Chromosomes and Gene Expression
Evolutionary Biology

The emergence and evolution of gene expression in genome regions replete with regulatory motifs

Reviewed Preprint
v2 • December 5, 2024
Revised by authors

Reviewed Preprint
v1 • July 1, 2024

Timothy Fuqua, Yiqiao Sun, Andreas Wagner

Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland • Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland • The Sante Fe Institute, Sante Fe, USA

eLife Assessment

This **important** study explores the relationship between the sequence of prokaryotic promoter elements and their activity using mutagenesis to generate thousands of mutant sequences. The evidence supporting these findings is **convincing**. This work will appeal to those interested in bacterial genetics, genome evolution, and gene regulation.

https://doi.org/10.7554/eLife.98654.2.sa4

Abstract

Gene regulation is essential for life and controlled by regulatory DNA. Mutations can modify the activity of regulatory DNA, and also create new regulatory DNA, a process called regulatory emergence. Non-regulatory and regulatory DNA contain motifs to which transcription factors may bind. In prokaryotes, gene expression requires a stretch of DNA called a promoter, which contains two motifs called -10 and -35 boxes. However, these motifs may occur in both promoters and non-promoter DNA in multiple copies. They have been implicated in some studies to improve promoter activity, and in others to repress it. Here, we ask whether the presence of such motifs in different genetic sequences influences promoter evolution and emergence. To understand whether and how promoter motifs influence promoter emergence and evolution, we start from 50 “promoter islands”, DNA sequences enriched with -10 and -35 boxes. We mutagenize these starting “parent” sequences, and measure gene expression driven by 240’000 of the resulting mutants. We find that the probability that mutations create an active promoter varies more than 200-fold, and is not correlated with the number of promoter motifs. For parent sequences without promoter activity, mutations created over 1’500 new -10 and -35 boxes at unique positions in the library, but only ~0.3% of these resulted in de-novo promoter activity. Only ~13% of all -10 and -35 boxes contribute to de-novo promoter activity. For parent sequences with promoter activity, mutations create new -10 and -35 boxes in 11 specific positions that partially overlap with preexisting ones to modulate expression. We also find that -10 and -35 boxes do not repress promoter activity. Overall, our work demonstrates how promoter motifs influence

promoter emergence and evolution. It has implications for predicting and understanding regulatory evolution, de-novo genes, and phenotypic evolution.

Introduction

Gene regulation is critical to the development and homeostasis of living beings (Prud'homme et al., 2007). It is controlled by DNA sequences that encode motifs for transcription factors (TFs) to bind and either activate or repress transcription (Jacob and Monod, 1961; Seshasayee et al., 2011). Mutations to these motifs can modulate gene expression, leading to the evolution of new phenotypes (Fuqua et al., 2020; Prud'homme et al., 2007). Mutations can also create new regulatory sequences from non-regulatory DNA, a process called regulatory emergence (Rebeiz et al., 2011; Villar et al., 2015).

A regulatory motif is a DNA sequence computationally predicted to be compatible with TF binding. Although it is commonly thought that regulatory motifs influence both regulatory emergence and evolution, pertinent empirical evidence is sparse and conflicting. On the one hand, such evidence shows that motifs can facilitate regulatory emergence and evolution. For example, recent experiments showed that a randomly synthesized DNA with a single motif for a particular TF favored the emergence of tissue-specific gene expression for that TF (Galupa et al., 2023). Likewise, transposable elements contain motifs compatible with the binding of particular TFs (Lynch et al., 2015; Villanueva-Cañas et al., 2019), and these motifs have biased gene regulatory networks towards utilizing these TFs in evolution (Lynch et al., 2015). In *E. coli*, an active promoter is more likely to overlap with multiple promoter motifs than individual motifs (Huerta and Collado-Vides, 2003).

On the other hand, there is also evidence that regulatory motifs can impede regulatory emergence and evolution. One reason for this is that TFs can compete in their binding to shared motifs (Chauhan et al., 2022; Lagator et al., 2022). For example, the TF Histone-like nucleoid-structuring protein (H-NS) binds to sequences similar to prokaryotic promoters to repress spurious transcription and the expression of foreign DNA (Forrest et al., 2022; Oshima et al., 2006). Another reason is that gene expression can be silenced when regulatory DNA contains multiple motifs for the same TF (Berman et al., 2002; Bykov et al., 2020; Loell et al., 2024; Perales et al., 2016). For example, in mice, retinal "silencers" contains clusters of motifs for the TF CRX, while retinal "enhancers" contain a mixture of CRX and other TF motifs (Loell et al., 2024). A third reason is that clusters of motifs for RNA polymerase binding can cause collisions between RNA polymerase molecules, leading to premature termination of transcription (Crampton et al., 2006; Hobson et al., 2012; Wang et al., 2023).

A canonical prokaryotic promoter recruits the RNA polymerase subunit, σ70 to transcribe downstream sequences (Burgess et al., 1969; Huerta and Collado-Vides, 2003; Paget and Helmann, 2003; van Hijum et al., 2009). The "standard model" (Lagator et al., 2022) of such promoters is that σ70 binds to two motifs: the -10 box, and the -35 box (Huerta and Collado-Vides, 2003; Pribnow, 1975; Urtecho et al., 2019). These motifs are thoroughly characterized (consensus TATAAT and TTGACA, respectively) and are optimally spaced 17±1 bp apart, with transcription beginning 3 bp downstream of the -10 box (Belliveau et al., 2018; Ireland et al., 2020; Kinney et al., 2010; Urtecho et al., 2023, 2019).

This standard model, however, is an oversimplification of promoter architecture (Lagator et al., 2022). Studies have also identified an AT-rich motif upstream of the -35 box called the UP element (consensus: AAAWWTWTTTTnnAAAA) (Estrem et al., 1998; Ross et al., 1993), which can increase promoter activity (Bracco et al., 1989; Meng et al., 2001). However, only ~3-19% of all characterized promoters have an UP-element (Estrem et al., 1998). Recent work also identified an AT-rich motif within the spacer region associated with promoter activity (Warman et

al., 2020). Other characterized motifs include variations of the -10 box such as the *extended -10 box / TGn motif* (TGnTATAAT), which can sometimes compensate for an absent -35 box (Burr et al., 2000; Mitchell et al., 2003; Yona et al., 2018). A symmetrical -10 box variant also leads to bidirectional promoter activity (consensus: TATWATA) if correctly spaced between two flanking -35 boxes located on opposing strands (Warman et al., 2021). Like in humans (Gotea et al., 2010) and invertebrates (Lifanov et al., 2003), most prokaryotic promoters also have multiple regulatory motifs (Gama-Castro et al., 2016; Huerta and Collado-Vides, 2003) and transcription start sites (Mendoza-Vargas et al., 2009). It is not clear whether these additional motifs influence promoter activity. If so, it may help to explain why it is difficult to predict promoter strength and activity (Lagator et al., 2022; Urtecho et al., 2023).

Recent work shows that mutations can help new promoters to emerge from promoter motifs or from sequences adjacent to such motifs (Bykov et al., 2020; Fuqua and Wagner, 2023; Yona et al., 2018). However, encoding -10 and -35 boxes is insufficient to drive complete transcription of a gene coding sequence. For instance, the *E. coli* genome contains clusters of -10 and -35 boxes that are bound by RNA polymerase and produce short oligonucleotide fragments, but rarely create complete transcripts. Such clusters are called *promoter islands*, and are strongly associated with horizontally-transferred DNA (Bykov et al., 2020; Panyukov and Ozoline, 2013; Purtov et al., 2014; Shavkunov et al., 2009).

There are two proposed explanations for why promoter islands do not create full transcripts. First, the TF H-NS may repress promoter activity in promoter islands. This is because in a *Δhns* background, transcript levels from the promoter islands increases (Purtov et al., 2014). However, mutagenizing a specific promoter island (*appY*) until it transcribes a GFP reporter, reveals that in-vitro H-NS binding does not significantly change when GFP levels increase (Bykov et al., 2020). Thus, it is not clear whether H-NS actually represses the complete transcription of these sequences. The second proposed explanation is that excessive promoter motifs silence transcription. In the aforementioned study, promoter activity increases when mutations improve a -10 box to better match its consensus (TAAAAAT➔TATACT), while simultaneously destroying surrounding -10 and -35 boxes (Bykov et al., 2020). However, we note that if these surrounding motifs never contributed to GFP fluorescence to begin with, then mutations could also simply have accumulated in them during random mutagenesis.

In this study, we define a promoter as a DNA sequence that drives the expression of a (fluorescent) protein, whose expression level, measured by its fluorescence, is greater than a defined threshold. We use a threshold of 1.5 arbitrary units (a.u.) of fluorescence. This definition does not distinguish between transcription and translation. We chose it because protein expression is usually more important than RNA expression whenever natural selection acts on gene expression, because it is the primary phenotype visible to natural selection (Jiang et al., 2023).

We systematically study the ability of promoter motifs to influence the evolution and emergence of promoter activity. To this end, we select 25 150 bp *template sequences* with and without promoter activity (*P1-P25*), from 25 independent promoter islands. For each template, we treated both strands as independent *parent sequences* (N=50 parent sequences). We then screened over 240,000 mutant variants (daughter sequences) derived from these parent sequences for promoter activity. In non-promoter parent sequences, the emergence of new -10 and -35 boxes in daughter sequences rarely resulted in de novo promoter activity. Conversely, for parent sequences with existing promoter activity, the promoter function was only modulated when new -10 and -35 boxes formed in daughter sequences which partially overlapped with existing boxes. Notably, we found no evidence that -10 and -35 boxes directly repress promoter activity, but H-NS does repress some of the promoter islands.

Results

Promoter islands are enriched with motifs for -10 and -35 boxes

Promoter islands are ~1kbps (kilo base pairs) long (Bykov et al., 2020; Panyukov and Ozoline, 2013; Purtov et al., 2014; Shavkunov et al., 2009), but for adequate mutational coverage, it is better to work with shorter sequences. We thus identified within 25 promoter islands, a region of 150 bp that showed the highest density of predicted -10 and -35 boxes for further analysis (**Fig 1A**). To this end, we used position-weight matrices (PWMs) – computational tools to identify these promoter motifs (Methods) (Hertz and Stormo, 1999). (These PWMs are derived from 145 of the top predicted -10 and -35 boxes from the database RegulonDB (Tierrafría et al., 2022), and are highly correlated with experimentally-derived - 10 and -35 boxes (Belliveau et al., 2018). See methods.) We refer to these as *template sequences* (P1-25), and treat each genetic strand as an independent *parent sequence* (N=50 parents). On average, each parent sequence contains ~5.32 -10 boxes and ~7.04 -35 boxes (**Fig S1**). 18 of these -10 boxes also include the TGn motif upstream of the hexamer.

Non-promoters vary widely in their potential to become promoters

Next we created a mutant library of *daughter* sequences from the 25 templates using an error-prone polymerase chain reaction (**Fig 1B**), and measured the extent to which these daughter sequences can drive gene expression using the well-established sort-seq procedure (Kinney et al., 2010; Peterman and Levine, 2016). Specifically, we cloned the library into a bidirectional fluorescence reporter plasmid (pMR1) to measure promoter activity on both DNA strands (top: GFP, bottom: RFP) (Guazzaroni and Silva-Rocha, 2014; Westmann et al., 2018). We then transformed the plasmid library into *E. coli* (DH5α, Takara Japan), sorted the bacteria using fluorescence activated cell sorting (FACS) into eight fluorescence bins (none, weak, moderate, and strong fluorescence, for both GFP and RFP) (**Fig 1C**, see also **Fig S2**), and sequenced the inserts from the sorted fluorescence bins (see methods).

Our library comprised 245'639 unique daughter sequences, with an average of 4'913 daughter sequences per template and ~2.5-point mutations per daughter (**Fig S2**). Each template has 3×150 possible mutational neighbors that differ by a single mutation. We define coverage as the percentage of these 450 bases and positions that are present in at least one daughter sequence for each parent. (Daughters contain 1-10 point mutations.) The mean coverage was 93% (minimum: 80%; maximum: 100%) (**Fig S3**). The mutation library therefore encompasses the majority of neighboring mutations for the 25 templates.

We next quantified how strongly each daughter sequence could drive gene expression from the top and bottom strand by calculating a fluorescence score for both GFP and RFP expression, respectively. These scores are bounded between a lowest score of 1.0 arbitrary units (a.u.), indicating no expression, and a highest score of 4.0 a.u., indicating maximal expression (methods). We declared a daughter sequence to have *promoter activity* or to be a *promoter* if its score was greater than or equal to 1.5 a.u., as this score lies at the boundary between no fluorescence and weak fluorescence based on the sort-seq bins (methods). Otherwise, we refer to a daughter sequence as having *no promoter activity* or being a *non-promoter*.

Because some sequences in our library are unmutated parent sequences, we determined that 10/50 of the parent sequences already encode promoter activity before mutagenesis. Specifically, three parents drove expression on the top strand (P19-GFP, P20-GFP, P22-GFP), and five did on the

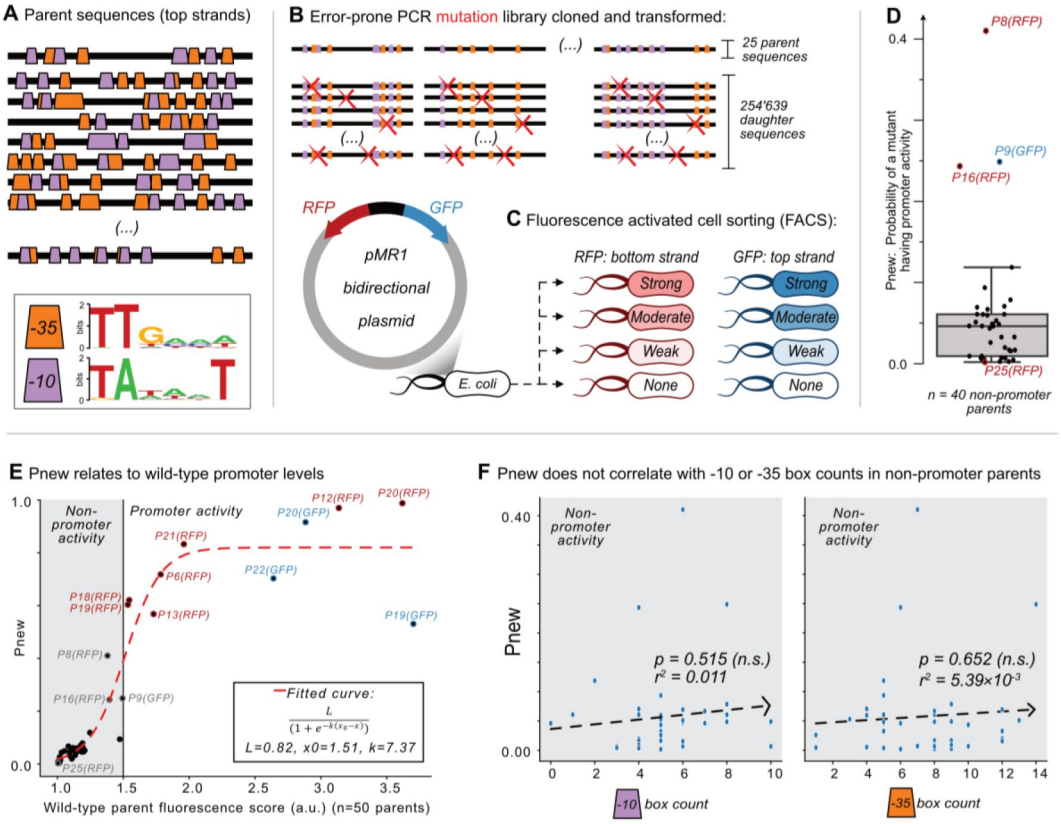


Figure 1.

Mutagenesis reveals vastly different probabilities of promoter emergence.

(A) The location of -10 and -35 boxes in a subset of the parent sequences. See **Fig S1** for the complete set (n=50). Orange trapezoids correspond to -35 boxes, and magenta trapezoids to -10 boxes, each identified using position-weight matrices (see sequence logos below and **Fig S1**). **(B-C)** The Sort-Seq protocol. **(B)** Top: we amplified 25 150bp template sequences from 25 promoter islands in the *E. coli* genome with an error-prone polymerase chain reaction, generating a mutagenesis library of 245'639 unique daughter sequences. Bottom: we cloned the library into the pMR1 reporter plasmid between a green fluorescent protein (GFP) coding sequence on the top strand (blue arrow) and a red fluorescent protein (RFP) coding sequence on the bottom strand (red arrow). We transformed the plasmid library into *E. coli.* **(C)** Using fluorescence-activated cell sorting (FACS), we sorted the transformed *E. coli* cells into 8 fluorescence bins: none, weak, moderate, and strong, for both RFP and GFP expression (see **Fig S2** for bins). We sequenced the plasmid inserts of cells from each bin, assigning a fluorescence score from 1.0-4.0 arbitrary units (a.u.), ranging from no fluorescence (1.0 a.u.) to strong fluorescence (4.0 a.u.) (see **Fig S4** for score distributions). **(D)** Pnew is the ratio of daughter sequences with a fluorescence score greater than 1.5 a.u. to the total number of daughter sequences for each parent. Circles show the Pnew values for each of 40 non-promoter parents (i.e., 50 parents – 3 top strand promoters – 7 bottom strand promoters = 40 non-promoter parents.) The height of the box represents the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR. **(E)** A scatterplot between the calculated fluorescence score (methods) for all of the parent sequences (n=50, both with and without promoter activity) and Pnew. The solid horizontal line at 1.5 a.u. marks our cutoff for whether a parent does not (gray shaded region) or does have promoter activity (no shading). The dashed line corresponds to a fitted sigmoidal curve, where L is equal to the upper asymptote, x0 is the inflection point, and k is the slope. See methods for fitting details. **(F)** Scatterplots comparing the number of -10 or -35 box counts per parent sequence and their respective Pnew values. The dashed line is the line of best fit calculated using the method of least squares. We test the null hypothesis that the slope of the correlation equals zero with the Wald Test. The r^2 value is the Pearson correlation coefficient. Calculation carried out using scipy.stats.linregress (version 1.8.1). Left: the number of -10 box counts (p=0.515, r^2=0.011). Right: the number of -35 box counts (p=0.652, r^2=5.39×10^{-3}).

bottom strand (P6-RFP, P12-RFP, P13-RFP, P18-RFP, P19-RFP, P20-RFP, P21-RFP). Two parents harbor bidirectional promoters (P19 and P20). The remaining 40 parent sequences are non-promoters, with an average fluorescence score of 1.39 a.u. We note that some of these parents have a fluorescence score higher than 1.39 a.u., but less than 1.50 a.u. such as P8-RFP (1.38 a.u.), P16-RFP (1.39 a.u.), P9-GFP (1.49 a.u.), and P1-GFP (1.47 a.u.). Whether these are truly "promoters" or not, is based solely on our threshold value of 1.5 a.u. We also note that 30% (15/50) of parents have the TGn motif upstream of a -10 box, but only 20% (3/15) of these parents have promoter activity (underlined with promoter activity: P4-RFP, P6-RFP, P8-RFP, P9-RFP, P10-RFP, P11-GFP, P12-GFP, P17-GFP, P18-GFP, P18-RFP, P19-RFP, P22-RFP, P24-GFP, P25-GFP, P25-RFP). See **Fig S4** for fluorescence score distributions for each parent and its daughters, and Data S2 for all daughter sequence fluorescence scores.

Promoter emergence correlates with minute differences in background promoter levels

Although mutating each of the 40 non-promoter parent sequences could create promoter activity, the likelihood P_{new} that a mutant has promoter activity, varies dramatically among parents. For each non-promoter parent, **Fig 1D** shows the percentage of active daughter sequences. The median P_{new} is 0.046 (std. ± 0.078), meaning that ~4.6% of all mutants have promoter activity. The lowest P_{new} is 0.002 (P25-GFP) and the highest 0.41 (P8-RFP), a 205-fold difference.

We hypothesized that these large differences in P_{new} could be explained by minute differences in the fluorescence scores of each parent, particularly if its score was below 1.5 a.u. Plotting the fluorescence scores of each parent (N=50) and their respective P_{new} values as a scatterplot (**Fig 1E**), we can fit these values to a sigmoid curve (see methods). This finding helps to explain why P8-RFP has a high P_{new} (0.41) and P25-GFP a low P_{new} (0.002), as their fluorescence scores are 1.380 and 1.009 a.u., respectively. The fact that the inflection point of the fitted curve is at 1.51 a.u. further justifies our use of 1.5 a.u. as a cutoff for promoter and non-promoter activity.

Promoter emergence does not correlate with simple sequence features

We asked what else could explain the differences in P_{new} values for the non-promoter parents. To this end, we calculated linear regressions between P_{new} values of the non-promoter parents and various sequence features. These include the number of daughter sequences, GC-content, -10 box counts and -35 box counts. We found that P_{new} does not significantly correlate with any of these features, including the number of -10 or -35 boxes (least squares linear regression, p=0.529, 0.930, 0.515, 0.652 respectively) (**Fig 1F,G** and **Fig S5A,B**).

We then asked whether any k-mers ranging from 1-6 bp correlated with the non-promoter P_{new} values (5,460 possible k-mers). 718 of these 1-6 bp k-mers are present 3 or more times in at least one non-promoter parent. We calculated a linear regression between the frequency of these 718 k-mers and each P_{new} value, and adjusted the p-values to respective q-values (Benjamini-Hochberg correction, FDR=0.05). This analysis revealed six k-mers: CTTC, GTTG, ACTTC, GTTGA, AACTTC, TAACTT which correlate with P_{new}. However, these correlations are heavily influenced by an outlying P_{new} value of 0.41 (P8-RFP) (**Fig S5C-H**), and upon removing P8-RFP from the analysis, no k-mer significantly correlates with P_{new} (data not shown).

Promoters emerge and evolve only from specific subsets of -10 and -35 boxes

Next, we asked where both existing and new promoters are located within the parent sequences. To this end, we first calculated for each parent and its respective daughters, the mutual information $I_i(b, f)$ between the nucleotide identity at each position i and the corresponding

fluorescence level *f* (**Fig 2A**). The essence of this calculation is to determine for every position *i*, the probability $p_i(b)$ of that position being an A,T,C, or G and the probability of each sequence having a (rounded) fluorescence score of 1,2,3, or 4 a.u. $p(f)$. Representing these probabilities as circles in a Venn-diagram, the overlap between the circles is the joint probability $p_i(b, f)$ that position *i* has a specific base *b* AND a particular fluorescence score *f*. The larger this joint probability is compared to the individual probabilities, the more important the nucleotide identity at position *i* is for fluorescence, i.e., for promoter activity. This calculation can reveal positions where RNA polymerase and transcription factors bind (Belliveau et al., 2018; Ireland et al., 2020; Kinney et al., 2010), as well as DNA regions where promoters most readily emerge (Fuqua and Wagner, 2023). The calculation effectively converts all 245'639 data points into maps that reveal where promoters either already exist or emerge in each of the parent sequences (see methods).

We identified 'hotspots' – regions of especially high mutual information – for each parent sequence (Fuqua and Wagner, 2023) (see methods and Data S3). For example, two such hotspots exist in parent P19-GFP, which is an active promoter that drives expression from both DNA strands (**Fig 2B**). One hotspot overlaps with a -35 box and the other overlaps with a -10 box. These hotspots indicate that fluorescence changes when these DNA sequences mutate. Because these sequences additionally match a canonical promoter, they are the likely cause of the promoter activity in P19-GFP (See **Fig S10A-B'** for further validation).

Across all 50 parent sequences, we identified a total of 68 hotspots (**Fig 2C**). 34 hotspots occur in the parent sequences with promoter activity. They correspond to locations where mutations either increase or decrease promoter activity. ~15% (5/34) of these hotspots overlap exclusively with -10 boxes, ~32% (11/34) with -35 boxes, and ~18% (6/34) overlap with both - 10 and -35 boxes. The remaining ~35% of hotspots overlap with neither a -10 nor a -35 box (12/34). To find out whether these overlaps could be explained by chance alone, we computationally scrambled the parent sequences while maintaining the hotspot coordinates. ~9% (3/34) of the hotspots overlapped with -10 boxes in the scrambled parents, ~12% (4/34) with -35 boxes, ~3% (1/34) with both -10 and -35 boxes, and the remaining ~76% (26/34) overlapping with neither. Hotspots in the biological parent sequences overlap to a significantly greater extent with the two classes of boxes than the scrambled parent sequences (chi-squared test, 3 degrees of freedom, χ^2=46.1, p=5.34×10^{-10}). This analysis shows that the overlap does not simply result from the high density of -10 and -35 boxes.

The remaining 34 hotspots occur in non-promoter parent sequences. Mutations in these locations solely increase fluorescence. Of these 34 hotspots, ~26% (9/34) overlap with -10 boxes, ~26% (9/34) with -35 boxes, ~15% with both -10 and -35 boxes (5/34), and ~32% with neither motif (11/34) (**Fig 2C-E**). Scrambling the parent sequences while maintaining hotspot coordinates resulted in ~24% (8/34) of hotspots overlapping with a -10 box, ~9% (3/34) with a -35 box, ~6% (2/34) with both -10 and -35 box, and the remaining ~62% (21/34) with neither. Again, these percentages of overlap with existing boxes are significantly greater for biological than for scrambled parent sequences (chi-squared test, 3 degrees of freedom, χ^2=21.4, p=8.75×10^{-5}). There is no significant difference between the motif overlap of promoters and non-promoters (chi-squared test, 3 degrees of freedom, χ^2=2.51, p=0.473). See **Fig S6** for mutual information plots for all parent sequences and Data S3 for a table with the hotspots and their coordinates.

