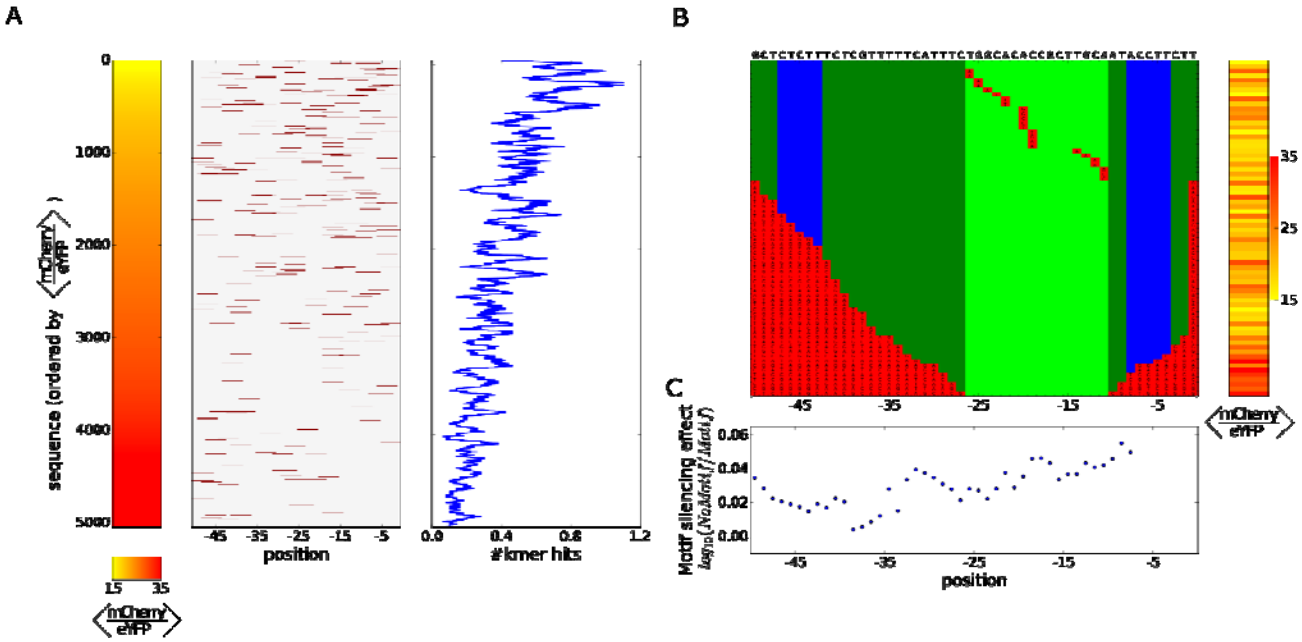

Figure 3

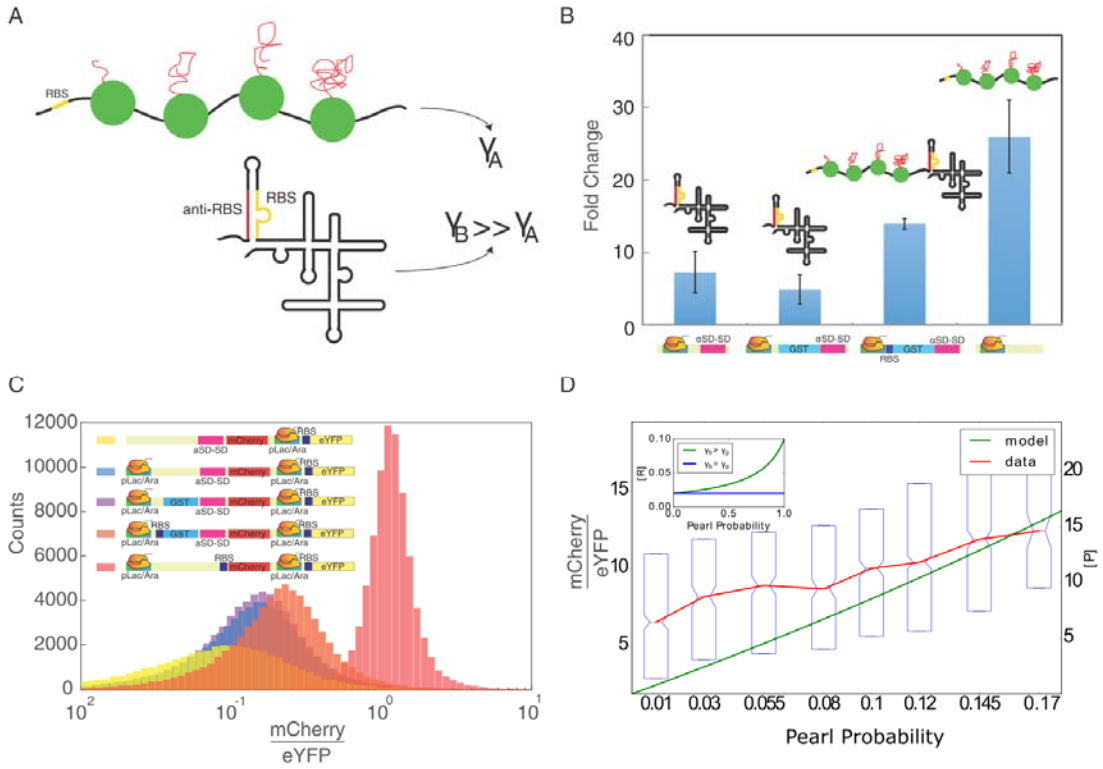