Despite this extensive overlap between promoter motifs and hotspots, each parent contains additional -10 and -35 boxes that do not overlap with hotspots. Of the 266 -10 boxes, only 37 (~14%) overlap within ±3 bp from a hotspot. Similarly, of the 352 -35 boxes, only 43 (~12%) overlap with a hotspot. For example, the non-promoter parent P3-RFP contains five -10 boxes, but only one overlaps with a hotspot (**Fig 2D**). Similarly, the non-promoter P18-GFP contains ten -10 and nine -35 boxes, but promoters only emerge in three hotspots of P18-GFP (**Fig 2E**).

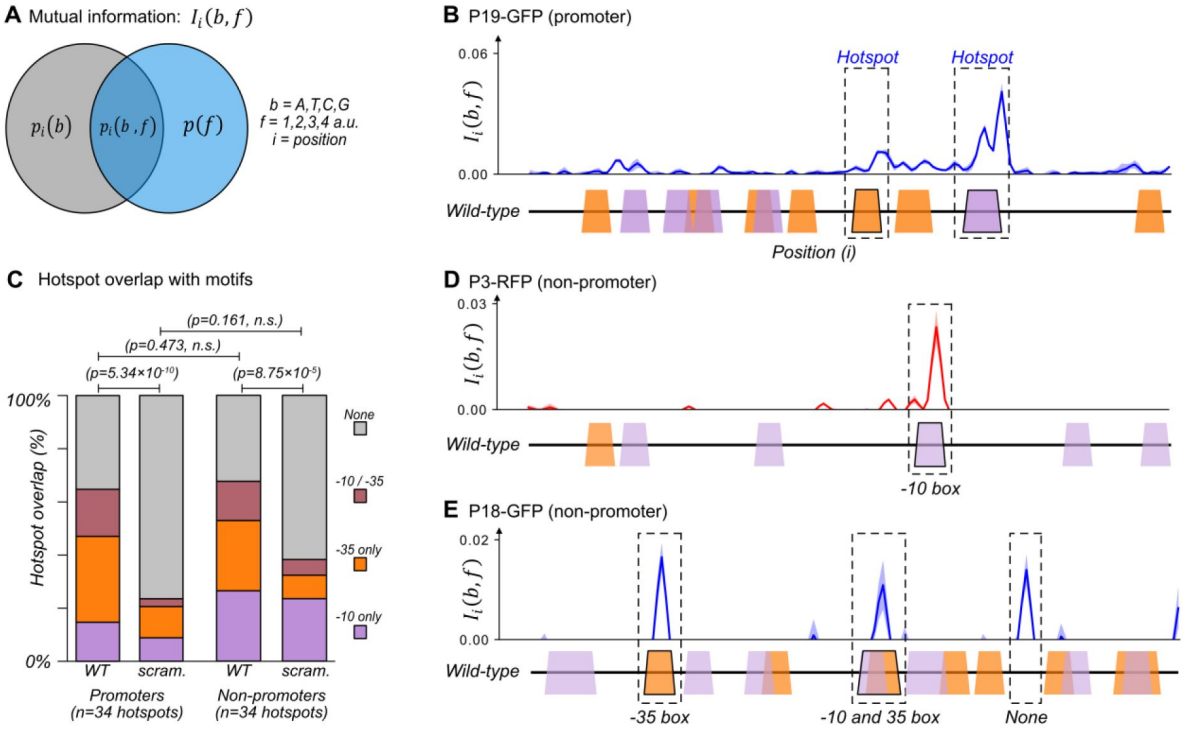


Figure 2.

The majority of promoters emerge and evolve within a subset of preexisting promoter motifs.

(A) We calculated the mutual information $I_i(b, f)$ between nucleotide identity (b = *A,T,C,G*) and fluorescence scores rounded to the nearest whole number (f = *1,2,3,4 a.u.*) for each position i in a parent sequence. In essence, the calculation compares the probability $p_i(b)$ of a base b occurring at position i, and the probability $p(f)$ that a sequence has a fluorescence score f. The joint probability $p_i(b, f)$ is the probability that a sequence with base b at position i has fluorescence score f. The greater the joint probability is compared to the individual probabilities, the more important the base at this position is for promoter activity. See methods. **(B)** An example of how to interpret mutual information using position-weight matrix (PWM) scores of predicted -10 and -35 boxes. Top: we plot the mutual information $I_i(b, f)$ for P19-GFP. P19-GFP is an active promoter. Solid line: mean mutual information. Shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Bottom: position-weight matrix (PWM) predictions for the -10 boxes (magenta trapezoids) and -35 boxes (orange trapezoids) along the wild-type parent sequence. We define hotspots as mutual information peaks greater than or equal to the 90th percentile of total mutual information (methods), and highlight them with dashed rectangles. **(C)** Stacked bar plots depicting the percentage of hotspots overlapping with -10 boxes only (magenta), -35 boxes (orange), both -10 and -35 boxes (red), or with neither (gray). We plot this information for the wild-type (WT) promoters and non-promoter parents, as well as scrambled (scram.) versions of the parents. Horizontal lines correspond to χ (chi)-squared tests between the counts in each group (3 degrees of freedom). **(D)** Analogous to (B) but for parent P3-RFP. P3-RFP is an inactive parent sequence. Hotspot overlaps with a -10 box. **(E)** Analogous to (B) but for parent P18-GFP. P18-GFP is an inactive parent sequence. Hotspots overlap with (from left to right) a -35 box, both a -10 and a -35 box, and neither (None). **Fig S6** shows analogous mutual information plots for daughters derived from each parent sequence.

In sum, while the majority (~67%) of new promoters evolve and emerge from -10 and -35 boxes, the presence of one such box does not indicate that promoter activity can emerge from it, or is encoded within it. Only ~13% of all -10 and -35 boxes in the parents contribute to promoter activity.

New -10 and -35 boxes readily emerge, but rarely lead to de-novo promoter activity

We hypothesized that promoters emerge when mutations create new -10 and -35 boxes. To test this hypothesis, we examined each hotspot and used PWMs to find out whether a subset of the daughter sequences gains a -10 or -35 box in the hotspot. We then asked whether gaining a -10 or -35 box leads to a significant increase in fluorescence that indicates the creation of a new promoter (Mann-Whitney U-test, Fig. S7 and methods). See Data S4 for the coordinates, fluorescence changes, and significance for the -10 and -35 boxes.

Mutations indeed created many new -10 and -35 boxes in our daughter sequences. On average, 39.5 and 39.4 new -10 and -35 boxes emerged at unique positions within the daughter sequences of each mutagenized parent (**Fig 3A,B**), with 1'562 and 1'576 new locations for -10 boxes and -35 boxes, respectively. ~22% (684/3'138) of these new boxes are spaced 15-20 bp away from their cognate box, and ~7.3% (114/1'562) of the new -10 boxes have the TGn motif upstream of them. However, only a mere 5 of the new -10 boxes and 4 of the new -35 boxes are significantly associated with increasing fluorescence by more than +0.5 a.u. (**Fig 3C,D**). That is, only ~0.3% of all new -10 and -35 boxes actually create new promoter activity. Thus, the creation of a -10 or -35 box, even in the appropriate proximity to the cognate box, rarely leads to de-novo promoter activity.

Promoters can emerge when mutations create motifs but not by destroying them

We briefly highlight two examples where gaining -10 and -35 boxes leads to de-novo promoter activity. Parent P16-RFP gains promoter activity when mutations create a -10 box 17 bp downstream of a preexisting -35 box (**Fig 3E**), increasing median fluorescence by +1.99 a.u. (**Fig 3F**) (two-tailed MWU test, $q=7.64\times10^{-32}$). The majority of these -10 boxes emerge when the sequence TAAACA (0.44 bits) mutates to TAAACt (7.28 bits). To validate this observation experimentally, we created this point mutation in a P16-RFP reporter, and compared its fluorescence levels with the wild-type P16-RFP using a plate reader (see methods). This experiment shows that fluorescence increases ~11.2-fold when the -10 box emerges (**Fig 3F'**).

We also highlight one hotspot where gaining a -35 box creates de-novo promoter activity. Specifically, parent P1-RFP gains promoter activity when mutations create a -35 box 18 bp upstream of a preexisting -10 box (**Fig 3G**), which increases fluorescence by +0.74 a.u. (**Fig 3H**, two-tailed MWU test, $q=4.25\text{x}10^{-42}$). The majority of new -35 boxes emerge when the sequence ATGAGT (0.00 bits) mutates to tTGAGT (3.93 bits). A validation construct with this point mutation shows a ~2.0-fold increase in fluorescence compared to the wild-type (**Fig 3H'**).

Fig S8 shows additional examples in which -35 and -10 boxes are created, and the results of their validation experiments. In two of these hotspots, our validation experiments revealed no substantial difference in gene expression as a result of the hotspot mutation (**Fig S8F'** and **Fig S8J'**). In both of these false positives, new -10 boxes emerge in locations without an upstream -35 box.

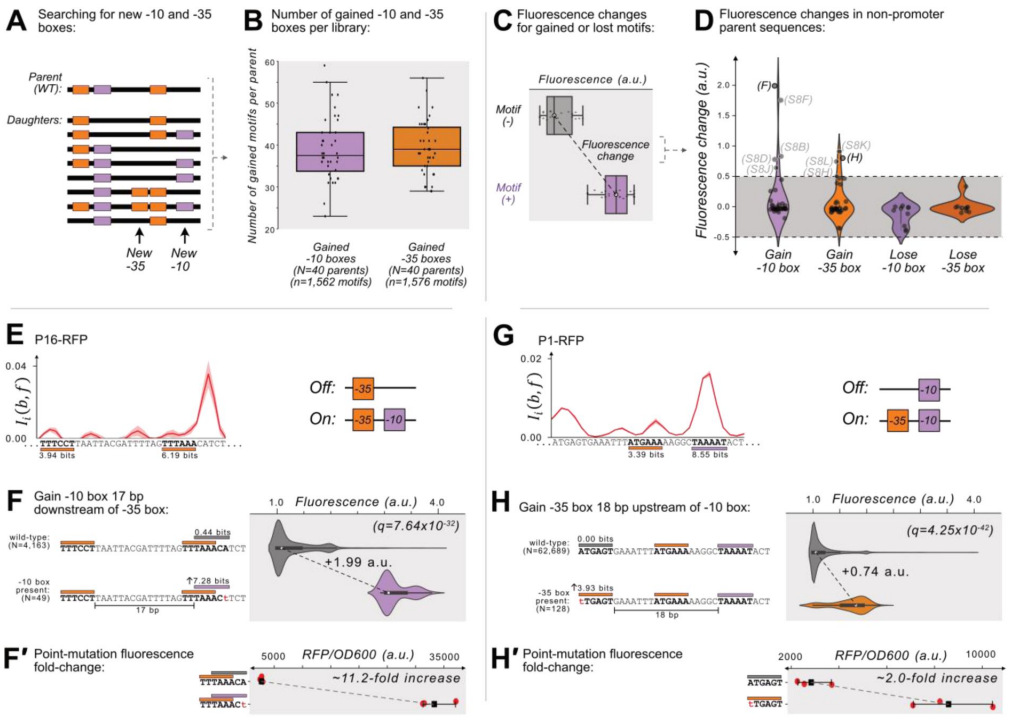


Figure 3.

Gaining -10 and -35 boxes rarely creates de-novo promoters.

(A) A cartoon for how we identify new -35 and -10 boxes (orange and magenta boxes, respectively) within each DNA sequence (thick, black, horizontal bars). Top: we first identify the -10 and -35 boxes in the wild-type parent sequence. Bottom: within all of the daughter sequences, we identify all locations where new -10 and -35 boxes appear, as shown with the arrows below the daughter sequences. We then count the new motifs gained for each parent, and plot the results in a box plot (**B**) for both new -10 and -35 boxes. Boxes represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR. **(C)** We define fluorescence change as the median difference between the fluorescence scores of sequences with vs without a -10 or -35 box. The fluorescence change is only considered valid if there is a significant difference between the central tendency of each distribution based on a two-sided Mann-Whitney U (MWU) test. See methods. **(D)** The change in fluorescence (arbitrary units, a.u.) when gaining or losing -10 and -35 boxes in mutual information hotspots of inactive parent sequences (see **Fig S7** for calculation overview). Dashed lines indicate an effect size threshold of ± 0.5 a.u. Each circle corresponds to a gain or loss of a box in a mutual information hotspot (see **Fig 2** for hotspots). Circles with letters in parentheses refer to the corresponding figure panels. The volume of each violin plot corresponds to a kernel density estimate of each distribution. Data available in Data S4. **(E)** Parent P16-RFP. Top: Mutual information $I_i(b, f)$ between nucleotide identity and fluorescence at each position. Solid red line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Bottom: position-weight matrix (PWM) predictions for the -10 boxes (magenta rectangles) and -35 boxes (orange rectangles) along the wild-type parent sequence. **(F)** Top: Parent P16-RFP and its PWM predictions from (E). We plot the fluorescence scores of all daughters without a -10 box in the region of interest (left, gray rectangle). Bottom: the most frequent genotype in the dataset where a -10 box is in the region of interest. We plot the fluorescence scores of all daughters with a -10 box in the region of interest. We tested the null hypothesis that the gain of the -10 box significantly increases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=7.64\times10^{-32}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(F')** Top: the fluorescence readouts of three colonies harboring the wild-type P16-RFP reporter measured using a plate reader. The horizontal axis shows the RFP readout normalized to the optical density of the culture (OD_{600}) during the reading (see methods). Each point corresponds to the fluorescence of an individual colony. Whiskers correspond to the minimum and maximum values and the dark square the mean fluorescence level. Bottom: analogous to top, but for three colonies harboring a single point mutation identified to change fluorescence in panel F. **(G)** Analogous to (E) but for the parent P1-RFP. **(H)** Analogous to (F) except for gaining a -35 box in the region of interest (two-tailed MWU test, $q=4.25\times10^{-42}$. **(H')** Analogous to F' but for the point mutation identified to change fluorescence in panel H. See **Fig S8** for additional examples.

To summarize, mutations readily create new -10 boxes and -35 boxes (see **Fig 3A-B**). This can create de-novo promoter activity (see **Fig 3E-H** and **Fig S8**), but it rarely does. Mutations also destroy -10 and -35 boxes originally in the parent sequences, but this does not create de-novo promoter activity (see **Fig 3D**).

Destroying -10 and -35 boxes in promoters exclusively lowers promoter activity

We next asked how gaining and losing -10 and -35 boxes can change expression driven by parent sequences that have promoter activity. To this end, we repeated the previous analysis but for the parents with promoter activity. See Data S4 for the coordinates, fluorescence changes, and significance for the -10 and -35 boxes.

We first identified 5 and 3 hotspots in which losing a -10 box or -35 box *decreases* expression (**Fig 4A**). These particular hotspots revealed where promoters are located in each parent. For example, for parent P12-RFP (**Fig S9C**), losing a -35 box decreases fluorescence by -1.57 a.u. (two-tailed MWU test, $q=6.28\times10^{-67}$) (**Fig S9D**), and losing a -10 box decreases fluorescence by -1.41 a.u. (two-tailed MWU test, $q=6.28\times10^{-91}$) (**Fig S9E**). Scrambling these motif sequences also lowers fluorescence by 62.6-fold and 49.7-fold respectively (**Fig S9D'** and **Fig S9E'**). Based on these data and the spacing (17 bp) of these two motifs, we can infer that they constitute the promoter in P12-RFP. See Fig. S9 for additional examples. We did not find any hotspots in which destroying a -10 box or a -35 box increases expression by more than 0.5 a.u. (see **Fig 4A**).

Gaining new motifs over existing motifs increases and decreases promoter activity

Our analysis found that across the 10 parents with promoter activity, 329 new -10 boxes and 332 new -35 boxes are created at unique positions within their daughter sequences. Despite this extensive gain of new boxes, only 4 of these new -10 boxes and 1 of these new -35 boxes further increases expression. More surprisingly, for 4 and 6 of the new -10 and -35 boxes, gaining the box *decreases* expression (n= 4+1+4+6, 15 hotspots with increased or decreased expression) (**Fig 4E**). To understand how this is possible, we examined each of these hotspots in detail.

We found that these mutations frequently create new boxes overlapping those we had identified as part of a promoter (**Fig S9**). This occurs when mutations create a -10 box overlapping a -10 box, a -35 box overlapping a -35 box, a -10 box overlapping a -35 box, or a -35 box overlapping a -10 box. We call the resulting event a “homo-gain” when the new box is of the same type as the one it overlaps, and otherwise a “hetero-gain”. In either case, the creation of the new box does not always destroy the original box.

6 of the 15 hotspots (40%) show homo-motif gains. For example in P12-RFP, a point mutation creates both a new -10 box that is +1 bp further from the -35 box, but also increases the PWM score of the original -10 box to match the consensus (TACAAT➔ TAtAAT). This point mutation increases median expression by +0.78 a.u. (**Fig 4D**) (two-tailed MWU test, $q=5.85\times10^{-13}$). A construct harboring this point mutation also exhibits a 0.3-fold increase in fluorescence compared to the wild-type (**Fig 4D'**). 3 of the 15 hotspots (20%) show hetero-motif gains. For example, for parent P22-GFP, where our mutational data had allowed us to map the active - 35 and -10 boxes (**Fig 4F**), mutations can change the -35 box into a -10 box, which decreases expression by -0.90 a.u. (**Fig 4G**) (two-tailed MWU test, $q=7.29\times10^{-7}$). A construct harboring this point mutation also exhibits a 13-fold decrease in fluorescence compared to the wild-type (**Fig 4G'**). Thus, while the loss in expression appears to be caused by gaining a -10 box, this gain also entails the loss of the promoter’s -35 box (see **Fig 4F** and **Fig S9J'**). See **Fig S10** for additional examples of the other homo- and hetero-motif gains.

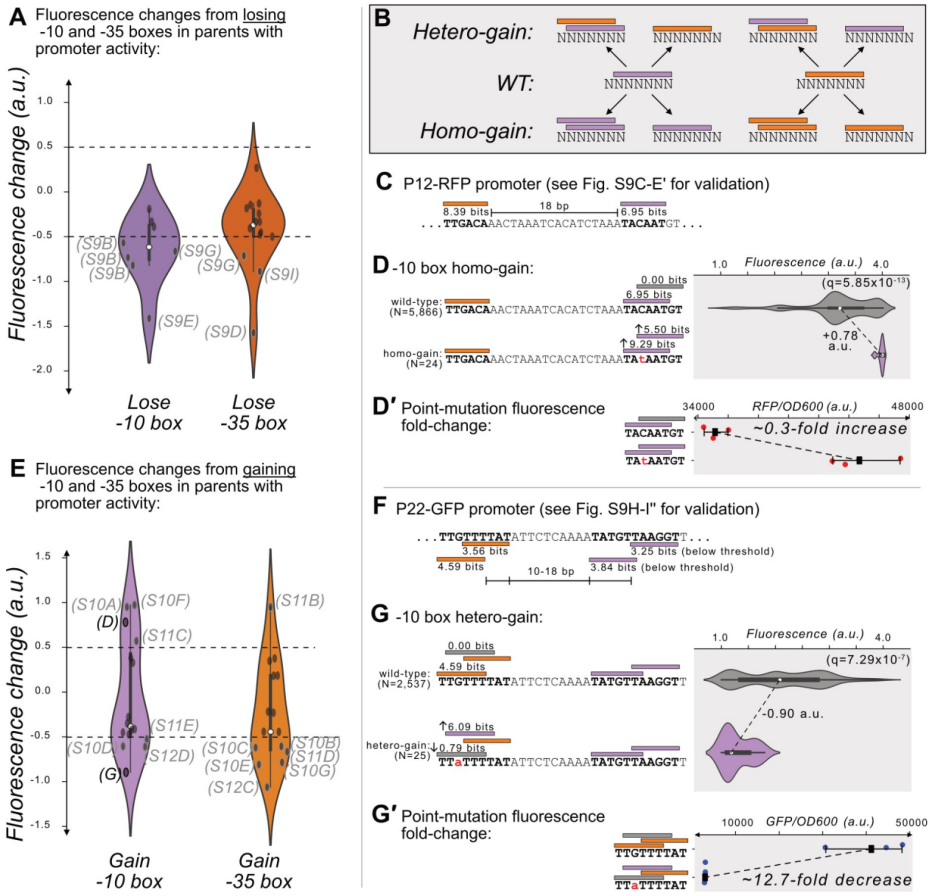


Figure 4.

Gaining -10 and -35 boxes modulates promoter activity.

(**A**) The change in fluorescence (arbitrary units, a.u.) when losing -10 and -35 boxes in hotspots in the active parent sequences. Dashed lines indicate an effect size threshold of ± 0.5 a.u. Each black point corresponds to a gain or loss of a -10 or -35 box in a mutual information hotspot. Outlined points with letters in parenthesis highlight the corresponding figure panels. The areas of the violin plots are the kernel density estimates (KDE) of each distribution. Within each violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. Data available in Data S4**. (B)** In parents with promoter activity, mutations frequently create new -10 (magenta rectangles) and -35 (orange rectangles) boxes over preexisting ones. (**C**) A model of the promoter located in parent P12-RFP. Position-weight matrix (PWM) predictions for the -10 boxes (magenta rectangles) and -35 boxes (orange rectangles) along with the wild-type parent sequence. See Fig. S9C-E' for the experiments validating this promoter. **(D)** Top: Parent P12-RFP and its PWM predictions from (C). We plot the fluorescence scores of all daughters without a -10 box in the region of interest (left, gray rectangle). Bottom: the most frequent genotype in the dataset where a -10 box is in the region of interest. We plot the fluorescence scores of all daughters with a -10 box in the region of interest. We tested the null hypothesis that the gain of the -10 box significantly increases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=5.85\times10^{-13}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(D')** Top: the fluorescence readouts of three colonies harboring the wild-type P12-RFP reporter, whose activity was measured using a plate reader. The horizontal axis shows red fluorescence normalized to the optical density of the culture (OD_{600}) during the reading (see methods). Each point corresponds to the fluorescence of an individual colony. Whiskers correspond to the minimum and maximum values and the dark square indicates the mean fluorescence level. Bottom: analogous to the top, but for three colonies harboring a point mutation identified to change fluorescence in the preceding panel. **(E)** Analogous to (A) but for gaining -10 and -35 boxes instead of losing them. **(F)** Analogous to (C) but for P22-GFP. Note that the two -10 boxes have PWM scores below the 3.98 bits threshold, but destroying these boxes decreases expression (see Fig. S9H-I'' for the experiments validating this promoter). **(G)** Analogous to (C) but for gaining a - 10 box in the gray highlighted region of interest on the top strand of P22. **(G')** Analogous to D' but for the wild-type and consensus mutant in panel (G).

4 of the 15 hotspots (~27%) are located in P20-GFP/RFP (Fig. S11), which encodes a bidirectional promoter. However, because our mutational data did not allow us to confidently map this promoter on either strand of P20, we cannot resolve why gaining -10 or -35 boxes may increase expression (2/4 hotspots) or decrease expression (2/4 hotspots, Fig. S11) of these parents.

To summarize, parents with promoter activity readily gain hundreds of additional -10 and -35 boxes, but this appears to primarily modulate expression only when the new -10 and -35 boxes overlap with the existing promoter.

Histone-like nucleoid-structuring protein (H-NS) represses P12-RFP and P22-GFP

The histone-like nucleoid structuring protein (H-NS) is a global transcriptional regulator (Dillon and Dorman, 2010; Hommais et al., 2001; Martínez-Antonio and Collado-Vides, 2003). H-NS binds to ~5% of the *E. coli* genome (Shimada et al., 2011), to horizontally transferred DNA (Navarre et al., 2007; Oshima et al., 2006), to transposons (Cooper et al., 2024), and to promoter islands (Purtov et al., 2014). To test whether H-NS represses the parent sequences, we transformed the parents into a *Δhns* background (KEIO collection, JW1225) (Baba et al., 2006) and compared their fluorescence levels with those in the DH5α (wild-type) background using a plate reader (see methods and Data S6).

We plot the fluorescence changes in **Fig 5A** as distributions for the 50 parents, where positive and negative values correspond to an increase or decrease in fluorescence in the *Δhns* background, respectively. Based on the null hypothesis that the parents are not regulated by H-NS, we classified outliers in these distributions (1.5 × the interquartile range) as H-NS-target candidates. We refer to these outliers as "candidates" because the fluorescence changes could also result from indirect *trans*-effects from the knockout (Mattioli et al., 2020; Metzger et al., 2016). This approach identified 6 candidates for H-NS targets (P2-GFP, P19-GFP, P20-GFP, P22-GFP, P12-RFP, and P20-RFP). For GFP, the largest change occurs in P22-GFP, increasing fluorescence ~1.6-fold in the mutant background (two-tailed t-test, $p=1.16\times10^{-8}$) (**Fig 5B**). For RFP, the largest change occurs in P12-RFP, increasing fluorescence ~0.5-fold in the mutant background (two-tailed t-test, $p=4.33\times10^{-10}$) (**Fig 5B**).

We note that for template P20, which is a bidirectional promoter, GFP expression increases ~2.6-fold in the *Δhns* background (two-tailed t-test, $p=1.59\times10^{-6}$). Simultaneously, RFP expression decreases ~0.42-fold in the *Δhns* background (two-tailed t-test, $p=4.77\times10^{-4}$) (**Fig S12A**). These findings suggest that H-NS also modulates the directionality of P20's bidirectional promoter through either cis- or trans-effects.

We hypothesized that H-NS binds to the candidates we had identified as candidate H-NS targets. To validate this hypothesis computationally, we first acquired a PWM for H-NS derived from RegulonDB (Tierrafría et al., 2022) (**Fig 5C**). We then searched for H-NS-specific hotspots within the 6 parent sequences, i.e., locations in which mutations that destroy H-NS motifs lead to significant fluorescence increases (Mann-Whitney U-test, **Fig S7** and methods). This approach identified 2 candidate motifs in P12-RFP and 1 in P22-GFP whose destruction are associated with fluorescence increases greater than 0.5 a.u. (**Fig 5D**). For the 4 remaining H-NS target candidates, no H-NS motifs are associated with fluorescence increases of more than 0.5 a.u. (**Fig S12B**). See Data S7 for the coordinates, fluorescence changes, and significance for the identified H-NS motifs.

For P22-GFP, a H-NS motif lies 77 bp upstream of the mapped promoter. Mutations which destroy this motif significantly increase fluorescence by +0.52 a.u. (two-tailed MWU test, $q=1.07\times10^{-3}$) (**Fig 5E**). For P12-RFP, one H-NS motif lies upstream of the mapped promoter's -35 box, and the other upstream of the mapped promoter's -10 box. Mutations that destroy these H-NS motifs

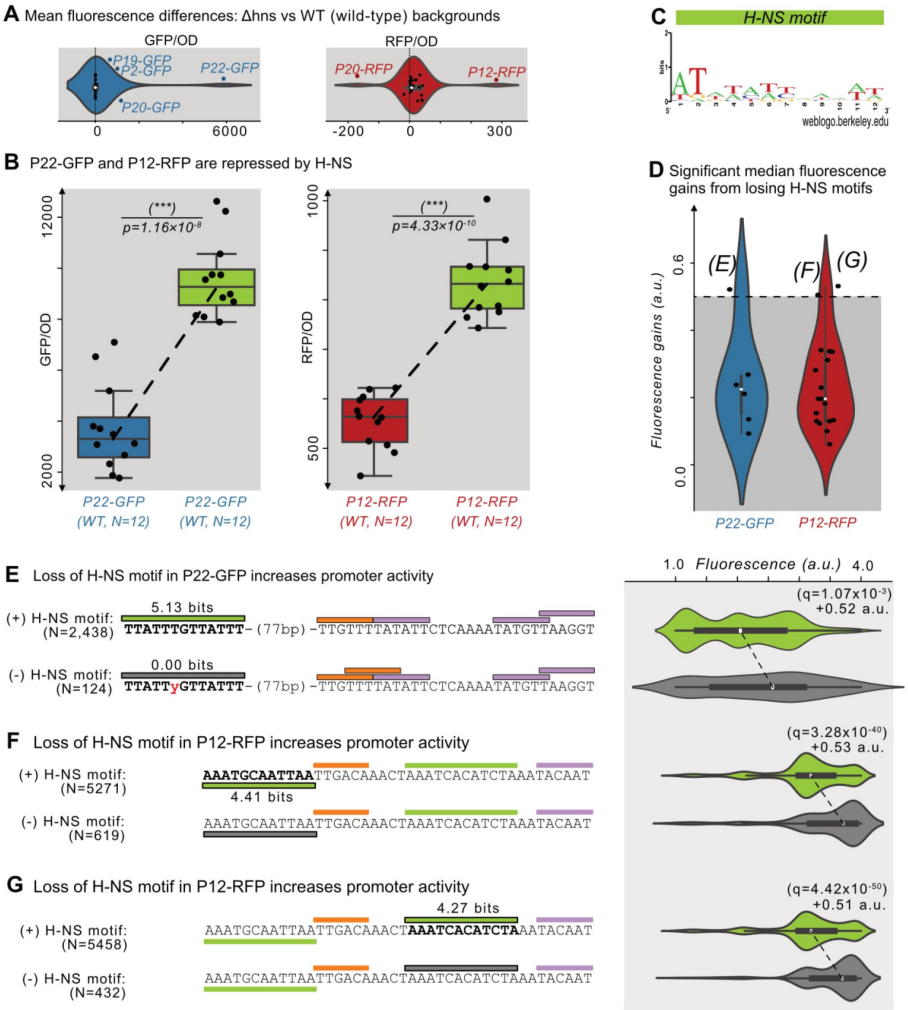


Figure 5.

Histone-like nucleoid-structuring protein (H-NS) represses P12-RFP and P22-GFP.

(A) The difference in average fluorescence levels for each parent sequence in a mutant background for the Histone-like nucleoid-structuring protein (H-NS) *Δhns* vs the wild-type background. Positive and negative values correspond to increases and decreases, respectively in fluorescence in the *Δhns* background. Fluorescence values measured using a plate reader (see methods). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. We classified parents outside of the whiskers as H-NS targets, which are outlined in blue or red, and labeled. Left: GFP fluorescence changes (N=25 parents). Right: RFP fluorescence changes (N=25 parents). **(B)** The fluorescence levels of bacterial colonies (N=12 each) harboring a parent sequence in the wild-type (left) vs the *Δhns* background. Box and whisker plots described in (A). We test the null hypothesis that the means in each background are the same using a two-tailed t-test. Left: P22-GFP levels in the wild-type vs the *Δhns* backgrounds ($p=1.16\times10^{-8}$). Right: P12-RFP levels ($p=4.33\times10^{-10}$). **(C)** A sequence logo derived from a position weight matrix for the transcription factor H-NS (Tierrafría et al., 2022). **(D)** The increase in fluorescence (arbitrary units, a.u.) when losing a H-NS motif in mutual information hotspots of P22-GFP and P12-RFP (see **Fig S7** for calculation overview and Data S7). Dashed lines indicate an effect size threshold of + 0.5 a.u. Each circle corresponds to a gain or loss of a motif in a mutual information hotspot (see **Fig 2** for hotspots). Outlined points with letters in parenthesis highlight the corresponding figure panels. See Figure S12B for additional parents. **(E)** Top: Parent P22-GFP and its PWM predictions (orange = -35 box, magenta = -10 box, green = H-NS motif of interest). We plot the fluorescence scores of all daughters with the H-NS motif. Bottom: the most frequent genotype in the dataset where the H-NS motif is absent. We plot the fluorescence scores of all daughters without the H-NS motif of interest. We tested the null hypothesis that losing the motif significantly increases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values are Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=1.07\times10^{-3}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(F,G)** Analogous to (E) but for the highlighted H-NS motif in P12-RFP.

significantly increase fluorescence by +0.53 and +0.51 a.u., respectively (two-tailed MWU test, $q=3.28\times10^{-40}$ and $q=4.42\times10^{-50}$) (**Fig 5F,G**). Based on these findings, we conclude that these motifs are bound by H-NS.

The binding of H-NS changes when new -10 and -35 boxes are gained

The H-NS sites in P12-RFP also overlap with 2 hotspots for which gaining a new -35 box (TTTTTT➔TTTwwT) and a -10 box (TAAATC➔TAAATt) significantly decreases fluorescence by -1.1 and -0.61 a.u. (two-tailed MWU test, $q=3.35\times10^{-5}$, $q=4.36\times10^{-4}$) (see **Fig 4F** and **Fig S12C,D**). Since these new -10 and -35 boxes do not overlap with the promoter we mapped on P12-RFP, the reduced fluorescence suggests that promoter motifs can repress promoter activity. However, we could not rule out an alternative hypothesis, namely that these mutations actually modulate H-NS binding.

To distinguish these two hypotheses, we engineered a construct with the new -35 box, and compared its fluorescence with that of wild-type P12-RFP. We observed a significant decrease in fluorescence (two-tailed t-test, $p=5.95\times10^{-7}$) (**Fig S12E**), confirming the previously observed association (see **Fig S12C**). This construct, however, also has significantly higher fluorescence in the *Δhns* background compared to the wild-type background (two-tailed t-test, p=0.0036) (**Fig S12E**). In addition, fluorescence of the mutant construct in the *Δhns* background does not significantly differ from that of P12-RFP in the wild type background (two-tailed t-test, p=0.276). These data show that the *Δhns* background rescues the fluorescence loss associated with gaining this new -35 box. Thus, expression lowers because mutations modulate H-NS binding, not by gaining the -35 box.

We repeated these experiments with a construct harboring the new -10 box. We observed a significant decrease in fluorescence compared to the wild-type P12-RFP (two-tailed t-test, p=0.0075) (**Fig S12F**), confirming the previously observed association (**Fig S12D**). This construct, however, also has significantly higher fluorescence in the *Δhns* vs the wild-type background (two-tailed t-test, p=0.0096) (**Fig S12F**). In addition, fluorescence of the mutant construct in the *Δhns* background did not differ from that of P12-RFP in the wild-type background (two-tailed t-test, p=0.310). These data show that the *Δhns* background also rescues the fluorescence loss associated with gaining a new -10 box. Thus, expression lowers because mutations modulate H-NS binding, not by gaining the -10 box.

To summarize, we present evidence that H-NS represses both P22-GFP and P12-RFP in *cis*. H-NS also modulates the bidirectionality of P20-GFP/RFP in cis or trans. In P22-GFP, the strongest H-NS motif lies upstream of the promoter. In P12-RFP, the strongest H-NS motifs lie upstream of the -10 and -35 boxes of the promoter. We note that there are 16 additional H-NS motifs surrounding the promoter in P12-RFP that may also regulate P12-RFP (**Fig S12G**). Mutations in two of these two H-NS motifs can create additional -10 and -35 boxes that appear to lower expression. However, the effects of these mutations are insignificant in the absence of H-NS, suggesting that these mutations actually modulate H-NS binding.

The UP-element does not strongly influence promoter activity in our dataset

The UP element is an additional AT-rich promoter motif that can lie stream of a -35 box in a promoter sequence (Estrem et al., 1998; Ross et al., 1993). We asked whether the creation of UP-elements also creates or modulates promoter activity in our dataset. To this end, we first identified a previously characterized position-weight matrix for the UP element (NNAAAWWTWTTTTNNWAAASYM, PWM threshold score = 19.2 bits) (Estrem et al., 1998) (**Fig S13A**). We then computationally searched for UP-element-specific hotspots within the parent

sequences, i.e., locations in which mutations that gain or lose UP-elements lead to significant fluorescence increases (Mann-Whitney U-test, **Fig S7** and methods. See Data S8 for the coordinates, fluorescence changes, and significance). The analysis did not identify any UP elements whose mutation significantly changes fluorescence.

We then repeated the analysis with a less stringent PWM threshold of 4.8 bits ($1/4^{th}$ of the PWM threshold score). This time, we identified 74 "UP-like" elements that are created or destroyed at unique positions within the parents. 23 of these motifs significantly change fluorescence when created or destroyed. However, even with this liberal threshold, none of these UP-like elements increase fluorescence by more than 0.5 a.u. when gained, or decrease fluorescence by more than 0.5 a.u. when lost (**Fig S13B**). This finding ultimately suggests that the UP element plays a negligible role in promoter emergence within our dataset.

Discussion

We found that non-promoter parents enriched with -10 and -35 vary dramatically (more than 200-fold) in the probability P_{new} that their mutants will have promoter (P_{new} = 0.002 to 0.41, see **Fig 1D**). When trying to understand the reasons for this variation, we found that P_{new} is not significantly correlated with the number of existing -10 or -35 boxes (least squares linear regression, p=0.515, 0.652 respectively). Instead, we present evidence that promoter emergence is best predicted by the level of background transcription each non-promoter parent produces, a phenomenon also referred to as "pervasive transcription" (Kapranov et al., 2007). From an evolutionary perspective, this would suggest that sequences which produce such pervasive transcripts – including the promoter islands (Panyukov and Ozoline, 2013) and the antisense strand of existing promoters (Dornenburg et al., 2010; Warman et al., 2021), may have a proclivity for evolving de-novo promoters compared to other sequences (Kapranov et al., 2007; Wade and Grainger, 2014).

A previous study randomly mutagenized the *appY* promoter island upstream of a GFP reporter, and isolated variants with increased and decreased GFP expression. The authors found that variants with higher GFP expression acquired mutations that 1) improve a -10 box to better match its consensus, and simultaneously 2) destroy other -10 and -35 boxes (Bykov et al., 2020). The authors concluded that additional -10 and -35 boxes repress expression driven by promoter islands. Our data challenge this conclusion in several ways.

First, we find that only ~13% of -10 and -35 boxes in promoter islands actually contribute to promoter activity. Extrapolating this percentage to the *appY* promoter island, ~87% (100% - 13%) of the motifs would not be contributing to its activity. Assuming the *appY* promoter island is not an outlier, this would insinuate that during random mutagenesis, these inert motifs might have accumulated mutations that do not change fluorescence. Indeed, Bykov et al. (Bykov et al., 2020) also found that a similar frequency of -10 and -35 boxes were destroyed in variants selected for lower GFP expression, which supports this argument. Second, we find no evidence that creating a -10 or -35 box lowers promoter activity in any of our 50 parent sequences. Third, we also find no evidence that destruction of a -10 or -35 box increases promoter activity without plausible alternative explanations, i.e. overlap of the destroyed box with a H-NS site, destruction of the promoter, or simultaneous creation of another motif as a result of the destruction. In sum, -10 and -35 boxes are not likely to repress promoter activity.

For the parents with promoter activity, we found that mutations that create new -10 and -35 boxes only affect transcription if these new boxes partially overlap with the -10 and -35 boxes of the parents' promoters. These new motifs affect expression in various ways. For example, they change the composition of the spacer between the boxes (Klein et al., 2021; Lagator et al., 2022; Urtecho et al., 2019), create new boxes that better match the consensus sequence of either box,

or even destroy the promoters themselves. In this way, they bring forth a range of expression changes. These observations could demonstrate how mutations can readily create promoters with a wide dynamic range of activity. This ability facilitates natural selection's fine-tuning of a promoter to precise expression levels, within a small sequence search space that is easily accessible by few DNA mutations. The finding also supports a previous conclusion that true promoters in sequences enriched with promoter motifs typically overlap with multiple -10 and -35 boxes (Huerta and Collado-Vides, 2003).

For the 40 parents without promoter activity, we found that mutations readily created over 3'000 new -10 or -35 boxes at unique positions, with 114 of these including "extended" -10 boxes, and 684 of these comprising a "standard" promoter with a canonical -35 and -10 box spaced 15-20 bp apart. However, fewer than 9 of these new boxes actually created detectable promoter activity (~0.3% of all new boxes). Additionally, we identified 74 UP-like-elements that were created or destroyed in the dataset, but none of these led to a large significant change in promoter activity.

These findings suggest that we are still underestimating the complexity of promoters. For instance, the -10 and -35 boxes, extended -10, and the UP-element may be one of many additional components underlying promoter architecture. Other components may include flanking sequences (Mitchell et al., 2003), which have been observed to play an important role in eukaryotic transcriptional regulation (Afek et al., 2014; Chiu et al., 2022; Farley et al., 2015; Gordân et al., 2013). Recent studies on *E. coli* promoters even characterize an AT-rich motif within the spacer sequence (Warman et al., 2020), and other studies use longer -10 and -35 box consensus sequences (Lagator et al., 2022). Another possibility is that there is much more transcriptional repression in the genome than anticipated (Singh et al., 2014). This would also coincide with the observed repression of H-NS in P22-GFP and P12-RFP, and accounts of H-NS-repression in the full promoter island sequences (Purtov et al., 2014).

These findings also call for future large-scale experiments like ours, together with more complex computational models of promoters, such as thermodynamic models, and computational tools such as machine learning to help us understand promoter architecture. (LaFleur et al., 2022; Lagator et al., 2022; Urtecho et al., 2023; Wang et al., 2020). Additionally, such studies should address the limitations of our own work. First, we use binary thresholding to determine i) the presence or absence of a motif, ii) whether a sequence has promoter activity or not, and iii) whether a part of a sequence is a hotspot or not. While chosen systematically, the thresholds we use for these decisions may cause us to miss subtle but important aspects of promoter evolution and emergence. Second, we measure protein expression through fluorescence as a readout for promoter activity. This readout combines transcription and translation. This means that we cannot differentiate between transcriptional and post-transcriptional regulation, including phenomena such as premature RNA termination (Song et al., 2022; Uptain and Chamberlin, 1997), post-transcriptional modifications (Mohanty and Kushner, 2006), and RNA-folding from riboswitch-like sequences (Mandal and Breaker, 2004).

Overall, our study demonstrates that -10 and -35 boxes neither prevent existing promoters from driving expression, nor do they prevent new promoters from emerging by mutation. The study shows how mutations can create new -10 and -35 boxes near or on top of preexisting ones to modulate expression. However, randomly creating a new -10 or -35 box will rarely create a new promoter, even if the new box is appropriately spaced upstream or downstream of a cognate box. Ultimately our study demonstrates that promoter models need to be further scrutinized, and that using mutagenesis to create de-novo promoters can provide new insights into promoter regulatory logic.

Data availability

A Jupyter notebook with all of the python scripts to recreate the analysis and figures in this study, as well as Data S1-S8 are available at the Github repository: *https://github.com/tfuqua95/promoter_islands*

Sequence reads are available at the Sequence Read Archive (SRA) with accession number: PRJNA1071572: *https://www.ncbi.nlm.nih.gov/bioproject/1071572*

Methods

Position weight matrices (PWMs)

PWMs for the -10 and -35 boxes are from (Belliveau et al., 2018), who derived the instances from 145 of the top predicted -10 and -35 boxes from the database, RegulonDB (Tierrafría et al., 2022). We converted them into PWMs using Biopython.motifs from the Biopython package (Cock et al., 2009) (v1.79). We deliberately assumed a uniform nucleotide composition (25% A, T, C, or G) in this process, because doing so allowed us to test PWM prediction performance in a variety of different sequence backgrounds. For every input sequence, a PWM returns an output score in bits. The higher this score is, the more likely it is that the input sequence can be bound by the protein the PWM represents.

To classify whether an output score is strong enough to be classified as a -10 or -35 box, we used the standard practice of setting a threshold equal to the information content of the PWM (Hertz and Stormo, 1999). The -10 box in our study has an information content of 3.98 bits and the -35 box of 3.39 bits. We classified input sequences with scores at least this high as - 10 or -35 box, respectively. In **Fig S9** we also plot "low-affinity" -10 and -35 boxes. For these motifs, we halved the thresholds (1.99 and 1.70 bits, respectively).

Plasmid MR1 (pMR1)

The plasmid MR1 (pMR1) is a variant of the plasmid RV2 (pRV2) in which the *kan* resistance gene has been swapped with the *cm* resistance gene (Guazzaroni and Silva-Rocha, 2014). Plasmid pMR1 encodes the BBa_J34801 ribosomal binding site (RBS, AAAGAGGAGAAA) 6 bp upstream of the start codon for GFP(LVA). The plasmid also encodes a putative RBS (AAGGGAGG) (Cazemier et al., 1999) 5 bp upstream of the start codon for mCherry on the opposite strand. The plasmid additionally contains the low-to-medium copy number origin of replication p15A (Westmann et al., 2018). A map of the plasmid is available on the Github repository: *https://github.com/tfuqua95/promoter_islands*

Amplifying parent sequences and plasmid MR1 with PCR

We acquired the genomic coordinates for the promoter island sequences from (Panyukov and Ozoline, 2013). We used a polymerase chain (PCR) reaction to amplify 78 parent sequences from the genome, using Q5 polymerase (NEB, USA product #M0491). The reaction amplifies each parent sequence from the genome, and additionally concatenates short sequences to the 5' and 3' ends of the inserts (5'-GGCTGAATTC...insert...GGATCCTTGC-3'). These overhangs allowed us to carry out the error-prone PCR with a single set of primers. See Data S1 for the list of primers and **Fig S14** for an overview of the PCR.

To amplify the parent sequences from the genome, we performed each PCR in a 50 uL reaction volume with the following reagents: 1 uL of each primer (2× primers, 2uL total, concentration 100 uMol), 5 uL Q5 reaction buffer, 1 uL 10 mM dNTPs (Thermo Scientific, USA, product #R0191), 1 uL of the genomic DNA extracted using a Wizard Genomic DNA Purification Kit according to the manufacturer's instructions (A1120, Promega USA), 1 uL Q5 high-fidelity DNA polymerase, and 40 uL of water. We carried out the PCR in a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) with 30 cycles at 55°C for annealing and 72°C for primer extension, both for 30 seconds. To purify the PCR products, we separated them on a 1% agarose gel, excised them, and used a Qiagen QIAquick Gel Purification Kit (Qiagen, Germany, product #28706) following the manufacturer's instructions. To verify the purification's success and estimate the final concentrations of each product, we used a Nanodrop One spectrophotometer (Thermo Scientific, USA).

To create the mutagenesis library of parent sequences, we pooled the 78 PCR products such that each product had the same concentration in the pool. We then performed a single error-prone polymerase chain reaction on the pooled products to create a mutagenesis library. To this end, we used GoTaq polymerase (Promega, USA, product #M3001). Specifically, we combined 1 uL of two PCR primers complementary to both the pMR1 plasmid and to the sequences concatenated to the ends of the initial PCR (1 uL forward and 1 uL reverse primer, each at a 100 uMol l^{-1}) (see Data S1 and **Fig S14**), 10 uL of GoTaq reaction buffer, 1 uL 10 mM dNTPs (Thermo Scientific, USA, product #R0191), 1 uL of pooled DNA, and 1 uL of GoTaq polymerase. We added water to a total of 98 uL, then added 2 uL of 15 mMol $MnCl_2$ to the reaction. We then excised and purified the product as described for the parent sequences.

We additionally used Q5 polymerase to create and amplify linear copies of pMR1 (see Data S1 for primer sequences and **Fig S14**). We carried out the necessary PCR as described for the parent sequences but for an extension time of 2 minutes and 30 seconds.

Molecular cloning

We cloned the error-prone PCR library into pMR1 at the EcoRI and BamHI sites using NEBuilder (New England Biolabs, USA, product #E2621) as follows: 100 ng of linear pMR1, 25 ng insert, 5 uL NEBuilder mastermix, and water to 10 uL total volume. To increase the reaction efficiency, we added 1 uL of T4 DNA ligase buffer (New England Biolabs, USA, product #B0202S). Using a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) we incubated each assembly reaction for 1 hour at 50°C.

Subsequently, we electroporated our product directly into *E. coli* DH5α electrocompetent cells (Takara, Japan, product #9027) by adding 2 uL of (unpurified) product to 2× 50 uL aliquots of *E. coli* (100 uL total) in a 2 mm electroporation cuvette (Cell Projects, England, product #EP-202), and electroporated the cells using a Bio-Rad MicroPulser (Bio-Rad, USA). For recovery, we then added 1 mL of Super Optimal Broth with Catabolite Repression Medium (SOC, provided with the electrocompetent cells (Takara, Japan, product #9027), shaking the cells at 230 RPM at 37°C for 1.5 hours (Infors HT, Switzerland, Multitron).

We then plated 5 uL of the bacterial culture onto a petri dish containing LB-agar supplemented with 100 ug/ml of chloramphenicol, and incubated the plate at 37°C overnight. In parallel, we added 9 mL of LB-chloramphenicol (100 ug/mL) to the remaining 995 uL of bacteria (in SOC), and incubated the resulting culture overnight. The following morning, we manually counted colonies on one quarter of the plate, and estimated the total number of clones (cells) in the original electroporated culture by multiplying this value by 4×200.

PCR-stitching for validation constructs

To validate the predictions of our random mutagenesis experiments, we created variants of inserts in our expression reporter plasmids pMR1 harboring either 1) individual point mutations in a promoter island, or 2) scrambled motif sequences within the promoter island fragments. To design the motifs for scrambling, we used the random.shuffle function from the python random library to maintain both the GC-content of the region and the spacing of surrounding elements. After computationally shuffling each motif, we checked that the shuffling did not result in the creation of a new -10 or -35 motif. If it did, we repeated the shuffling until it did not. We then proceeded to create the mutants in-vivo.

For each insert, we carried out two PCRs. For the first PCR, we use the forward primer: “pMR1_construct_gibson_assembly_forward”, and a specific reverse primer labeled “bottom” which contains the mutation(s) of interest (primers in Data S1). The “bottom” primers include either the point mutations or the scrambled motif sequences. For the second PCR, we use the reverse primer: “pMR1_construct_gibson_assembly_reverse”, and a specific forward primer labeled “top” in Data S1, which is the reverse complement of the “bottom primer” (primers in Data S1). For both PCRs, the amplification template was 1 uL from a mini prep of the template sequence (P1-P25) already cloned into pMR1 as described in “Amplifying parent sequences and plasmid MR1 with PCR.”