Figure 4

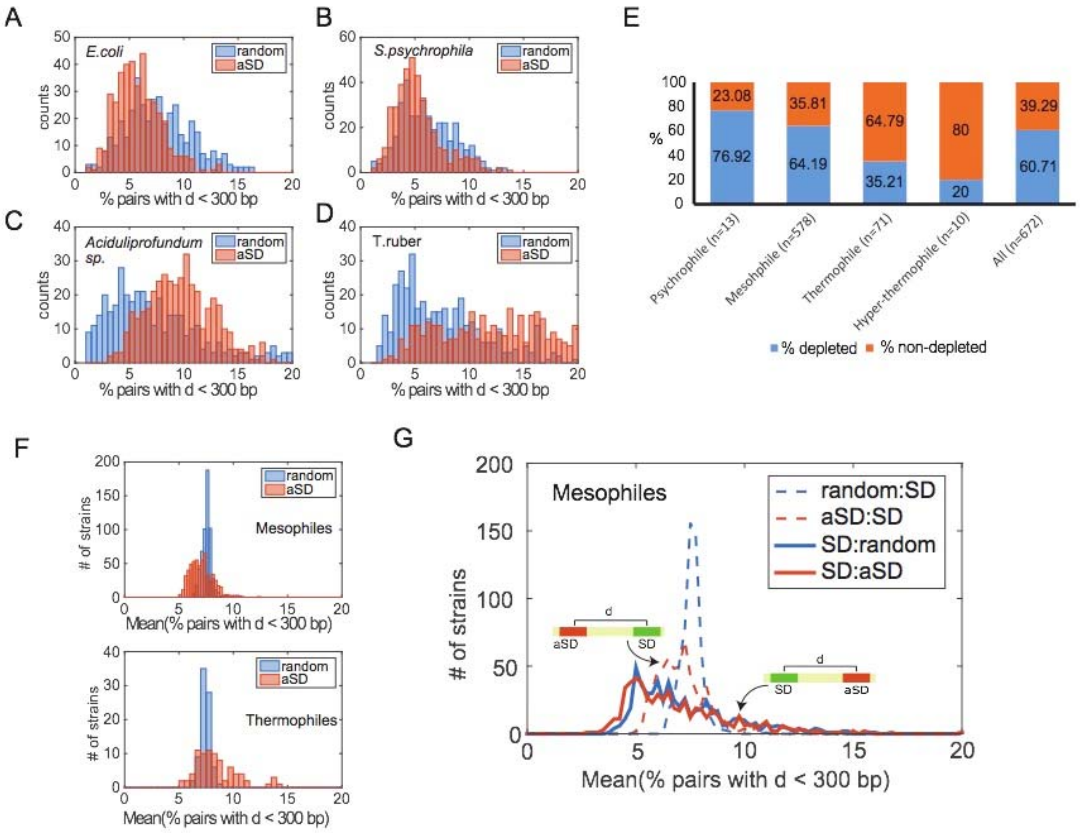

Figure 5