Because we used plasmids harboring the template sequence for PCR amplification, the product of each PCR includes inserts with the mutations, and plasmids with the wild-type inserts. To remove the template from the reaction mixture, we directly added 5 uL of rCutsmart buffer (NEB, USA, product #B6004S) and 1 uL of DpnI (NEB, USA, product #R0176S) to each reaction mix, and incubated the reactions at 36°C for 1 hour followed by a 70°C incubation for 15 minutes using a thermocycler (C1000 Touch Thermal Cycler, Bio-Rad, USA). We then purified the products using a Qiagen QIAquick Gel Purification Kit (Qiagen, Germany, product #28706), as described by the manufacturer.

We assembled the first and second PCR products with the pMR1 plasmid using 100 ng of linear pMR1, 25 ng of the first PCR product, 25 ng of the second PCR product, 5 uL NEBuilder mastermix, and water to a total volume of 10 uL (New England Biolabs, USA, product #E2621). The assembly and transformations were then carried out as described in “Molecular cloning.” To confirm the mutations, we selected a single colony and used the primer pMR1_F2 to sequence the insert by Sanger sequencing.

Measuring fluorescence with a plate reader

We streaked bacteria from glycerol stocks onto LB-agar plates with the appropriate selection antibiotics. For the plasmids that we transformed into the DH5α strain (wild-type), this included 100 ug/ml of chloramphenicol. For the plasmids that we transformed into the Δhns strain, this included the 100 ug/ml of chloramphenicol with an additional 50ug/ml of kanamycin. The plates were left to incubate overnight (~16 hours) at 37°C. We then selected individual colonies and inoculated them into 200 uL cultures of LB with 100 ug/ml of chloramphenicol at the aforementioned concentrations. We grew the cultures overnight (~16 hours) at 37°C, shaking at 230 RPM with an Infors HT incubator (Multitron, Switzerland).

We centrifuged the cultures for 10 minutes at 3’200g (Eppendorf, Germany, #5810R), removed the supernatants, and resuspended the pellets in 200 uL of Dulbeco’s Phosphate Buffer Solution (PBS) (Sigma, USA, D8537). We repeated this wash step, and resuspending the pellets once again in 200 uL of Dulbeco’s PBS.

We pipetted 200 uL of the cultures into a 96-well microplate (Sarstedt AG, Germany, #3924) and measured the fluorescence using a Tecan Spark plate reader (Tecan, Switzerland). We measured GFP with an excitation of 485 nm (bandwidth 20 nm) and emission of 535 nm (bandwidth 20 nm). We measured red fluorescence with an excitation of 560 nm (bandwidth 20 nm) and emission of 620 nm (bandwidth 20 nm). We additionally measured the optical density at 600 nm (bandwidth 3.5 nm) (OD_{600}). We report all fluorescence readings from the plate reader using the fluorescence counts normalized to the OD_{600}. See Data S6 for the readings.

Sort-seq controls

We synthesized three control promoter sequences (Data S1) as fiducial markers for defining the fluorescence sorting gates for sort-seq. The insert DNA sequences were chemically synthesized (Integrated DNA Technologies, USA). We then amplified these promoters, cloned them into plasmid MR1, and transformed *E.coli* with them as described in "Molecular Cloning." The first is the promoter iGem bba_j23110 on the top DNA strand, which drives moderate GFP expression from plasmid pMR1. The second is the reverse complement of this promoter, which drives moderate RFP expression from the bottom DNA strand of pMR1. The third and final (negative) control is the wild-type pMR1 plasmid without any insert in its multiple cloning sequence.

Sort-Seq

To perform sort-seq, we first added 100 uL of the glycerol stock of our mutagenesis library to 1 mL of LB-chloramphenicol (100 ug/ml). We then allowed the resulting culture to shake at 230 RPM and 37°C overnight (~16 hours) on an Infors HT (Multitron, Switzerland) shaking incubator. Subsequently, we centrifuged the culture and washed the cell pellet twice with Dulbeco's Phosphate Buffered Saline (PBS) (Sigma, USA, D8537). We then resuspended the pellet in PBS.

To detect fluorescence caused by GFP and RFP expression, we measured green fluorescence using a 488 nm laser (FITC-H) at 750 volts, and red fluorescence with a 633 nm laser (PE-H) at 510 volts on an Aria III cell sorter (BD Biosciences, USA). We used the following fluorescence gates for sort-seq (Figure S2).

RFP bin #1 (none, i.e., no fluorescence): We defined a minimum boundary to not sort cell debris, salts, empty droplets, and other impurities, as the larger of two PE-H values. These values are the minimum PE-H of the negative control (empty pMR1 plasmid) and the minimum PE-H of the opposite fluorophore positive control (GFP in this case). The upper boundary of this bin is the highest PE-H value detected for these same controls.

RFP bin #4 (high): We defined the lower boundary of this bin using the mean fluorescence level of the positive RFP control. This bin does not have an upper bound, because it encompasses the highest levels of fluorescence.

RFP bins #2 and #3 (low and moderate): the lower bound of bin #2 begins at the upper bound of bin #1. The upper bound of bin #3 ends at the lower bound of bin #4. The upper bound of bin #2, which is identical to the lower bound of bin #3, equals the average of the lower boundary of bin #4 and the upper bound of bin #1.

We defined the bins for GFP analogously, but with GFP and RFP controls swapped, with the following exception: Because the GFP positive control produces a bimodal FITC-H distribution, we defined the lower bound of bin #4 for green fluorescence as the peak of the higher mode of this distribution. See **Fig S2** for an illustration of the bins.

We sorted the library using two consecutive rounds of sorting on consecutive days. On the first day, we sorted 2'160'000, 71'545, 28'399, and 22'720 cells into GFP bins none, low, moderate, and high, respectively; and 2'160'000, 540'000, 101'636, and 76'898 cells into RFP bins none, low,

moderate, and high, respectively. After sorting, we added 1 mL of SOC medium (Sigma, USA, product #CMR002K) to the sorted cultures, incubating the cells at 37°C and shaking for two hours at 230 RPM (Infors HT, Switzerland, Multitron). Following this recovery period, we added 9 mL of LB-Chloramphenicol (100 ug/ml chloramphenicol) to each culture and incubated them overnight (~16 hours) under the same conditions.

On day two, we once again washed the cultures and re-sorted cells in each culture into the same fluorescence bin. That is, we sorted cells from GFP-none into GFP-none, GFP-low into GFP-low, and so forth. This resorting step ensures that each genotype fluoresces within the same boundaries as on the previous day, and reduces the likelihood of propagating sorting errors into subsequent analyses. The resulting number of cells for each bin were as follows: GFP-none: 6'480'000, GFP-low: 216'000, GFP-moderate: 90'000, GFP-high: 69'000, RFP-none: 6'480'000, RFP-low: 540'000, RFP-moderate: 306'000, RFP-high 231'000 cells. We allowed the cells to recover as described at the end of day one.

After ~16 hours of overnight culture, we created glycerol stocks for each culture (8 cultures total) as described in "Molecular Cloning." From the remaining culture volume, we isolated the plasmids using a Qiagen QIAprep Spin Miniprep kit (Qiagen, Germany, product #27104) as described by the manufacturer. We amplified the plasmid inserts using PCR as described in "Molecular Cloning", using primers (Data S1) that add a unique multiplexing barcode onto the 5'-end of each PCR product from each bin. In addition, we isolated the inserts from the unsorted library, and endowed them too with a unique barcode. This procedure yielded 9 (=8+1) sequence samples, each with a unique barcode. We then pooled equimolar amounts of these samples, and sequenced them using Illumina paired-end sequencing (Eurofins GmbH, Germany).

Illumina sequencing

The amplicon pool was sequenced by Eurofins Genomics (Eurofins GmbH, Germany) using a NovaSeq 6000 (Illumina, USA) sequencer, with an S4 flow cell, and a PE150 (Paired-end 150 bp) run. In total, 282'843'000 reads and 84'852'900'000 bases were sequenced. Raw sequencing reads can be found here: *https://www.ncbi.nlm.nih.gov/bioproject/1071572*

Processing sequencing reads

Using Flash2 (Magoč and Salzberg, 2011), we merged the paired-end reads that resulted from Illumina sequencing. Each paired-end read contains palindromic cut-sites for 5'-EcoRI (GAATTC) and 3'-BamHI (GGATCC). To orient all of the paired-end reads in the same direction, we searched each paired-end read for both cut-sites, discarding any sequence that did not encode both, and took the reverse complement of the read if the BamHI site was upstream of the EcoRI site. We then identified the tagmented barcode sequence upstream of the EcoRI site, classifying from which fluorescence bin the read originated. We then cropped the sequences to remove the barcodes, BamHI, and EcoRI sites, and counted the total number of reads within every bin. We refer to each of these unique paired-end read sequences as *daughter sequences*.

To focus on point mutations rather than insertions and deletions, we removed daughter sequences with a length different from the wild-type 150 bp. To minimize the influence of sequencing errors in our dataset, we removed daughter sequences that did not occur at least once in the unsorted library, and daughter sequences with less than thirty reads in the entire dataset.

To map each daughter sequence to its respective template parent sequence, we calculated the Hamming distance (i.e. the number of nucleotide differences between two sequences) between each daughter sequence and the 78 parent sequences, and identified the parent as the sequence with the smallest Hamming distance. To minimize our likelihood of mapping the daughter sequences to the wrong parent sequence, we removed daughter sequences from the dataset for which the closest Hamming distance was greater than 10. Because the error-prone PCR created

different numbers of daughter sequences per parent sequence, we also removed all parent and their respective daughter sequences if there were less than 2'000 unique daughter sequences. After these calculations and filtration steps, our dataset contained 245'639 daughter sequences derived from 25 parent sequences.

To calculate fluorescence scores (*F*) we used in the next analysis step, we calculated (*F*) using Equation (1):

$$F = \frac{\sum_{f=1}^{4}(f \times Reads_f)}{\sum_{f=1}^{4}(Reads_f)} \quad (1)$$

Here, *f* is an integer index corresponding to each of the four fluorescence bins, i.e., *f* = 1,2,3,4 corresponds to no fluorescence, low, moderate, and high fluorescence, respectively. $Reads_f$ is the number of times the sequence was identified in each fluorescence bin *f*. We calculated *F* for both green and red fluorescence. Each score is in arbitrary units (a.u.) of fluorescence.

See Data S2 for a csv file with the daughter sequences and their respective fluorescence values, which we analyzed further with the code on the GitHub repository: *https://github.com/tfuqua95/promoter_islands*.

Mutual information

To calculate mutual information between nucleotides and fluorescence scores, we first randomly split the dataset into three equal-sized pseudo-replicates (r1, r2, and r3) to minimize the effects of sampling biases as well as sorting errors when calculating mutual information. For each pseudo-replicate, we calculated the mutual information $I_i(b, f)$ between the nucleotide identity *b* and fluorescence score *f* score at each position *i* (*1<i<150*), of every daughter sequence using *equation (2)*, as defined in ref. (Kinney et al., 2010):

$$I_i(b,f) = \sum_b \sum_f p_i(b,f) \log_2 \frac{p_i(b,f)}{p_i(b) \times p(f)} - \frac{(n_b - 1)(n_f - 1)\log_2 e}{2N} + O(N^{-2}) \quad (2)$$

In the left term on the right-hand side of equation (2), *b* indexes each nucleotide base (*b* = A,T,C,G) and *f* denotes the fluorescence score (see Processing sequencing reads) rounded to the nearest whole number (*f=1,2,3,4*). The expression $p_i(b)$ denotes the relative frequency of base *b* (A, T, C, or G) at position *i*. *p*(*f*)) is the relative frequency of a daughter sequence with a (rounded) fluorescence score of 1,2,3, or 4. $p_i(b, f)$ is the observed joint probability, i.e., the relative frequency of a daughter sequence having a score of *f* and base *b* at position *i*.

The right term on the right-hand side of equation (2) is a correction term for differences in degrees of freedom (9 degrees of freedom), where n_b = 4 is the number of possible bases n_f = 4 is the number of possible (rounded) fluorescence scores, and *N* is the size of the library. The fewer degrees of freedom there are and the larger the library is, the smaller the correction term.

For the mutual information plots shown in the text, we computed a Gaussian filter with the ndimage (alpha = 1) function in scipy (v1.8.1). In these plots, the solid line is always the average mutual information between r1, r2, and r3 at each position, and the lightly shaded region indicates ± 1 standard deviation between r1, r2, and r3 at each position.

Identifying mutual information hotspots

We calculated the mutual information for each parent and its daughters (see Mutual Information), and used the following procedure to automatically identify mutual information hotspots. First, we subtracted the standard deviation from the mean mutual information at each nucleotide position *i*, and truncated all values smaller than 0.0005 bits to 0 bits. We refer to the resulting quantity as *adjusted* mutual information. We then established the 90th percentile of this adjusted mutual information at each nucleotide position *i* as a threshold to identify mutual information hotspots.

Specifically, we identified mutual information hotspots as nucleotide positions where the adjusted mutual information 1) lies above this threshold, and 2) is greater than the mutual information at the positions *i*-1 and *i*+1. See Data S3 for hotspot coordinates.

Analysis of motif gain and loss

To identify hotspots in which gaining and losing a -10 or -35 box associates with a significant change in gene expression (fluorescence), we performed the following procedure. First, we drew a 6 bp sliding window starting at the beginning of each parent sequence. Within this window, we calculated the respective -10 and -35 PWM score for every daughter sequence (see **Fig S7A**). For each daughter sequence, if the binding score was greater than a designated threshold (-35 box = 3.39 bits, -10 box = 3.98 bits, see "Position Weight Matrices"), we added the fluorescence score to a "Motif (+)" list. If the PWM score was below this threshold, we added the score to a "Motif (-)" list (see **Fig S7B**).

If each list contained more than 10 values, we then tested the null hypothesis that the fluorescence scores have the same central tendency, using a two-sided Mann-Whitney U test (mannwhitneyu function from scipy.stats, v1.8.1). If the p-value was less than 0.05 and was located within ±3 bp from a hotspot, we recorded the value in Data S4. We then advanced the 6 bp window in steps of one nucleotide position until it had reached the final end of the parent sequence, and repeated this calculation at each location of the sliding window. See **Fig S7**.

We performed this analysis for both GFP and RFP fluorescence scores on their respective DNA strands for both -10 and -35 boxes and for all 25 parent sequences (Data S4). To account for multiple hypothesis testing, we calculated corrected q-values using a Benjamini-Hochberg correction (false-discovery rate 0.05).

For subsequent analyses, we only focused on significant associations found within ± 3 bp from the peak of each hotspot (see "Identifying promoter emergence hotspots" and "Mutual information"), and associations where the median fluorescence difference between the lists box present and box absent is larger than the absolute value of 0.5 arbitrary units (a.u.).

We carried out the same analyses for H-NS (Data S7) and the UP-element (Data S8).

Sigmoid curve-fitting

To fit the distribution of Pnew to a sigmoid function, we used the function curve_fit from the python module scipy.optimize (version 1.8.1). The sigmoid function we employ is based on the equation

$$Pnew = \frac{L}{(1 + e^{-k(x_0 - x)})} \qquad (3)$$

Here, L corresponds to the upper asymptote, x_0 is the inflection point, k is the slope of the curve, and x corresponds to a fluorescence value measured from a parent sequence. After fitting L, x_0, and k, we tested the null hypotheses that the parameters L, x_0, and k are equal to 0 using a non-linear least squares regression (p<Z, p<Z, and 1.43×10^{-9} for L, x_0, and k, respectively.

Acknowledgements

The European Molecular Biology Organization (EMBO) supports TF with a postdoctoral fellowship (ALTF 963-2021). TF is also supported by a University of Zurich Postdoc Fellowship (FK-23-120). AW and his group is supported by the Swiss National Science Foundation (grants 310030_208174).

We would like to thank the Wagner lab for all discussions and conversations, even if they were not always pertinent to science. We additionally thank Baxter for inspiring Tim with the analysis pipeline during a painfully slow and cold morning walk.

Author contributions

TF and AW conceived the study and designed experiments. YS and TF carried out experiments. TF, YS, and AW analyzed the data. TF wrote the scripts for computational analysis. TF generated the figures. TF and AW wrote the paper.

Conflicts of interest

The authors declare no competing interests.

Supplementary figures

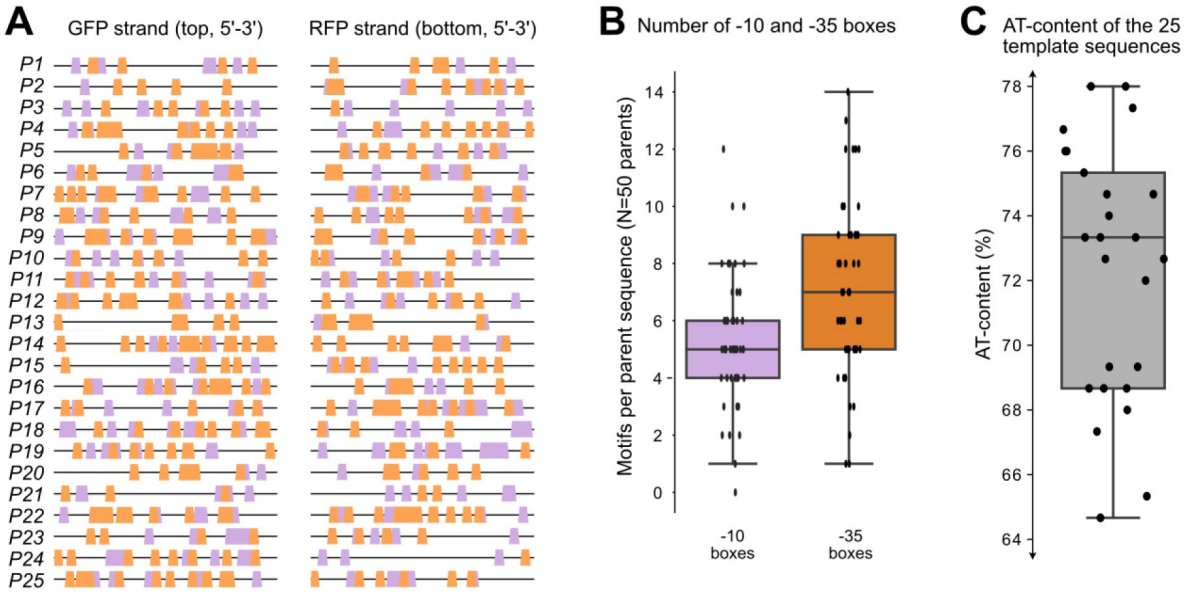


Figure S1.

Promoter island sequences.

(A) The locations and arrangements of -10 and -35 boxes located across both the top and bottom strands of the 150 base pair DNA template sequences (P1-P25). Orange trapezoids: -35 boxes. Magenta trapezoids: -10 boxes. Left: top (GFP) strand. Right: bottom (RFP) strand. Note: the plot is shown from 5'-3' for both strands. We refer to each strand on each template as an individual "parent sequence (N=50)." **(B)** Distributions of the number of motifs found in the 50 parent sequences. We identified motifs with position-weight matrices (PWMs, methods). Magenta: number of -10 boxes in each parent sequence (n=50). Orange: number of - 35 boxes in each parent sequence (n=50). **(C)** Distribution of AT-content for the 25 template sequences. Boxes in B) and C) represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR

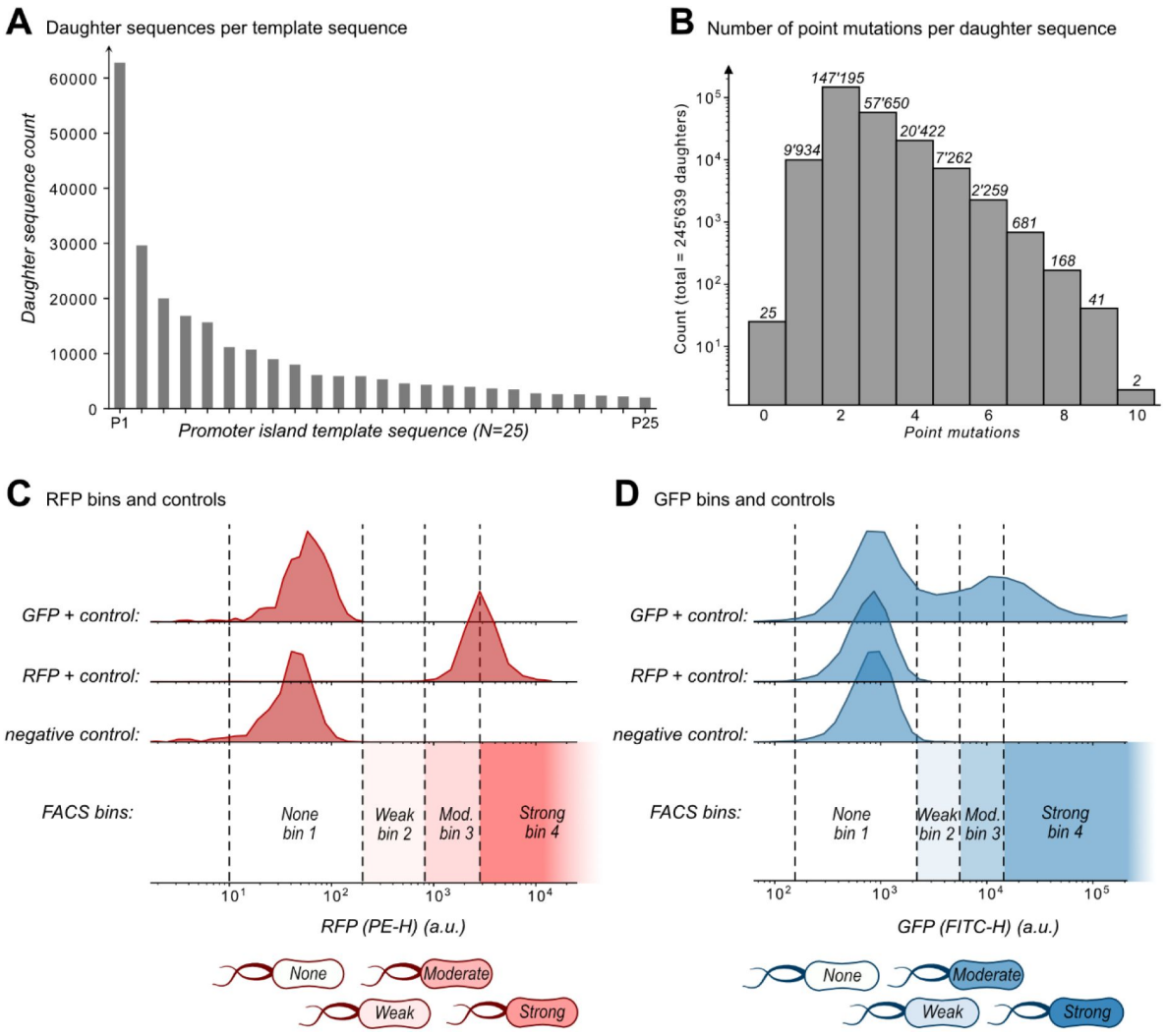


Figure S2.

Mutagenesis library and sort-seq bins.

(A) A histogram with the number of unique daughter sequences per template sequence (N=245’639 daughters and N=25 template sequences). **(B)** The number of point mutations per daughter sequence (N=245’639 daughter sequences). The number above each bin is the number of daughters in each bin. **(C)** We defined four red fluorescence bins (1, 2, 3, 4, corresponding to none, weak, moderate, and strong fluorescence, respectively) using three controls: GFP+, RFP+, and a negative control (see Methods). Plots show histograms of the fluorescent readouts from 10’000 cells for each control. For bin 1 (none), we defined a minimum boundary as the larger of two PE-H values. These values are the minimum PE-H of the negative control (empty pMR1 plasmid) and the minimum PE-H of the opposite fluorophore positive control (GFP in this case). The upper boundary of bin 1 is the highest PE-H value detected for these same controls. For bin 4 (strong), we defined the lower boundary as the mean fluorescence level of the positive RFP control. This bin does not have an upper bound, because it encompasses the highest levels of fluorescence. For RFP bins 2 (weak) and 3 (moderate), the lower bound of bin 2 is identical to the upper bound of bin 1. The upper bound of bin 3 is identical to the lower bound of bin 4. The upper bound of bin 2, which is identical to the lower bound of bin 3, equals the average of the lower boundary of bin 4 and the upper bound of bin 1. **(D)** Analogous to (C) except with GFP and RFP controls swapped, with the following exception: Because the GFP positive control produces a bimodal FITC-H distribution, we defined the lower bound of bin 4 for green fluorescence as the peak of the higher mode of this distribution.

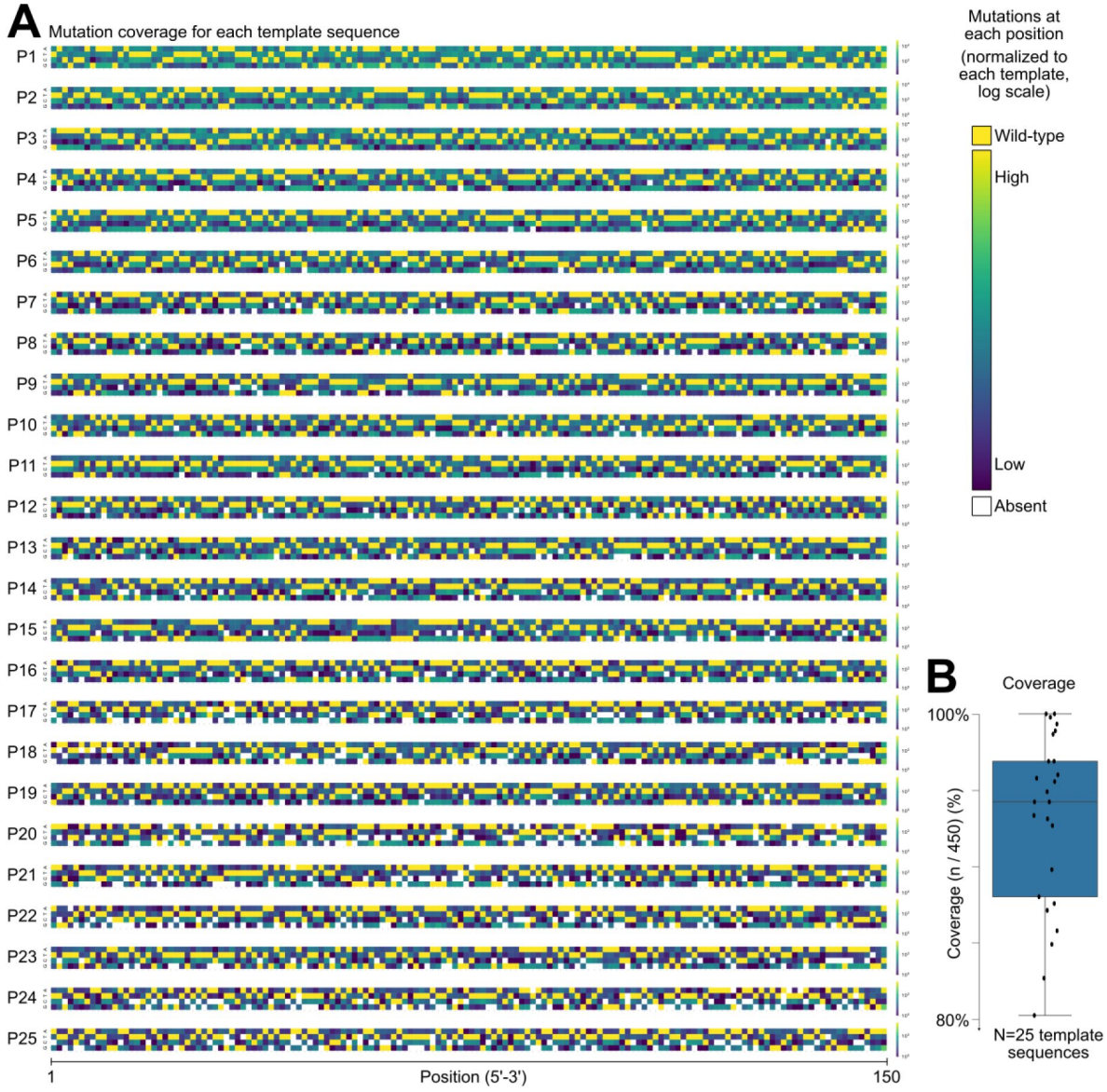


Figure S3.

Mutational coverage for the template sequences.

(A) Heatmaps depicting the frequency of each nucleotide (A,T,C,G, y-axis) at each position (5'-3', x-axis) for all of the daughter sequences of each template sequence (P1-P25, N=25 templates). Daughter sequences contain 1-10 point mutations (see **Fig S2B**). The color of each square in the heatmap corresponds to the frequency of each nucleotide occurring in the library at its respective position, where less frequent nucleotides are shown in dark magenta. Highly frequent nucleotides are in yellow (log-scale). These include the nucleotides of the parent sequence, due to our modest mutation rate (~2.0 point mutations per daughter sequence). Light gray squares indicate mutations absent from the library. **(B)** We calculate coverage as the percentage of all possible mutations of the template sequences (A,T,C,G and positions 1 through 150, 3×150 = 450 total neighboring mutations). In the boxplot, the box represents the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR.

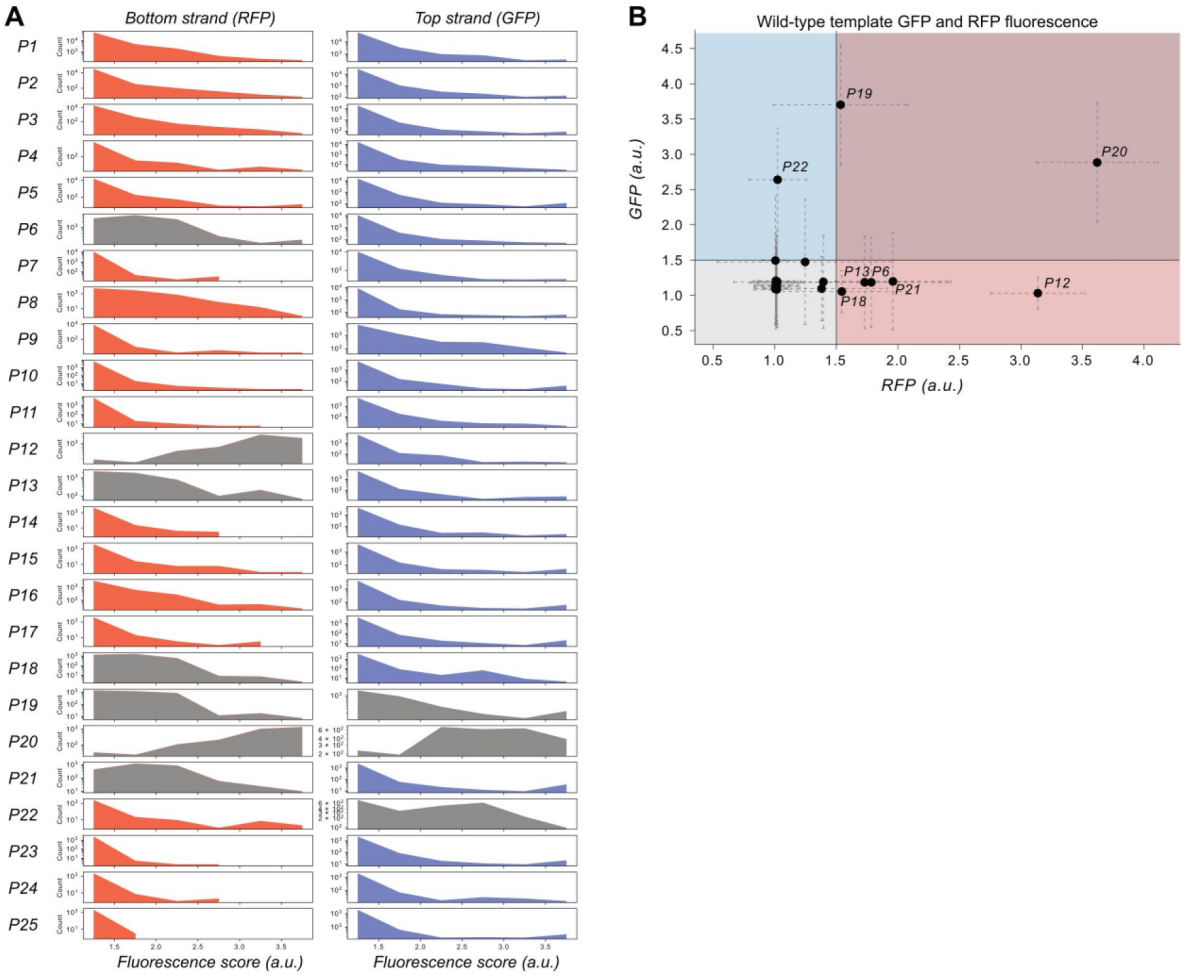


Figure S4.

Histogram of the fluorescence score distributions for each parent and its respective daughter sequences.

(A) Rows: template sequences (P1-P25). Left column: bottom strand (RFP fluorescence scores); Right column: top strand (GFP fluorescence scores.) Each panel is a histogram of the frequency of fluorescence scores (arbitrary units, a.u.) from the sort-seq experiment. Gray histograms indicate those parent sequences that already encode promoter activity. Note: daughter sequence counts vary among parents, and y-axis is unique to each panel. **(B)** A scatterplot comparing of RFP fluorescence scores (horizontal axis) and GFP fluorescence scores (vertical axis, a.u.: arbitrary units). Each point represents the fluorescence score of one the 25 template sequences (P1-P25). Length of dashed lines correspond to the standard deviation of the fluorescence values (see methods). Points that are labeled have a fluorescence score for GFP or RFP greater than 1.5 a.u. Points in the gray quadrant are non-promoters for both GFP and RFP. Points in the blue quadrant are promoters driving GFP expression only. Points in the red quadrant are promoters driving RFP expression only. Points in the purple quadrant drive expression of both GFP and RFP.

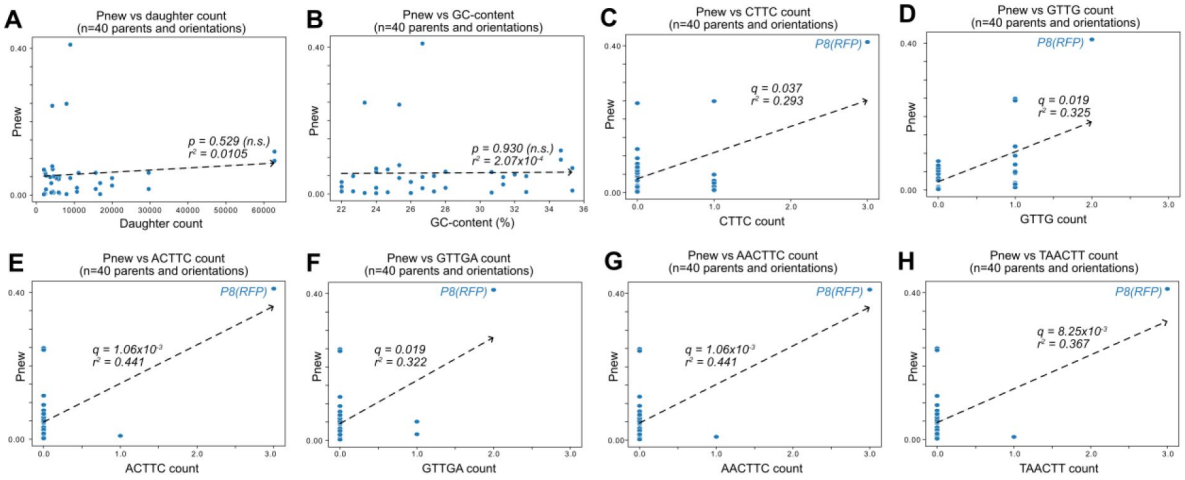


Figure S5.

Correlation of P_{new} with sequence composition.

(A) Scatterplots of P_{new} (see **Fig 1**) and the number of daughter sequences per parent. P_{new} is the ratio of daughter sequences with a fluorescence score greater than 1.5 a.u. to the total number of daughter sequences for each parent. Each point represents a parent without promoter activity (n=40). The dashed line is the line of best fit calculated using the method of least squares. We test the null hypothesis that the slope of the correlation equals zero with the Wald Test. The r^2 value is the Pearson correlation coefficient. Calculation carried out using scipy.stats.linregress (version 1.8.1) (p=0.529, r^2=0.0105, not significant (n.s.)). **(B)** Analogous to (A) but comparing the GC-content of each parent sequence with P_{new} (p=0.930, r^2=2.07×10^{-4}, n.s.). **(C-H)** We searched for k-mers that correlate with P_{new} and identified six such k-mers with significant correlations. These k-mers are shown on the horizontal axes of the panels. Q-values correspond to p-values corrected with a Benjamini-Hochberg correction for multiple testing. The outlier at P8-RFP is labeled in each panel. Without P8-RFP, none of these correlations would be significant. **(C)** Analogous to (A) but for the k-mer CTTC (q=0.037, r^2=293). **(D)** Analogous to (A) but for the k-mer GTTG (q=0.019, r^2=0.325). **(E)** Analogous to (A) but for the k-mer ACTTC (q=1.06×10^{-3}, r^2=0.441). **(F)** Analogous to (A) but for the k-mer GTTGA (q=0.019, r^2=0.322). **(G)** Analogous to (A) but for the k-mer AACTTC (q=1.06×10^{-3}, r^2=0.441). **(H)** Analogous to (A) but for the k-mer TAACTT (q=8.25×10^{-3}, r^2=0.367).

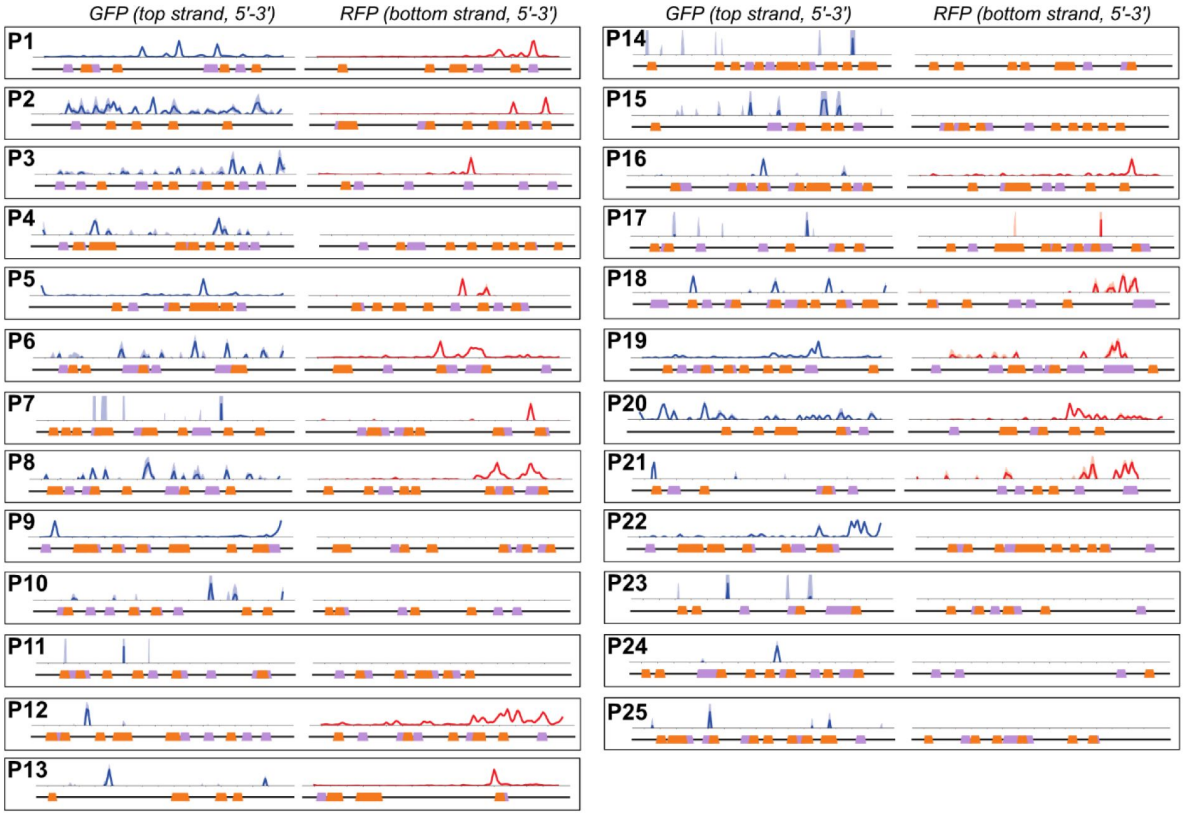


Figure S6.

Mutual information and promoter motifs in the parent sequences.

Each panel corresponds to a unique parent sequence, with its numerical identifier in bold. Within each panel there are two plots, corresponding to the top strand (left), and the bottom strand (right), both shown from 5′ (left) to 3′ (right). Within each panel, we show on top the mutual information $I_i(b, f)$ between nucleotide identity and fluorescence levels at every position (Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods)). The bottom of each panel shows position-weight matrix (PWM) predictions for the occurrence of -10 box motifs (magenta trapezoids) and -35 box motifs (orange trapezoids) along the wild-type parent sequence. See methods.

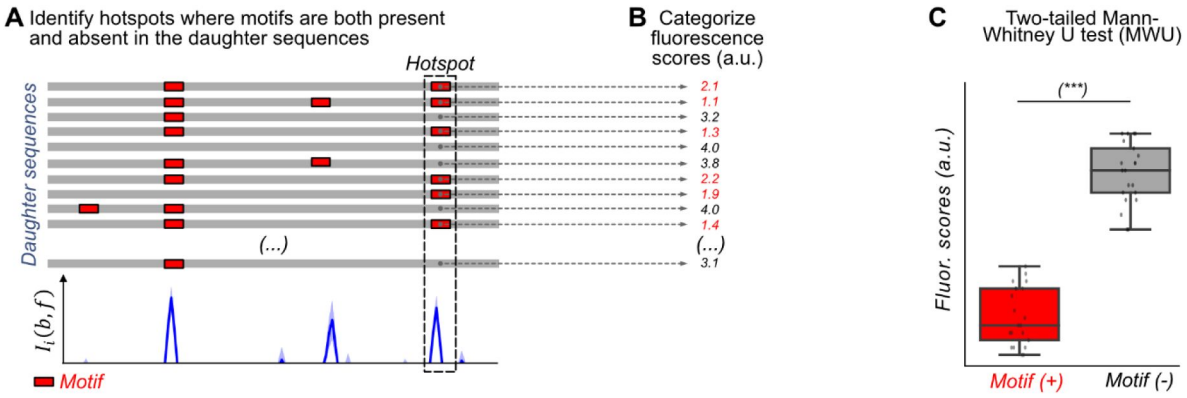


Figure S7.

Identifying where motifs are gained and lost in hotspots that are associated with changes in fluorescence.

These panels do not show data from any parent sequence or hotspot, but show hypothetical data to illustrate how we identified the associations of base identity and fluorescence in our analysis. **(A)** Top: a cartoon illustration of all daughter sequences from a single parent. Red boxes correspond to position-weight matrix (PWM) predicted motifs. Bottom: Mutual information between fluorescence levels and nucleotide identity at each position calculated from the respective daughter sequences of a given parent sequence. We examine for each mutual information hotspot if -35 or -10 box motifs are present or absent in each daughter sequence. The dashed rectangle highlights a hotspot at the 3'-end of the daughter sequences discussed in B) and C). **(B)** We categorize the fluorescence scores for the daughter sequences based on whether they have a motif (red) or not (black) in the hotspot. **(C)** We test the null hypothesis that the fluorescence scores in each group have the same central tendency using a two-sided Mann-Whitney U (MWU) test. See methods.

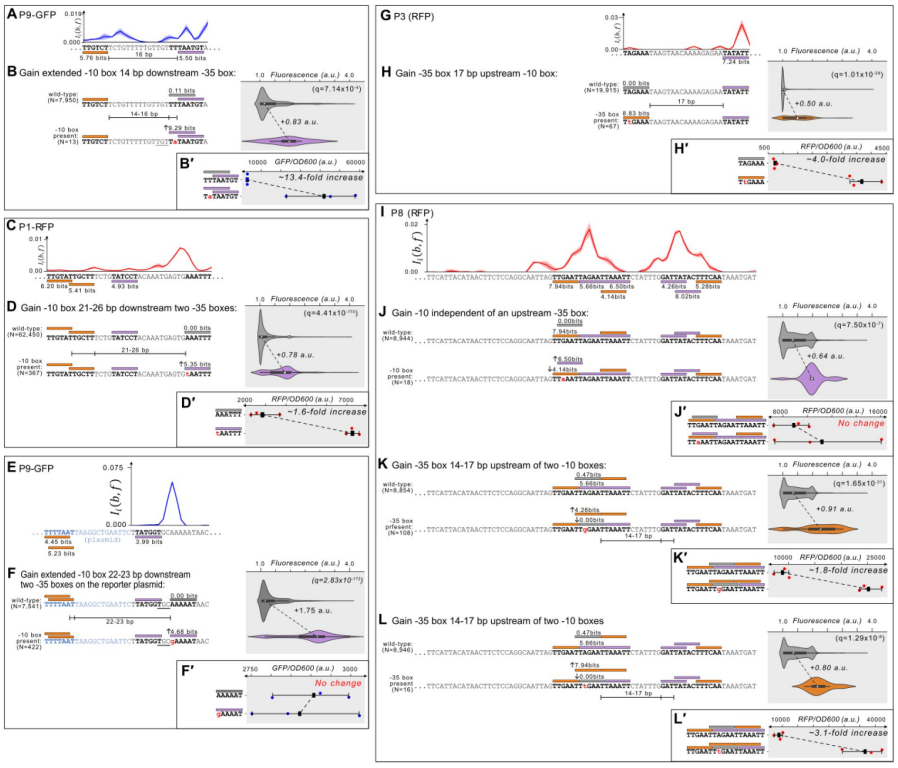


Figure S8.

Additional examples of which gaining -10 and -35 boxes creates de-novo promoters.

(A) Parent P9-GFP. Top: Mutual information $I_i(b, f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Bottom: position-weight matrix (PWM) predictions for the -10 boxes (magenta rectangles) and -35 boxes (orange rectangles) along the wild-type parent sequence. **(B)** Top: Parent P9-GFP and its PWM predictions from (A). We plot the fluorescence scores of all daughters without a -10 box in the region of interest (left, gray rectangle). Bottom: the most frequent genotype in the dataset where a -10 box is in the region of interest. We plot the fluorescence scores of all daughters with a -10 box in the region of interest. We tested the null hypothesis that the gain of the motif significantly increases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=7.14\times10^{-4}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(B')** Top: the fluorescence readouts of three colonies harboring the wild-type P9-GFP reporter construct measured using a plate reader. The y-axis corresponds to the fluorescence readout normalized to the optical density of the culture (OD_{600}) during the reading (see methods). Each point corresponds to the fluorescence of an individual colony. Whiskers correspond to the minimum and maximum values and the dark square the mean fluorescence level. Bottom: analogous to top, but for three colonies harboring the degenerate consensus sequence identified to change fluorescence in panel B. **(C)** Analogous to (A) but for parent P1-RFP. **(D)** Analogous to (B) but for gaining a -10 box in the gray region of interest on P1-RFP (two-tailed MWU test, $q=4.41\times10^{-110}$). **(D')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (D). **(E)** Analogous to (A) but for parent P9-GFP. **(F)** Analogous to (B) but for gaining a -10 box in the gray region of interest on P9-GFP (two-tailed MWU test, $q=2.83\times10^{-173}$). **(F')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (E). **(G)** Analogous to (A) but for parent P3-RFP. **(H)** Analogous to (B) but for gaining a -35 box in the gray region of interest on P3-RFP (two-tailed MWU test, $q=1.01\times10^{-29}$). **(H')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (H). **(I)** Analogous to (A) but for parent P8-RFP. **(J)** Analogous to (B) but for gaining a -10 box in the gray region of interest on P8-RFP (two-tailed MWU test, $q=7.50\times10^{-7}$). **(J')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (J). **(K)** Analogous to (B) but for gaining a -35 box in the gray region of interest on P8-RFP (two-tailed MWU test, $q=1.65\times10^{-31}$). **(K')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (K). **(L)** Analogous to (B) but for gaining a -35 box in the gray region of interest on P8-RFP (two-tailed MWU test, $q=1.29\times10^{-8}$). **(L')** Analogous to (B') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (L).

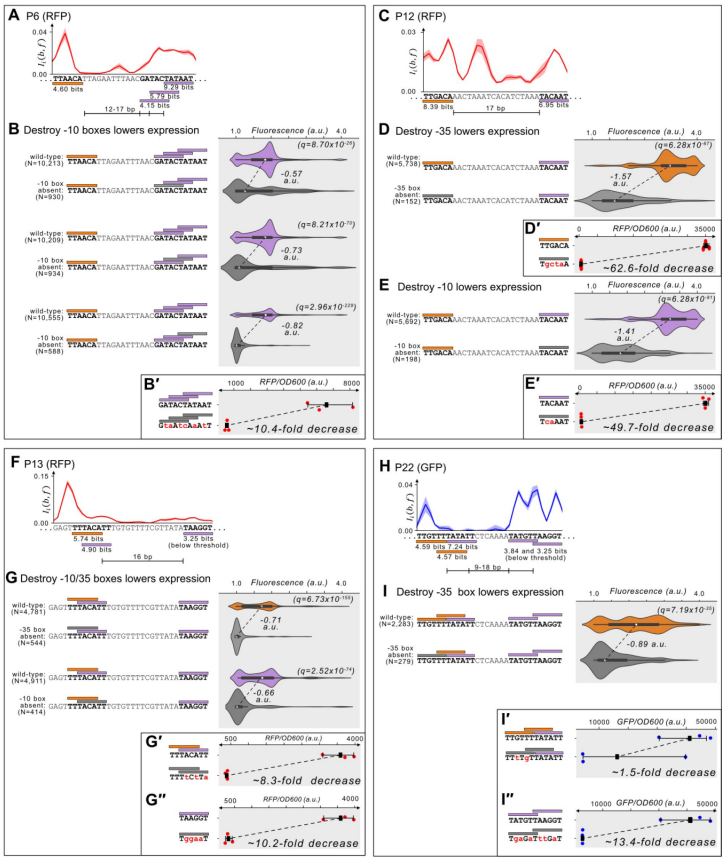


Figure S9.

Mapping promoters in active parents.

(A) Parent P6-RFP. Top: Mutual information $I_i(b, f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Bottom: position-weight matrix (PWM) predictions for the -10 boxes (magenta rectangles) and -35 boxes (orange rectangles) along the wild-type parent sequence. **(B)** Top: Parent P6-RFP and its PWM predictions from (A). We plot the fluorescence scores of all daughters with a -10 box. Bottom: the most frequent genotype in the dataset where a -10 box is missing in the gray region of interest. We plot the fluorescence scores of all daughters without a -10 box in the region of interest. We tested the null hypothesis that the loss of the motif significantly decreases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, from top to bottom: $q=8.70\times10^{-26}$, $q=8.21\times10^{-70}$, $q=2.96\times10^{-229}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(B')** Top: the fluorescence readouts of three colonies harboring the wild-type P6-RFP reporter construct measured using a plate reader. The y-axis corresponds to the fluorescence readout normalized to the optical density of the culture (OD_{600}) during the reading (see methods). Each point corresponds to the fluorescence of an individual colony. Whiskers correspond to the minimum and maximum values and the dark square the mean fluorescence level. Bottom: analogous to top, but for three colonies harboring a scrambled sequence where the three regions of interest were tested (See methods for scrambling procedure). **(C)** Analogous to (A) but for parent P12-RFP. **(D)** Analogous to (B) but for losing a -35 box in the orange region of interest on P12-RFP (two-tailed MWU test, $q=4.41\times10^{-110}$). **(D')** Analogous to (B') but the fluorescence levels from three colonies with and without the scrambled -35 box sequence. **(E)** Analogous to (B) but for losing a -10 box in the magenta region of interest on P12-RFP. **(E')** Analogous to (B') but for the fluorescence levels from three colonies with and without the scrambled -10 box sequence. **(F)** Analogous to (A) but for parent P13-RFP. **(G)** Analogous to (B) but for losing an overlapping -10/35 box in the orange and magenta region of interest on P13-RFP. **(G')** Analogous to (B') but for the fluorescence levels from three colonies with and without the scrambled -10/35 box sequence. **(G'')** Analogous to (B') but for the fluorescence levels from three colonies with and without a scrambled low-affinity -10 box sequence. **(H)** Analogous to (A) but for parent P22-GFP. **(I)** Analogous to (B) but for losing a -35 box in the orange region of interest on P22-GFP. **(I')** Analogous to (B') but for the fluorescence levels from three colonies with and without two scrambled -35 box sequences. **(I'')** Analogous to (B') but for the fluorescence levels from three colonies with and without two scrambled low-affinity - 10 box sequences.

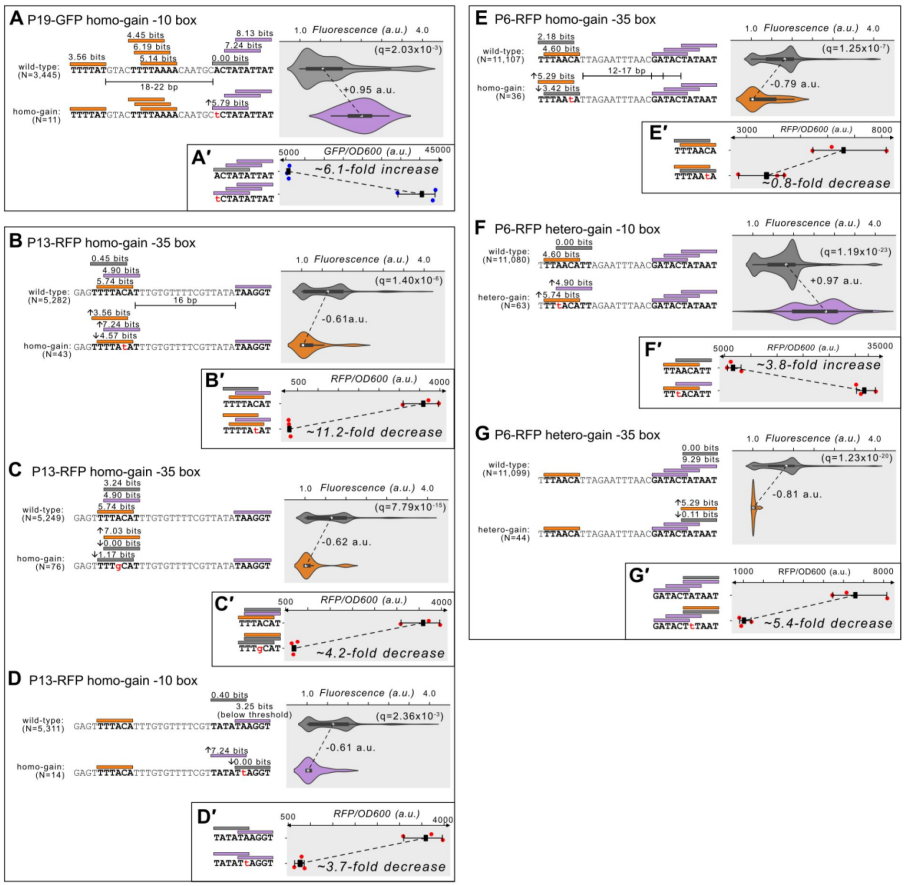


Figure S10.

Additional examples of mutations modulating promoter activity.

(A) Top: Parent P19-GFP and its PWM predictions (orange rectangle = -35 box, magenta rectangle = -10 box, gray rectangle = region of interest). We plot the fluorescence scores of all daughters without a -10 box in the region of interest. Bottom: the most frequent genotype in the dataset where a -10 box is in the region of interest. We plot the fluorescence scores of all daughters with a -10 box in the region of interest. We tested the null hypothesis that the gain of the motif significantly increases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=2.03\times10^{-3}$). Within the violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. **(A')** Top: the fluorescence readouts of three colonies harboring the wild-type P19-GFP reporter construct measured using a plate reader. The y-axis corresponds to the fluorescence readout normalized to the optical density of the culture (OD_{600}) during the reading (see methods). Each point corresponds to the fluorescence of an individual colony. Whiskers correspond to the minimum and maximum values and the dark square the mean fluorescence level. Bottom: analogous to top, but for three colonies harboring the degenerate consensus sequence identified to change fluorescence in panel A. **(B)** Analogous to (A) but for gaining a -35 box in the gray region of interest on P13-RFP (two-tailed MWU test, $q=1.40\times10^{-6}$). **(B')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (B). **(C)** Analogous to (A) but for gaining a -35 box in the gray region of interest on P13-RFP (two-tailed MWU test, $q=7.79\times10^{-15}$). **(C')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (C). **(D)** Analogous to (A) but for gaining a -10 box in the gray region of interest on P13-RFP (two-tailed MWU test, $q=2.36\times10^{-3}$). **(D')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (D). **(E)** Analogous to (A) but for gaining a -35 box in the gray region of interest on P6-RFP (two-tailed MWU test, $q=1.25\times10^{-7}$). **(E')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (E). **(F)** Analogous to (A) but for gaining a -10 box in the gray region of interest on P6-RFP (two-tailed MWU test, $q=1.19\times10^{-23}$). **(F')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (F). **(G)** Analogous to (A) but for gaining a -35 box in the gray region of interest on P6-RFP (two-tailed MWU test, $q=1.23\times10^{-20}$). **(G')** Analogous to (A') but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified in (G).

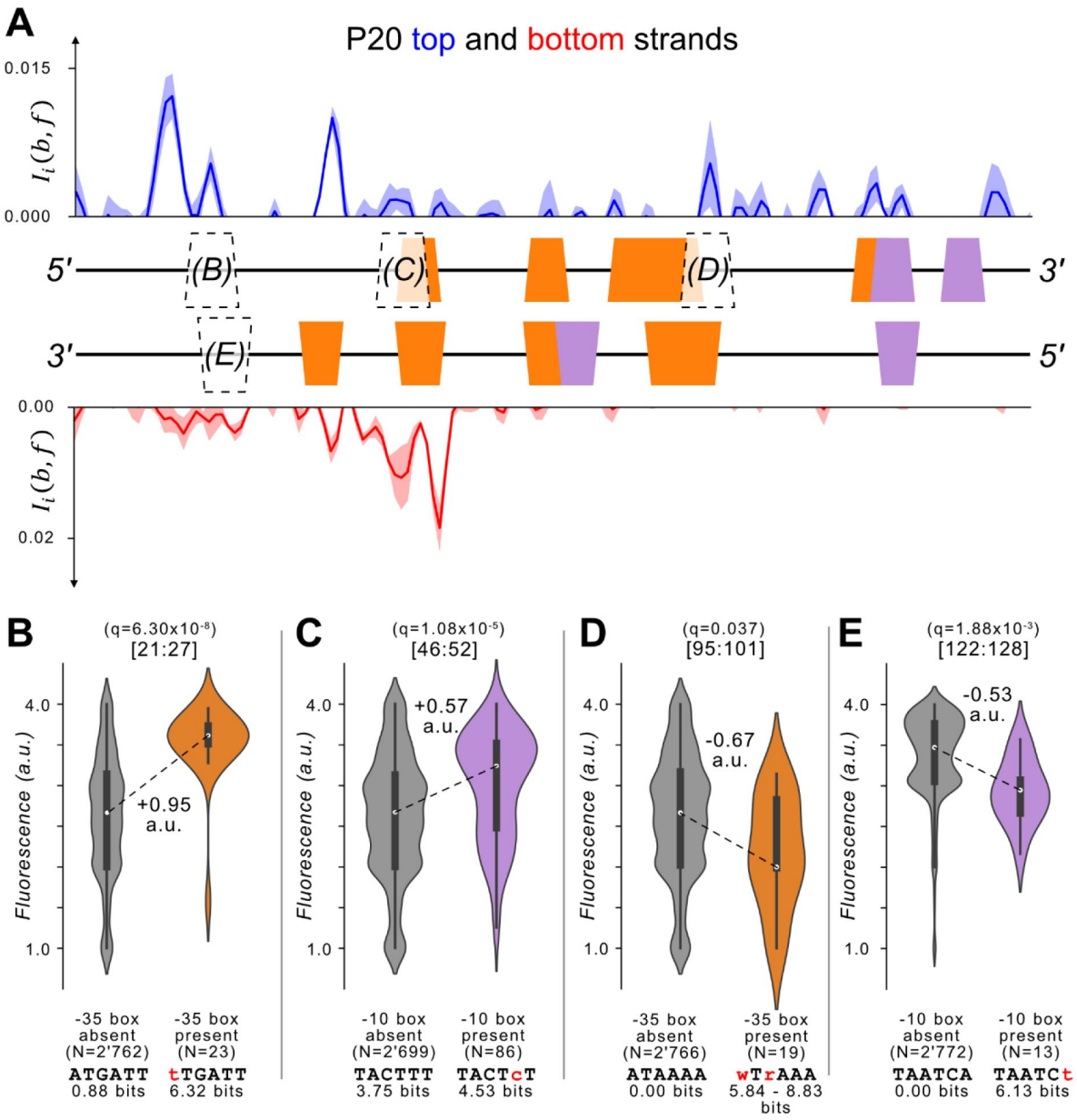


Figure S11.

Template P20.

(A) We plot the mutual information for the top (blue) and bottom (red) strands of template P20. Top: P20-GFP. We plot the mutual information $I_i(b, f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (methods). Middle: position-weight matrix (PWM) predictions for the -10 boxes (magenta trapezoids) and -35 boxes (orange trapezoids) along the wild-type parent sequence. Outlined trapezoids with letters in parenthesis indicate regions of interest for the subsequent figure panel. Bottom: the mutual information for parent P20-RFP as described for P20-GFP. **(B)** We test the null hypothesis that the fluorescence scores have the same central tendency for daughter sequences with vs without -35 boxes at position 21:27 using a two-sided Mann-Whitney U (MWU) test ($q=6.30\times10^{-8}$). The areas of the violin plots are the kernel density estimates (KDE) of each distribution. Within each violin plot is a box plot, where the box represents the interquartile range (IQR) and the white circle the median. Whiskers correspond to 1.5×IQR. The DNA sequences are degenerate consensus sequences in each respective category. Values underneath them correspond to position weight matrix (PWM) bit scores of the degenerate consensus sequence**. (C)** Analogous to (B) but for gaining a -10 box at position 46:52 (two-tailed MWU test, $q=1.08\times10^{-5}$). **(D)** Analogous to (B) but for gaining a -35 box at position 95:101 (two-tailed MWU test, q=0.037). **(E)** Analogous to (B) but for gaining a -10 box at position 122:128 (two-tailed MWU test, $q=1.88\times10^{-3}$).

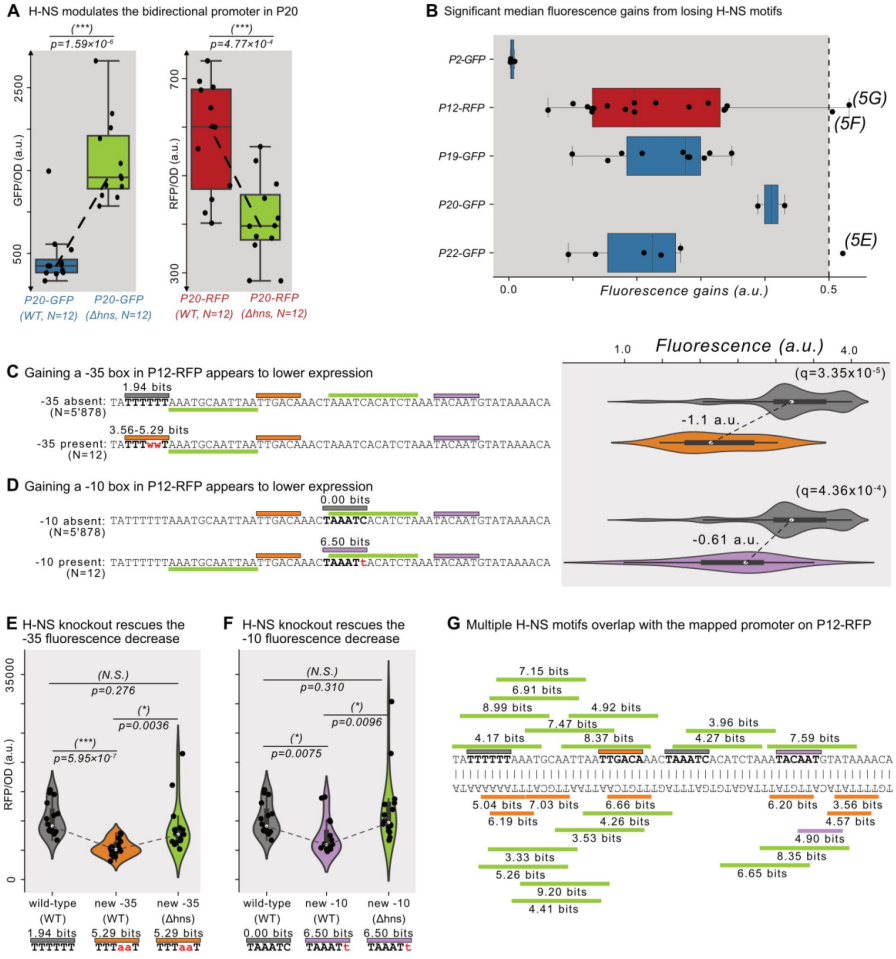


Fig S12.

Additional H-NS sites.

(A) The fluorescence levels of bacteria colonies (N=12 each) harboring a parent sequence in the wild-type (left) vs the *Δhns* background. Boxes represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR. We test the null hypothesis that the means in each background are the same using a two-tailed t-test. Left: P20-GFP levels in the wild-type vs the *Δhns* backgrounds ($p=1.59\times10^{-6}$). Right: P20-RFP levels ($p=4.77\times10^{-4}$). **(B)** The increase in fluorescence (arbitrary units, a.u.) when losing a H-NS motif in mutual information hotspots of P22-GFP and P12-RFP (see **Fig S7** for calculation overview). The dashed line indicates an effect size threshold of + 0.5 a.u. Each circle corresponds to a gain or loss of a motif in a mutual information hotspot (see **Fig 2** for hotspots). Outlined points with letters in parentheses highlight the corresponding figure panels. **(C)** The unmutated sequence of P12-RFP. Orange rectangles = -35 boxes, magenta rectangles = -10 boxes, green rectangles = H-NS motifs. Boxes above the DNA sequence correspond to the same strand as the displayed DNA sequence. Boxes below are on the opposite strand. The gray region corresponds to the region of interest and its current PWM score for the -35 box. Top: Parent P12-RFP and its PWM predictions. We plot the fluorescence scores of all daughters without a -35 box in the gray region of interest. Bottom: the most frequent genotype in the dataset where a -35 box is gained in the region of interest. We plot the fluorescence scores of all daughters with a -35 box in the region of interest. We tested the null hypothesis that gaining a -35 box significantly decreases fluorescence (two-tailed Mann-Whitney U [MWU] test). The q-values correspond to Benjamini-Hochberg-corrected p-values (methods) (two-tailed MWU test, $q=3.35\times10^{-5}$). Within the violin plot is a box plot, in which the box represents the interquartile range (IQR) and the white circle represents the median. Whiskers correspond to 1.5×IQR. **(D)** Analogous to (C) but for the region of interest highlighted in gray, corresponding to a new -10 box (two-tailed MWU test, $q=4.36\times10^{-4}$). **(E)** Analogous to (A) but for constructs from the parent P12-RFP with or without the -35 box shown in (C), in either the wild-type background or the *Δhns* background. N=16 colonies each. **(F)** Analogous to (E) but for constructs with or without the -10 box shown in (D). **(G)** PWM motif predictions and their respective fluorescence scores in bits for P12-RFP. Orange boxes = -35 boxes, magenta boxes = -10 boxes, green boxes = H-NS motifs.

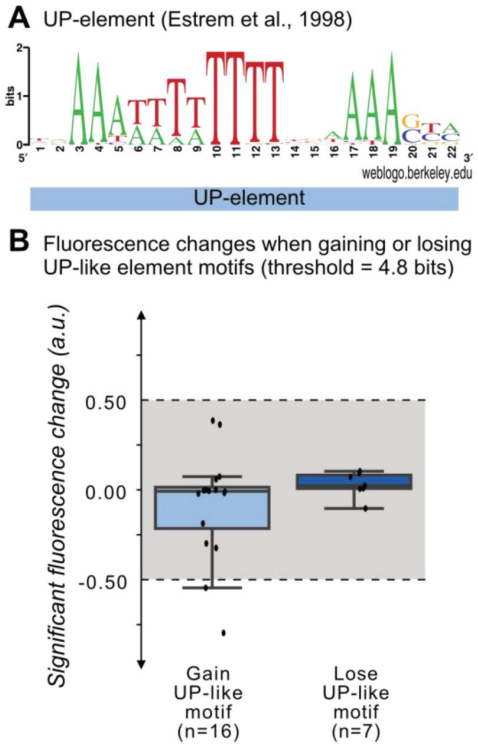


Figure S13.

The change in fluorescence when gaining and losing UP-like motifs.

(A) A sequence logo for the UP-element derived from a position weight matrix for this element (Estrem et al., 1998). **(B)** Change in fluorescence (arbitrary units, a.u.) when losing or gaining UP element motifs (threshold =4.8 bits) in mutational information hotspots of the parent sequences. Each point corresponds to a gain or loss of an UP-element motif in a mutual information hotspot. The boxes in the box plots represent the interquartile range (IQR), the white circle the median, and the whiskers to 1.5×IQR.

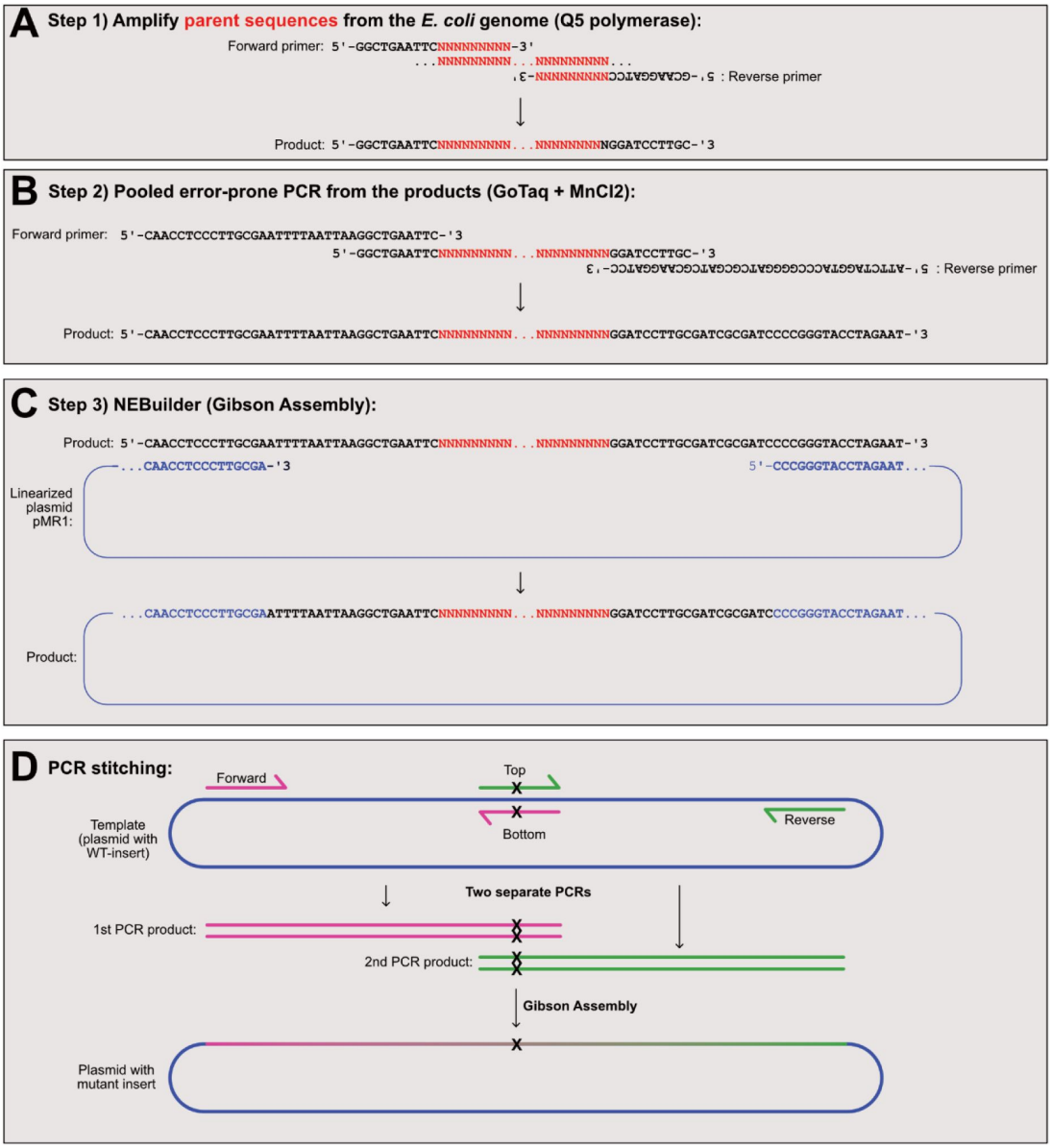


Fig S14.

Molecular cloning of parent sequences.

(A) We first amplify the parent sequences from the *E. coli* genome using Q5 polymerase. Top: the forward and reverse primers contain constant overhang regions, and unique sequences (red N's) unique to the parent sequence. Bottom: the product is a unique parent sequence with the sequences 5'-GGCTGAATTC...insert...GGATCCTTGC-3' concatenated to the flanks. **(B)** We pooled the parent sequences amplified in (A) together for the error-prone polymerase chain reaction using GoTaq and $MnCl_2$. Top: the forward and reverse primers contain constant overhang regions homologous to the pMR1 region for Gibson Assembly. Bottom: the product is the parent sequence flanked by sequences homologous to the pMR1 plasmid. **(C)** We carry out a Gibson Assembly reaction using NEBuilder. Top: we combine the products from (B) with a linear copy of the pMR1 plasmid. Bottom: the final assembly product is the pMR1 plasmid with a mutant library of parent sequences. **(D)** We use PCR stitching to introduce mutations into promoter island inserts of plasmid pMR1. Top: We carry out two PCRs (magenta primers and green primers). The primers "Top" and "Bottom" are complementary and have the same point mutation(s) not present in the template. Middle: the PCR products are double-stranded DNA sequences which share a homologous region harboring the mutation(s). We carry out a Gibson Assembly reaction using NEBuilder with the pMR1 plasmid and the two products. Bottom: the assembly reaction combines the two PCR products together with the desired mutation(s) into the pMR1 plasmid.

References

Afek A, Schipper JL, Horton J, Gordân R, Lukatsky DB (2014) **Protein–DNA binding in the absence of specific base-pair recognition** *Proceedings of the National Academy of Sciences* **111**:17140–17145 https://doi.org/10.1073/pnas.1410569111

Avsec Ž *et al.* (2021) **Base-resolution models of transcription-factor binding reveal soft motif syntax** *Nat Genet* **53**:354–366 https://doi.org/10.1038/s41588-021-00782-6

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection** *Mol Syst Biol* **2**:2006–8 https://doi.org/10.1038/msb4100050

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

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome** *Proc Natl Acad Sci U S A* **99**:757–762 https://doi.org/10.1073/pnas.231608898

Bracco L, Kotlarz D, Kolb A, Diekmann S, Buc H (1989) **Synthetic curved DNA sequences can act as transcriptional activators in Escherichia coli** *EMBO J* **8**:4289–4296

Burgess RR, Travers AA, Dunn JJ, Bautz EK (1969) **Factor stimulating transcription by RNA polymerase** *Nature* **221**:43–46 https://doi.org/10.1038/221043a0

Burr T, Mitchell J, Kolb A, Minchin S, Busby S (2000) **DNA sequence elements located immediately upstream of the –10 hexamer in Escherichia coli promoters: a systematic study** *Nucleic Acids Res* **28**:1864–1870

Bykov A, Glazunova O, Alikina O, Sukharicheva N, Masulis I, Shavkunov K, Ozoline O (2020) **Excessive Promoters as Silencers of Genes Horizontally Acquired by Escherichia coli** *Front Mol Biosci* **7** https://doi.org/10.3389/fmolb.2020.00028

Cazemier AE, Verdoes JC, van Ooyen AJJ, Op den Camp HJM. (1999) **Molecular and Biochemical Characterization of Two Xylanase-Encoding Genes from Cellulomonas pachnodae** *Appl Environ Microbiol* **65**:4099–4107

Chauhan V, Bahrudeen MNM, Palma CSD, Baptista ISC, Almeida BLB, Dash S, Kandavalli V, Ribeiro AS (2022) **Analytical kinetic model of native tandem promoters in E. coli** *PLOS Computational Biology* **18** https://doi.org/10.1371/journal.pcbi.1009824

Chiu T-P, Li J, Jiang Y, Rohs R (2022) **It is in the flanks: Conformational flexibility of transcription factor binding sites** *Biophysical Journal* **121**:3765–3767 https://doi.org/10.1016/j.bpj.2022.09.020

Cock PJA *et al.* (2009) **Biopython: freely available Python tools for computational molecular biology and bioinformatics** *Bioinformatics* **25**:1422–1423 https://doi.org/10.1093/bioinformatics/btp163

Cooper C, Legood S, Wheat RL, Forrest D, Sharma P, Haycocks JRJ, Grainger DC (2024) **H-NS is a bacterial transposon capture protein** *Nat Commun* **15** https://doi.org/10.1038/s41467-024-51407-5

Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH (2006) **Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy** *Nucleic Acids Research* **34**:5416–5425 https://doi.org/10.1093/nar/gkl668

Dillon SC, Dorman CJ (2010) **Bacterial nucleoid-associated proteins, nucleoid structure and gene expression** *Nat Rev Microbiol* **8**:185–195 https://doi.org/10.1038/nrmicro2261

Dornenburg JE, DeVita AM, Palumbo MJ, Wade JT (2010) **Widespread Antisense Transcription in Escherichia coli** *mBio* **1** https://doi.org/10.1128/mbio.00024-10

Estrem ST, Gaal T, Ross W, Gourse RL (1998) **Identification of an UP element consensus sequence for bacterial promoters** *Proc Natl Acad Sci U S A* **95**:9761–9766

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS (2015) **Suboptimization of developmental enhancers** *Science* **350**:325–328 https://doi.org/10.1126/science.aac6948

Forrest D, Warman EA, Erkelens AM, Dame RT, Grainger DC (2022) **Xenogeneic silencing strategies in bacteria are dictated by RNA polymerase promiscuity** *Nat Commun* **13** https://doi.org/10.1038/s41467-022-28747-1

Fuqua T *et al.* (2020) **Dense and pleiotropic regulatory information in a developmental enhancer** *Nature* **587**:235–239 https://doi.org/10.1038/s41586-020-2816-5

Fuqua Wagner (2023) **Mobile DNA is replete with hotspots for the de novo emergence of gene regulation** *bioRxiv* https://doi.org/10.1101/2023.10.22.563463

Galupa R *et al.* (2023) **Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development** *Developmental Cell* **58**:51–62 https://doi.org/10.1016/j.devcel.2022.12.003

Gama-Castro S *et al.* (2016) **RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond** *Nucleic Acids Res* **44** https://doi.org/10.1093/nar/gkv1156

Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML (2013) **Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape** *Cell Rep* **3**:1093–1104 https://doi.org/10.1016/j.celrep.2013.03.014

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I (2010) **Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers** *Genome Res* **20**:565–577 https://doi.org/10.1101/gr.104471.109

Guazzaroni M-E, Silva-Rocha R (2014) **Expanding the Logic of Bacterial Promoters Using Engineered Overlapping Operators for Global Regulators** *ACS Synth Biol* **3**:666–675 https://doi.org/10.1021/sb500084f

Hertz GZ, Stormo GD (1999) **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences** *Bioinformatics* **15**:563–577 https://doi.org/10.1093/bioinformatics/15.7.563

Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ (2012) **RNA polymerase II collision interrupts convergent transcription** *Mol Cell* **48**:365–374 https://doi.org/10.1016/j.molcel.2012.08.027

Hommais F, Krin E, Laurent-Winter C, Soutourina O, Malpertuy A, Le Caer JP, Danchin A, Bertin P (2001) **Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS** *Mol Microbiol* **40**:20–36 https://doi.org/10.1046/j.1365-2958.2001.02358.x

Huerta AM, Collado-Vides J (2003) **Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals** *Journal of Molecular Biology* **333**:261–278 https://doi.org/10.1016/j.jmb.2003.07.017

Inukai S, Kock KH, Bulyk ML (2017) **Transcription factor–DNA binding: beyond binding site motifs** *Curr Opin Genet Dev* **43**:110–119 https://doi.org/10.1016/j.gde.2017.02.007

Ireland WT, Beeler SM, Flores-Bautista E, McCarty NS, Röschinger T, Belliveau NM, Sweredoski MJ, Moradian A, Kinney JB, Phillips R (2020) **Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time** *eLife* **9** https://doi.org/10.7554/eLife.55308

Jacob F, Monod J (1961) **Genetic regulatory mechanisms in the synthesis of proteins** *Journal of Molecular Biology* **3**:318–356 https://doi.org/10.1016/S0022-2836(61)80072-7

Jiang D, Cope AL, Zhang J, Pennell M (2023) **On the Decoupling of Evolutionary Changes in mRNA and Protein Levels** *Mol Biol Evol* **40** https://doi.org/10.1093/molbev/msad169

Kapranov P *et al.* (2007) **RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription** *Science* **316**:1484–1488 https://doi.org/10.1126/science.1138341

Kinney JB, Murugan A, Callan CG, Cox EC (2010) **Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence** *Proceedings of the National Academy of Sciences* **107**:9158–9163 https://doi.org/10.1073/pnas.1004290107

Klein CA, Teufel M, Weile CJ, Sobetzko P (2021) **The bacterial promoter spacer modulates promoter strength and timing by length, TG-motifs and DNA supercoiling sensitivity** *Sci Rep* **11** https://doi.org/10.1038/s41598-021-03817-4

Krieger G, Lupo O, Wittkopp P, Barkai N (2022) **Evolution of transcription factor binding through sequence variations and turnover of binding sites** *Genome Res* **32**:1099–1111 https://doi.org/10.1101/gr.276715.122

LaFleur TL, Hossain A, Salis HM (2022) **Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria** *Nat Commun* **13** https://doi.org/10.1038/s41467-022-32829-5

Lagator M, Sarikas S, Steinrueck M, Toledo-Aparicio D, Bollback JP, Guet CC, Tkačik G (2022) **Predicting bacterial promoter function and evolution from random sequences** *eLife* **11** https://doi.org/10.7554/eLife.64543

Lässig M, Mustonen V, Walczak AM (2017) **Predicting evolution** *Nat Ecol Evol* **1**:1–9 https://doi.org/10.1038/s41559-017-0077

Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA (2003) **Homotypic Regulatory Clusters in Drosophila** *Genome Res* **13**:579–588 https://doi.org/10.1101/gr.668403

Loell KJ, Friedman RZ, Myers CA, Corbo JC, Cohen BA, White MA (2024) **Transcription factor interactions explain the context-dependent activity of CRX binding sites** *PLoS Comput Biol* **20** https://doi.org/10.1371/journal.pcbi.1011802

Lynch VJ *et al.* (2015) **Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy** *Cell Reports* **10**:551–561 https://doi.org/10.1016/j.celrep.2014.12.052

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

Mandal M, Breaker RR (2004) **Gene regulation by riboswitches** *Nat Rev Mol Cell Biol* **5**:451–463 https://doi.org/10.1038/nrm1403

Martínez-Antonio A, Collado-Vides J (2003) **Identifying global regulators in transcriptional regulatory networks in bacteria** *Current Opinion in Microbiology* **6**:482–489 https://doi.org/10.1016/j.mib.2003.09.002

Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, Melé M (2020) **Cis and trans effects differentially contribute to the evolution of promoters and enhancers** *Genome Biology* **21** https://doi.org/10.1186/s13059-020-02110-3

Mendoza-Vargas A *et al.* (2009) **Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli** *PLoS One* **4** https://doi.org/10.1371/journal.pone.0007526

Meng W, Belyaeva T, Savery NJ, Busby SJW, Ross WE, Gaal T, Gourse RL, Thomas MS (2001) **UP element-dependent transcription at the Escherichia coli rrnB P1 promoter: positional requirements and role of the RNA polymerase α subunit linker** *Nucleic Acids Res* **29**:4166–4178

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ (2016) **Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression** *Molecular Biology and Evolution* **33**:1131–1146 https://doi.org/10.1093/molbev/msw011

Mitchell JE, Zheng D, Busby SJW, Minchin SD (2003) **Identification and analysis of 'extended –10' promoters in Escherichia coli** *Nucleic Acids Res* **31**:4689–4695

Mohanty BK, Kushner SR (2006) **The majority of Escherichia coli mRNAs undergo post-transcriptional modification in exponentially growing cells** *Nucleic Acids Res* **34**:5695–5704 https://doi.org/10.1093/nar/gkl684

Navarre WW, McClelland M, Libby SJ, Fang FC (2007) **Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA** *Genes Dev* **21**:1456–1471 https://doi.org/10.1101/gad.1543107

Oshima T, Ishikawa S, Kurokawa K, Aiba H, Ogasawara N (2006) **Escherichia coli Histone-Like Protein H-NS Preferentially Binds to Horizontally Acquired DNA in Association with RNA Polymerase** *DNA Research* **13**:141–153 https://doi.org/10.1093/dnares/dsl009

Paget MS, Helmann JD (2003) **The σ70family of sigma factors** *Genome Biology* **4** https://doi.org/10.1186/gb-2003-4-1-203

Panyukov VV, Ozoline ON (2013) **Promoters of Escherichia coli versus Promoter Islands: Function and Structure Comparison** *PLOS One* **8** https://doi.org/10.1371/journal.pone.0062601

Perales M, Rodriguez K, Snipes S, Yadav RK, Diaz-Mendoza M, Reddy GV (2016) **Threshold-dependent transcriptional discrimination underlies stem cell homeostasis** *Proc Natl Acad Sci U S A* **113**:E6298–E6306 https://doi.org/10.1073/pnas.1607669113

Peterman N, Levine E (2016) **Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations** *BMC Genomics* **17** https://doi.org/10.1186/s12864-016-2533-5

Pribnow D (1975) **Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter** *Proc Natl Acad Sci U S A* **72**:784–788

Prud'homme B, Gompel N, Carroll SB (2007) **Emerging principles of regulatory evolution** *Proceedings of the National Academy of Sciences* **104**:8605–8612 https://doi.org/10.1073/pnas.0700488104

Purtov YA, Glazunova OA, Antipov SS, Pokusaeva VO, Fesenko EE, Preobrazhenskaya EV, Shavkunov KS, Tutukina MN, Lukyanov VI, Ozoline ON (2014) **Promoter islands as a platform for interaction with nucleoid proteins and transcription factors** *J Bioinform Comput Biol* **12** https://doi.org/10.1142/S0219720014410066

Rebeiz M, Jikomes N, Kassner VA, Carroll SB (2011) **Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences** *Proceedings of the National Academy of Sciences* **108**:10036–10043 https://doi.org/10.1073/pnas.1105937108

Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, Severinov K, Gourse RL (1993) **A Third Recognition Element in Bacterial Promoters: DNA Binding by the α Subunit of RNA Polymerase** *Science* **262**:1407–1413 https://doi.org/10.1126/science.8248780

Seshasayee ASN, Sivaraman K, Luscombe NM (2011) **An overview of prokaryotic transcription factors : a summary of function and occurrence in bacterial genomes** *Subcell Biochem* **52**:7–23 https://doi.org/10.1007/978-90-481-9069-0_2

Shavkunov KS, Masulis IS, Tutukina MN, Deev AA, Ozoline ON (2009) **Gains and unexpected lessons from genome-scale promoter mapping** *Nucleic Acids Res* **37**:4919–4931 https://doi.org/10.1093/nar/gkp490

Shimada T, Bridier A, Briandet R, Ishihama A (2011) **Novel roles of LeuO in transcription regulation of E. coli genome: antagonistic interplay with the universal silencer H-NS** *Mol Microbiol* **82**:378–397 https://doi.org/10.1111/j.1365-2958.2011.07818.x

Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC (2014) **Widespread suppression of intragenic transcription initiation by H-NS** *Genes Dev* **28**:214–219 https://doi.org/10.1101/gad.234336.113

Song E, Uhm H, Munasingha PR, Hwang S, Seo Y-S, Kang JY, Kang C, Hohng S (2022) **Rho-dependent transcription termination proceeds via three routes** *Nat Commun* **13** https://doi.org/10.1038/s41467-022-29321-5

Stern DL, Orgogozo V (2008) **The loci of evolution: how predictable is genetic evolution?** *Evolution* **62**:2155–2177 https://doi.org/10.1111/j.1558-5646.2008.00450.x

Tierrafría VH *et al.* (2022) **RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12** *Microbial Genomics* **8** https://doi.org/10.1099/mgen.0.000833

Uptain SM, Chamberlin MJ (1997) **Escherichia coli RNA polymerase terminates transcription efficiently at rho-independent terminators on single-stranded DNA templates** *Proceedings of the National Academy of Sciences* **94**:13548–13553 https://doi.org/10.1073/pnas.94.25.13548

Urtecho G, Insigne KD, Tripp AD, Brinck MS, Lubock NB, Acree C, Kim H, Chan T, Kosuri S (2023) **Genome-wide Functional Characterization of Escherichia coli Promoters and Sequence Elements Encoding Their Regulation** *eLife* **12** https://doi.org/10.7554/eLife.92558

Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S (2019) **Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli** *Biochemistry* **58**:1539–1551 https://doi.org/10.1021/acs.biochem.7b01069

van Hijum SAFT, Medema MH, Kuipers OP (2009) **Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation** *Microbiology and Molecular Biology Reviews* **73**:481–509 https://doi.org/10.1128/MMBR.00037-08

Villanueva-Cañas JL, Horvath V, Aguilera L, González J (2019) **Diverse families of transposable elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster** *Nucleic Acids Res* **47**:6842–6857 https://doi.org/10.1093/nar/gkz490

Villar D *et al.* (2015) **Enhancer evolution across 20 mammalian species** *Cell* **160**:554–566 https://doi.org/10.1016/j.cell.2015.01.006

Wade JT, Grainger DC (2014) **Pervasive transcription: illuminating the dark matter of bacterial transcriptomes** *Nat Rev Microbiol* **12**:647–653 https://doi.org/10.1038/nrmicro3316

Wang L, Watters JW, Ju X, Lu G, Liu S (2023) **Head-on and co-directional RNA polymerase collisions orchestrate bidirectional transcription termination** *Molecular Cell* **83**:1153–1164 https://doi.org/10.1016/j.molcel.2023.02.017

Wang Y, Wang H, Wei L, Li S, Liu L, Wang X (2020) **Synthetic promoter design in Escherichia coli based on a deep generative network** *Nucleic Acids Research* **48**:6403–6412 https://doi.org/10.1093/nar/gkaa325

Warman EA, Forrest D, Guest T, Haycocks JJRJ, Wade JT, Grainger DC (2021) **Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA-sequence symmetry** *Nat Microbiol* **6**:746–756 https://doi.org/10.1038/s41564-021-00898-9

Warman EA, Singh SS, Gubieda AG, Grainger DC (2020) **A non-canonical promoter element drives spurious transcription of horizontally acquired bacterial genes** *Nucleic Acids Res* **48**:4891–4901 https://doi.org/10.1093/nar/gkaa244

Westmann CA, Alves L de F, Silva-Rocha R, Guazzaroni M-E (2018) **Mining Novel Constitutive Promoter Elements in Soil Metagenomic Libraries in Escherichia coli** *Frontiers in Microbiology* **9**

Yona AH, Alm EJ, Gore J (2018) **Random sequences rapidly evolve into de novo promoters** *Nat Commun* **9** https://doi.org/10.1038/s41467-018-04026-w

Yue F *et al.* (2014) **A comparative encyclopedia of DNA elements in the mouse genome** *Nature* **515**:355–364 https://doi.org/10.1038/nature13992

Author information

Timothy Fuqua
Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland
ORCID iD: 0000-0003-4005-3329

Yiqiao Sun
Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland

Andreas Wagner
Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland, The Sante Fe Institute, Sante Fe, USA

For correspondence: andreas.wagner@ieu.uzh.ch

Editors

Reviewing Editor
Vincent Lynch
University at Buffalo, State University of New York, Buffalo, United States of America

Senior Editor
Alan Moses
University of Toronto, Toronto, Canada

Reviewer #1 (Public review):

Summary:

This study by Fuqua et al. studies the emergence of sigma70 promoters in bacterial genomes. While there have been several studies to explore how mutations lead to promoter activity,

this is the first to explore this phenomena in a wide variety of backgrounds, which notably contain a diverse assortment of local sigma70 motifs in variable configurations. By exploring how mutations affect promoter activity in such diverse backgrounds, they are able to identify a variety of anecdotal examples of gain/loss of promoter activity and propose several mechanisms for how these mutations are interacting within the local motif landscape. Ultimately, they show how different sequences have different probabilities of gaining/losing promoter activity and may do so through a variety of mechanisms.

Major strengths and weaknesses of the methods and results:

This study uses Sort-Seq to characterize promoter activity, which has been adopted by multiple groups and shown to be robust. Furthermore, they use a slightly altered protocol which allows measurements of bi-directional promoter activity. This combined with their pooling strategy allows them to characterize expression of many different backgrounds in both directions in extremely high-throughput which is impressive! A second key approach this study relies on is the identification of promoter motifs using position weight matrices (PWMs). While these methods are prone to false positives, the authors implement a systematic approach which is standard in the field. However, drawing these types of binary definitions (is this a motif? yes/no) should always come with the caveat that gene expression is quantitative traits that we oversimplify when drawing boundaries.

Their approach to randomly mutagenize promoters allowed them to find many examples of different types of evolutions that may occur to increase or decrease promoter activity. They have supported these with validations in more controlled backgrounds which convincingly support their proposed mechanisms for promoter evolution.

An appraisal of whether the authors achieved their aims, and whether the results support their conclusions:

The authors express a key finding that the specific landscape of promoter motifs in a sequence affect the likelihood that local mutations create or destroy regulatory elements. The authors have described many examples, including several that are non-obvious, and show convincingly that different sequence backgrounds have different probabilities for gaining or losing promoter activity. This overarching conclusion is supported by trend and mechanistic data which show differences in probabilities of evolving promoters, as well as the mechanisms underlying these evolutions. Furthermore, these mutations are well described and presented, showing the strength of emergent promoter motifs and their specific spacings from existing motifs within the sequence.

Impact of the work on the field, and the utility of the methods and data to the community:

This study enhances our understanding of the diverse mechanisms by which promoters can evolve or devolve, potentially improving models that predict mutational outcomes. While this study reveals complex mutational patterns, modeling them could significantly advance our ability to predict bacterial evolutionary trajectories and interpret genomes, bringing us closer to that goal.

Recent work in the field of bacterial gene regulation has raised interest in bidirectional promoter regions. While the authors do not discuss how mutations that raise expression in one direction may affect another, they have created an expansive dataset which may enable other groups to study this interesting phenomenon. Also, their variation of the Sort-Seq protocol will be a valuable example for other groups who may be interested in studying bidirectional expression. Lastly, this study may be of interests to groups studying eukaryotic regulation as it can inform how the evolution of transcription factor binding sites influences short-range interactions with local regulator elements.

Any additional context to understand the significance of the work:

Predicting whether a sequence drives promoter activity is a challenging task. By learning the types of mutations that create or destroy promoters, this study provides valuable insights for computational models aimed at predicting promoter activity.

Comments on revised version:

I am satisfied with the extensive changes made by the author. This manuscript is excellent.

I very much like the change in figures to incorporate the sequence information. It is great to see clear representations of the emergent sigma70 motifs and their spacing relative to existing motifs. This addition significantly improves the clarity of the findings.

The validation of mutations on a clean background is well-executed, and the results are convincing. I appreciate the effort put into validating their results. The additional analyses that include TGn and UP-element motifs are also well done and highly relevant, as these elements are known to compensate for weaker or absent -35 sequences.

Most or all perceived inconsistencies from the previous version have been resolved. While I don't think the fluorescence threshold of 1.5 a.u. for promoter activity is justified, the authors do acknowledge this shortcoming, and even empirically-derived thresholds are still technically arbitrary.

I particularly enjoyed Figure 1E, thank you for entertaining my analysis request! Also, the H-NS story is a nice addition showing how transcription factors influence this evolution

Overall, this revised manuscript is an excellent contribution to the field, and I have no further recommendations for improvement.

https://doi.org/10.7554/eLife.98654.2.sa3

Reviewer #2 (Public review):

Summary:

Fuqua et al investigated the relationship between prokaryotic box motifs and the activation of promoter activity using a mutagenesis sequencing approach. From generating thousands of mutant daughter sequences from both active and non-active promoter sequences they were able to produce a fantastic dataset to investigate potential mechanisms for promoter activation. From these large numbers of mutated sequences, they were able to generate mutual information with gene expression to identify key mutations relating to the activation of promoter island sequences.

Strengths:

The data generated from this paper is an important resource to address this question of promoter activation. Being able to link promoter modulated gene expression to mutational changes in previously nonactive promoter regions is exciting. This approach allows future large-scale studies to investigate evolutionary processes relating to changes in gene regulation in a statistically robust manner. Here there is a focus on the -10 and -35 boxes but other elements and interactions were explored including; H-NS binding, UP-element and TGn. Alongside this, the method of identifying key mutations using mutual information in this paper is well done and should be a standard in future studies for identifying regions of interest.

Weaknesses:

While the generation of the data is superb, as the authors have stated clearly themselves, there is a lot of scope for future studies to understand both causal relationships and utilise the data more effectively. The authors look at changes in regulatory expression based on a few observations that are treated independently but occur concurrently. While this study has backed up findings experimentally this may not always be possible. Previously this reviewer had suggested addressing this using complementary approaches such as analysis focusing on identifying important motifs, using something like a glm lasso regression to identify significant motifs, and then combining with mutational hotspot information would be more robust. The authors tried to implement such an approach in response to the review, but its complexity became beyond the scope. I look forward to the development of such methods that allow more complete exploration of similar datasets.

Comments on revised version:

The authors addressed all my previous comments. I believe the study is much improved and thank them for the time and effort they put into addressing the comments.

https://doi.org/10.7554/eLife.98654.2.sa2

Reviewer #3 (Public review):

This work brings a computational approach to the study of promoters and transcription. The paper is improved but there are still factual errors and implausible explanations. I am not convinced by the response from the authors, concerning the promoter -35 element, in their rebuttal.

Comments on author rebuttal:

- We respectfully but strongly disagree that our analysis has misrepresented the true nature of -35 boxes. First, accounting for more A's at position 5 in the PWM is not going to lead to a "critical error." This is because positions 4-6 of the motif barely have any information content (bits) compared to positions 1-3 (see Fig 1A).

The analysis does misrepresent the consensus -35 element, which is, unequivocally, TTGACA. I agree that positions 4-6 of the element are less well-conserved.

- This assertion is not just based on our own PWM, but based on ample precedent in the literature. In PMID 14529615, TTG is present in 38% of all -35 boxes, but ACA only in 8%.

This does not mean that TTGACA is not the consensus, or that "ACA" is not important at promoters where it's present.

- In PMID 29388765, with the -10 instance TATAAT, the -35 instance TTGCAA yields stronger promoters compared to the -35 instance TTGACA (See their Figure 3B).

This is a known phenomenon and results from "perfect" promoters being limited at the point of RNA polymerase promoter escape (because the RNAP struggles to "let go" of perfect promoters). This does not mean the TTGACA is not the consensus. Indeed, and this is a key point, it is evident in the figure the authors refer to that TTGACA stimulates more transcription than alternative -35 sequences when -10 elements are not perfect.

- In PMID 29745856 (Figure 2), the most information content lies in positions 1-3, with the A and C at position 5 both nearly equally represented, as in our PWM.

The motif shown in this paper suffers from exactly the same issue as the paper under review; the variable spacing between the -35 hexamer and -10 element isn't taken into account by

MEME.

- In PMID 33958766 (Figure 1) an experimentally-derived -35 box is even reduced to a "partial" -35 box which only includes positions 1 and 2, with consensus: TTnnnn.

This paper does not show an "experimentally-derived -35 box" in Figure 1 (or anywhere else, as far as I can see).

- In addition, we did not derive the PWMs as the reviewer describes. The PWMs we use are based on computational predictions that are in excellent agreement with experimental results. Specifically, the PWMs we use are from PMID 29728462, which acquired 145 -10 and -35 box sequences from the top 3.3% of computationally predicted boxes from Regulon DB.

The paper mentioned states "for the genomic RNAP logo, sequences were taken from computationally predicted RNAP binding sites on RegulonDB" so these are not experimentally defined promoters? It's not obvious from the paper, or regulon DB, which sequences these are or how they were predicted.

- Thank you for pointing out that our original submission was incomplete in this regard. We address these concerns by new analyses, including some new experiments. First, Rho dependent termination is associated with the RUT motif, which is very rich in Cytosines (PMID: 30845912). Given that our sequences confer between 65%-78% of AT-content, canonical rho dependent termination is unlikely. However, we computationally searched for rho-dependent terminators using the available code from PMID: 30845912, but the algorithm did not identify any putative RUTs. Because this analysis was not informative, we did not include it in the paper.

I don't believe it is the case that Rho absolutely requires a RUT sequence. My understanding is that, if an RNA is not translated, Rho will intervene (e.g. see PMID: 18487194).

- We respectfully disagree that the reviewer's point is pertinent because what the reviewer is referring to is the likelihood that the sequence is a promoter, which indeed increases with AT content, but we are focused on the likelihood that a sequence becomes a promoter through DNA mutation

I disagree that this distinction is relevant. An AT-rich sequence will much more closely resemble a promoter by chance than a GC rich sequence. As an extreme example, the sequence TTTTTT can be converted into a reasonable -10 element by one change (to TATTTT) but the sequence GGGGGG can't.

https://doi.org/10.7554/eLife.98654.2.sa1

Author response:

The following is the authors' response to the original reviews.

We performed multiple new experiments and analyses in response to the reviewers concerns, and incorporated the results of these analyses in the main text, and in multiple substantially revised or new figures. Before embarking on a point-by-point reply to the reviewers' concerns, we here briefly summarize our most important revisions.

First, we addressed a concern shared by Reviewers #1-3 about a lack of information about our DNA sequences. To this end, we redesigned multiple figures (Figures 3, 4, 5, S8, S9, S10, S11, and S12) to include the DNA sequences of each tested promoter, the specific mutations that occurred in it, the resulting changes in position-weight-matrix (PWM) scores, and the spacing between promoter motifs. Second, Reviewers #1 and #2 raised concerns about a lack of validation of our computational predictions and the resulting incompleteness of the

manuscript. To address this issue, we engineered 27 reporter constructs harboring specific mutations, and experimentally validated our computational predictions with them. Third, we expanded our analysis to study how a more complete repertoire of other sigma 70 promoter motifs such as the UP-element and the extended -10 / TGn motif affects gene expression driven by the promoters we study. Fourth, we addressed concerns by Reviewer #3 about the role of the Histone-like nucleoid-structuring protein (H-NS) in promoter emergence and evolution. We did this by performing both experiments and computational analyses, which are now shown in the newly added Figure 5. Fifth, to satisfy Reviewer #3's concerns about missing details in the Discussion, we have rewritten this section, adding additional details and references.

We next describe these and many other changes in a point-by-point reply to each reviewer's comments. In addition, we append a detailed list of changes to each section and figure to the end of this document.

> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *This study by Fuqua et al. studies the emergence of sigma70 promoters in bacterial genomes. While there have been several studies to explore how mutations lead to promoter activity, this is the first to explore this phenomenon in a wide variety of backgrounds, which notably contain a diverse assortment of local sigma70 motifs in variable configurations. By exploring how mutations affect promoter activity in such diverse backgrounds, they are able to identify a variety of anecdotal examples of gain/loss of promoter activity and propose several mechanisms for how these mutations interact within the local motif landscape. Ultimately, they show how different sequences have different probabilities of gaining/losing promoter activity and may do so through a variety of mechanisms.*

We thank Reviewer #1 for taking the time to read and provide critical feedback on our manuscript. Their summary is fundamentally correct.

> *Major strengths and weaknesses of the methods and results:*
>
> *This study uses Sort-Seq to characterize promoter activity, which has been adopted by multiple groups and shown to be robust. Furthermore, they use a slightly altered protocol that allows measurements of bi-directional promoter activity. This combined with their pooling strategy allows them to characterize expressions of many different backgrounds in both directions in extremely high throughput which is impressive! A second key approach this study relies on is the identification of promoter motifs using position weight matrices (PWMs). While these methods are prone to false positives, the authors implement a systematic approach which is standard in the field. However, drawing these types of binary definitions (is this a motif? yes/no) should always come with the caveat that gene expression is a quantitative trait that we oversimplify when drawing boundaries.*

The point is well-taken. To clarify this and other issues, we have added a section on the limitations of our work to the Discussion. Within this section we include the following sentences (lines 675-680):

"Additionally, future studies will be necessary to address the limitations of our own work. First, we use binary thresholding to determine i) the presence or absence of a motif, ii) whether a sequence has promoter activity or not, and iii) whether a part of a sequence is a hotspot or not. While chosen systematically, the thresholds we use for these decisions may cause us to miss subtle but important aspects of promoter evolution and emergence."

> *Their approach to randomly mutagenizing promoters allowed them to find many anecdotal examples of different types of evolutions that may occur to increase or decrease promoter activity. However, the lack of validation of these phenomena in more controlled backgrounds may require us to further scrutinize their results. That is, their explanations for why certain mutations lead or obviate promoter activity may be due to interactions with other elements in the 'messy' backgrounds, rather than what is proposed.*

Thank you for raising this important point. To address it, we have conducted extensive new validation experiments for the newest version of this manuscript. For the “anecdotal” examples you described, we created 27 reporter constructs harboring the precise mutation that leads to the loss or gain of gene expression, and validated its ability to drive gene expression. The results from these experiments are in Figures 3, 4, 5, and Supplemental Figures S8-S11, and are labeled with a ' (prime) symbol.

These experiments not only confirm the increases and decreases in fluorescence that our analysis had predicted. They also demonstrate, with the exception of two (out of 27) falsepositive discoveries, that background mutations do not confound our analysis. We mention these two exceptions (lines 364-367):

“In two of these hotspots, our validation experiments revealed no substantial difference in gene expression as a result of the hotspot mutation (Fig S8F' and Fig S8J'). In both of these false positives, new -10 boxes emerge in locations without an upstream -35 box.”

> *An appraisal of whether the authors achieved their aims, and whether the results support their conclusions:*
>
> *The authors express a key finding that the specific landscape of promoter motifs in a sequence affects the likelihood that local mutations create or destroy regulatory elements. The authors have described many examples, including several that are non-obvious, and show convincingly that different sequence backgrounds have different probabilities for gaining or losing promoter activity. While this overarching conclusion is supported by the manuscript, the proposed mechanisms for explaining changes in promoter activity are not sufficiently validated to be taken for absolute truth. There is not sufficient description of the strength of emergent promoter motifs or their specific spacings from existing motifs within the sequence. Furthermore, they do not define a systematic process by which mutations are assigned to different categories (e.g. box shifting, tandem motifs, etc.) which may imply that the specific examples are assigned based on which is most convenient for the narrative.*

To summarize, Reviewer #1 criticizes the following three aspects of our work in this comment. 1) The mechanisms we proposed are not sufficiently validated. 2) The description of motifs, spacing, and PWM scores are not shown. 3) How mutations are classified into different categories (i.e. box-shifting, tandem motifs, etc.) is not systematically defined.

These are all valid criticisms. In response, we performed an extensive set of follow-up experiments and analyses, and redesigned the majority of the figures. Here is a more detailed response to each criticism:

(1) Proposed mechanisms for explaining changes in promoter activity are not sufficiently validated. We engineered 27 reporter constructs harboring the specific mutations in the parents that we had predicted to change promoter activity. For each, we compared their fluorescence levels with their wild-type counterpart. The results from these experiments are in Figures 3 and 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12, and are labeled with a ' (prime) symbol.

(2) No sufficient description of the strength of emergent promoter motifs or their specific spacings. We redesigned the figures to include the DNA sequences of the parent sequences, as well as the degenerate consensus sequences for each mutation. We additionally now highlight the specific motif sequences, their respective PWM scores, and by how much the score changes upon mutation. Finally, we annotated the spacing of motifs. These changes are in Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12.

We note that in many cases, high-scoring PWM hits for the same motif can overlap (i.e. two -10 motifs or two -35 motifs overlap). Additionally, the proximity of a -35 and -10 box does not guarantee that the two boxes are interacting. Together, these two facts can result in an ambiguity of the spacer size between two boxes. To avoid any reporting bias, we thus often report spacer sizes as a range (see Figure panels 4F, S8D, S8F-L, S9A, S9H, S10A, and S10E). The smallest spacer we annotate is in Figure 4F with 10 bp, and the largest is in Figure S8D with 26 bp. Any more "extreme" distances are not annotated and for the reader to decide if an interaction is present or not.

(3) No systematic process by which mutations are assigned to different categories such as box shifting, tandem motifs, etc. We opted to reformulate these categories completely, because the phenotypic effects of a previously mentioned "tandem motif" was actually a byproduct of H-NS repression (see the newly added Figure S12).

We also agree that the categories were ambiguous. We now introduce two terms: homo-gain and hetero-gain of -10 and -35 boxes. The manuscript now clearly defines these terms, and the relevant passage now reads as follows (lines 430-435):

"We found that these mutations frequently create new boxes overlapping those we had identified as part of a promoter

(Fig S9). This occurs when mutations create a -10 box overlapping a -10 box, a -35 box overlapping a -35 box, a -10 box overlapping a -35 box, or a -35 box overlapping a -10 box. We call the resulting event a "homo-gain" when the new box is of the same type as the one it overlaps, and otherwise a "hetero-gain". In either case, the creation of the new box does not always destroy the original box."

> *Impact of the work on the field, and the utility of the methods and data to the community: From this study, we are more aware of different types of ways promoters can evolve and devolve, but do not have a better ability to predict when mutations will lead to these effects. Recent work in the field of bacterial gene regulation has raised interest in bidirectional promoter regions. While the authors do not discuss how mutations that raise expression in one direction may affect another, they have created an expansive dataset that may enable other groups to study this interesting phenomenon. Also, their variation of the Sort-Seq protocol will be a valuable example for other groups who may be interested in studying bidirectional expression. Lastly, this study may be of interest to groups studying eukaryotic regulation as it can inform how the evolution of transcription factor binding sites influences short-range interactions with local regulator elements. Any additional context to understand the significance of the work:*
>
> *The task of computationally predicting whether a sequence drives promoter activity is difficult. By learning what types of mutations create or destroy promoters from this study, we are better equipped for this task.*

We thank Reviewer #1 again for their time and their thoughtful comments.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *Fuqua et al investigated the relationship between prokaryotic box motifs and the activation of promoter activity using a mutagenesis sequencing approach. From generating thousands of mutant daughter sequences from both active and non-active promoter sequences they were able to produce a fantastic dataset to investigate potential mechanisms for promoter activation. From these large numbers of mutated sequences, they were able to generate mutual information with gene expression to identify key mutations relating to the activation of promoter island sequences.*

We thank Reviewer #2 for reading and providing a thorough review of our manuscript.

> *Strengths:*
>
> *The data generated from this paper is an important resource to address this question of promoter activation. Being able to link the activation of gene expression to mutational changes in previously nonactive promoter regions is exciting and allows the potential to investigate evolutionary processes relating to gene regulation in a statistically robust manner. Alongside this, the method of identifying key mutations using mutual information in this paper is well done and should be standard in future studies for identifying regions of interest.*

Thank you for your kind words.

> *Weaknesses:*
>
> *While the generation of the data is superb the focus only on these mutational hotspots removes a lot of the information available to the authors to generate robust conclusions. For instance.*
>
> *(1) The linear regression in S5 used to demonstrate that the number of mutational hotspots correlates with the likelihood of a mutation causing promoter activation is driven by three extreme points.*

A fair criticism. In response, we have chosen to remove the analysis of this trend from the manuscript entirely. (Additionally, Pnew and mutual information calculations both relied on the fluorescence scores of daughter sequences, so the finding was circular in its logic.)

> *(2) Many of the arguments also rely on the number of mutational hotspots being located near box motifs. The context-dependent likelihood of this occurring is not taken into account given that these sequences are inherently box motif rich. So, something like an enrichment test to identify how likely these hot spots are to form in or next to motifs.*

Another good point. To address it, we carried out a computational analysis where we randomly scrambled the nucleotides of each parent sequence while maintaining the coordinates for each mutual information “hotspot.” This scrambling results in significantly less overlap with hotspots and boxes. This analysis is now depicted in Figure 2C and described in lines 272-296.

> *(3) The link between changes in expression and mutations in surrounding motifs is assessed with two-sided Mann Whitney U tests. This method assumes that the sequence motifs are independent of one another, but the hotspots of interest occur either in 0, 3, 4, or 5s in sequences. There is therefore no sequence where these hotspots can be*

> *independent and the correlation causation argument for motif change on expression is weakened.*

This is a fair criticism and a limitation of the MWU test. To better support our reasoning, we engineered 27 reporter constructs harboring the specific mutations in the parents that we had predicted to change promoter activity. For each, we compared their fluorescence levels with their wild-type counterpart. The results from these experiments are in Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12 and are labeled with a ' (prime) symbol.

These experiments not only confirm the increases and decreases in fluorescence that our analysis had predicted. They also demonstrate, with the exception of two (out of 27) falsepositive discoveries, that background mutations do not confound our analysis. We mention these two exceptions (lines 364-367):

“In two of these hotspots, our validation experiments revealed no substantial difference in gene expression as a result of the hotspot mutation (Fig S8F' and Fig S8J'). In both of these false positives, new -10 boxes emerge in locations without an upstream -35 box.”

> *(4) The distance between -10 and -35 was mentioned briefly but not taken into account in the analysis.*

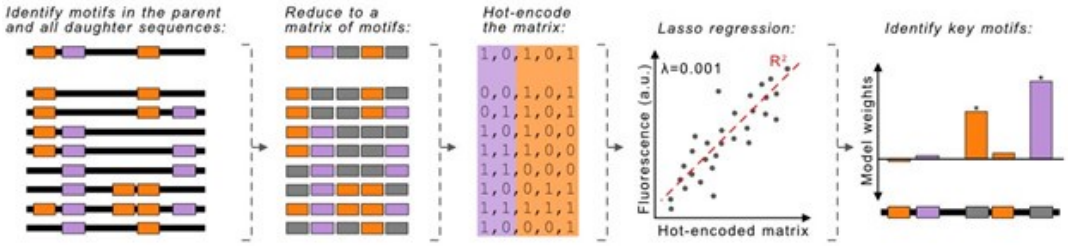
We have now included these spacer distances where appropriate. These changes are in Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12.

We note that in many cases, high-scoring PWM hits for the same motif can overlap (i.e. two -10 motifs or two -35 motifs overlap). Additionally, the proximity of a -35 and -10 box does not guarantee that the two boxes are interacting. Together, these two facts can result in an ambiguity of the spacer size between two boxes. To avoid any reporting bias, we thus often report spacer sizes as a range (see Figure panels 4F, S8D, S8F-L, S9A, S9H, S10A, and S10E). The smallest spacer we annotate is in Figure 4F with 10 bp, and the largest is in Figure S8D with 26 bp. More “extreme” distances are not annotated, and for the reader to decide if an interaction is present or not.

> *The authors propose mechanisms of promoter activation based on a few observations that are treated independently but occur concurrently. To address this using complementary approaches such as analysis focusing on identifying important motifs, using something like a glm lasso regression to identify significant motifs, and then combining with mutational hotspot information would be more robust.*

This is a great idea, and we pursued it as part of the revision. For each parent sequence, we mapped the locations of all -10 and -35 box motifs in the daughters, then reduced each sequence to a binary representation, either encoding or not encoding these motifs, also referred to as a “hot-encoded matrix.” We subsequently performed a Lasso regression between the hot-encoded matrices and the fluorescence scores of each daughter sequence. The regression then outputs “weights” to each of the motifs in the daughters. The larger a motif’s weight is, the more the motif influences promoter activity. The Author response image 1 describes our workflow.

Author response image 1.



We really wanted this analysis to work, but unfortunately, the computational model does not act robustly, even when testing multiple values for the hyperparameter lambda (λ), which accounts for differences in model biases vs variance.

The regression assigns strong weights almost exclusively to -10 boxes, and assigns weak to even negative weights to -35 boxes. While initially exciting, these weights do not consistently align with the results from the 27 constructs with individual mutations that we tested experimentally. This ultimately suggests that the regression is overfitting the data.

We do think a LASSO-regression approach can be applied to explore how individual motifs contribute to promoter activity. However, effectively implementing such a method would require a substantially more complex analysis. We respectfully believe that such an approach would distract from the current narrative, and would be more appropriate for a computational journal in a future study.

Because this analysis was inconclusive, we have not made it part of the revised manuscript. However, we hope that our 27 experimentally validated new constructs with individual mutations are sufficient to address the reviewer's concerns regarding independent verification of our computational predictions.

> *Other elements known to be involved in promoter activation including TGn or UP elements were not investigated or discussed.*

Thank you for highlighting this potentially important oversight. In response, we have performed two independent analyses to explore the role of TGn in promoter emergence in evolution. First, we computationally searched for -10 boxes with the bases TGn immediately upstream of them in the parent sequences, and found 18 of these "extended -10 boxes" in the parents (lines 143145):

"On average, each parent sequence contains ~5.32 -10 boxes and ~7.04 -35 boxes (Fig S1). 18 of these -10 boxes also include the TGn motif upstream of the hexamer."

However, only 20% of these boxes were found in parents with promoter activity (lines 182-185):

"We also note that 30% (15/50) of parents have the TGn motif upstream of a -10 box, but only 20% (3/15) of these parents have promoter activity (underlined with promoter activity: P4-RFP, P6-RFP, P8-RFP, P9-RFP, P10-RFP, P11GFP, P12-GFP, P17-GFP, P18-GFP, P18-RFP, P19-RFP, P22-RFP, P24-GFP, P25-GFP, P25-RFP). "

Second, we computationally searched through all of the daughter sequences to identify new -10 boxes with TGn immediately upstream. We found 114 -10 boxes with the bases TGn upstream. However, only 5 new -10 boxes (2 with TGn) were associated with increasing fluorescence (lines 338-345):

"On average, 39.5 and 39.4 new -10 and -35 boxes emerged at unique positions within the daughter sequences of each mutagenized parent (Fig 3A,B), with 1'562 and 1'576 new locations for -10 boxes and -35 boxes, respectively. ~22% (684/3'138) of these new boxes are spaced 15-20 bp away from their cognate box, and ~7.3% (114/1'562) of the new -10 boxes have the TGn motif upstream of them. However, only a mere five of the new -10 boxes and four of the new 35 boxes are significantly associated with increasing fluorescence by more than +0.5 a.u. (Fig 3C,D)."

In addition, we now study the role of UP elements. This analysis showed that the UP element plays a negligible role in promoter emergence within our dataset. It is discussed in a new subsection of the results (lines 591-608).

Collectively, these additional analyses suggest that the presence of TGn plus a -10 box is insufficient to create promoter activity, and that the UP element does not play a significant role in promoter emergence or evolution.

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *Like many papers in the last 5-10 years, this work brings a computational approach to the study of promoters and transcription, but unfortunately disregards or misrepresents much of the existing literature and makes unwarranted claims of novelty. My main concerns with the current paper are outlined below although the problems are deeply embedded.*

We thank Reviewer #3 for taking the time to review this manuscript. We have made extensive changes to address their concerns about our work.

> *Strengths:*
>
> *The data could be useful if interpreted properly, taking into account i) the role of translation ii) other promoter elements, and iii) the relevant literature.*
>
> *Weaknesses:*
>
> *(1) Incorrect assumptions and oversimplification of promoters.*
>
> *- There is a critical error on line 68 and Figure 1A. It is well established that the -35 element consensus is TTGACA but the authors state TTGAAA, which is also the sequence represented by the sequence logo shown and so presumably the PWM used. It is essential that the authors use the correct -35 motif/PWM/consensus. Likely, the authors have made this mistake because they have looked at DNA sequence logos generated from promoter alignments anchored by either the position of the -10 element or transcription start site (TSS), most likely the latter. The distance between the TSS and -10 varies. Fewer than half of E. coli promoters have the optimal 7 bp separation with distances of 8, 6, and 5 bp not being uncommon (PMID: 35241653). Furthermore, the distance between the -10 and -35 elements is also variable (16,17, and 18 bp spacings are all frequently found, PMID: 6310517). This means that alignments, used to generate sequence logos, have misaligned -35 hexamers. Consequently, the true consensus is not represented. If the alignment discrepancies are corrected, the true consensus emerges. This problem seems to permeate the whole study since this obviously incorrect consensus/motif has been used throughout to identify sequences that resemble -35 hexamers.*

We respectfully but strongly disagree that our analysis has misrepresented the true nature of -35 boxes. First, accounting for more A's at position 5 in the PWM is not going to lead to a

"critical error." This is because positions 4-6 of the motif barely have any information content (bits) compared to positions 1-3 (see Fig 1A). This assertion is not just based on our own PWM, but based on ample precedent in the literature. In PMID 14529615, TTG is present in 38% of all -35 boxes, but ACA only in 8%. In PMID 29388765, with the -10 instance TATAAT, the -35 instance TTGCAA yields stronger promoters compared to the -35 instance TTGACA (See their Figure 3B).

In PMID 29745856 (Figure 2), the most information content lies in positions 1-3, with the A and C at position 5 both nearly equally represented, as in our PWM. In PMID 33958766 (Figure 1) an experimentally-derived -35 box is even reduced to a "partial" -35 box which only includes positions 1 and 2, with consensus: TTnnnn.

In addition, we did not derive the PWMs as the reviewer describes. The PWMs we use are based on computational predictions that are in excellent agreement with experimental results. Specifically, the PWMs we use are from PMID 29728462, which acquired 145 -10 and -35 box sequences from the top 3.3% of computationally predicted boxes from Regulon DB. See PMID 14529615 for the computational pipeline that was used to derive the PWMs, which independently aligns the -10 and -35 boxes to create the consensus sequences. The -35 PWMs significantly and strongly correlates with an experimentally derived -35 box (see Supporting Information from Figure S4 of Belliveau et al., PNAS 2017. Pearson correlation coefficient = 0.89). Within the 145 -35 boxes, the exact consensus sequence (TTGACA) that Reviewer #3 is concerned about is present 6 times in our matrix, and has a PWM score above the significance threshold. In other words, TTGACA, is classified to be a -35 box in our dataset.

We now provide DNA sequences for each of the figures to improve accessibility and reproducibility. A reader can now use any PWM or method they wish to interpret the data.

> *- An uninformed person reading this paper would be led to believe that prokaryotic promoters have only two sequence elements: the -10 and -35 hexamers. This is because the authors completely ignore the role of the TG motif, UP element, and spacer region sequence. All of these can compensate for the lack of a strong -35 hexamer and it's known that appending such elements to a lone -10 sequence can create an active promoter (e.g. PMIDs 15118087, 21398630, 12907708, 16626282, 32297955). Very likely, some of the mutations, classified as not corresponding to a -10 or -35 element in Figure 2, target some of these other promoter motifs.*

Thank you for bringing this oversight to our attention. We have performed two independent analyses to explore the role of TGn in promoter emergence in evolution. First, we computationally searched for -10 boxes with the bases TGn immediately upstream of them in the parent sequences, and found 18 of these "extended -10 boxes" in the parents (lines 143145):

"On average, each parent sequence contains ~5.32 -10 boxes and ~7.04 -35 boxes (Fig S1). 18 of these -10 boxes also include the TGn motif upstream of the hexamer."

However, only 20% of these boxes were found in parents with promoter activity (lines 182-185):

"We also note that 30% (15/50) of parents have the TGn motif upstream of a -10 box, but only 20% (3/15) of these parents have promoter activity (underlined with promoter activity: P4-RFP, P6-RFP, P8-RFP, P9-RFP, P10-RFP, P11GFP, P12-GFP, P17-GFP, P18-GFP, P18-RFP, P19-RFP, P22-RFP, P24-GFP, P25-GFP, P25-RFP)."

Second, we computationally searched through all of the daughter sequences to identify new -10 boxes with TGn immediately upstream. We found 114 -10 boxes with the bases TGn

upstream. However, only 5 new -10 boxes (2 with TGn) were associated with increasing fluorescence (lines 338-345):

“On average, 39.5 and 39.4 new -10 and -35 boxes emerged at unique positions within the daughter sequences of each mutagenized parent (Fig 3A,B), with 1’562 and 1’576 new locations for -10 boxes and -35 boxes, respectively. ~22% (684/3’138) of these new boxes are spaced 15-20 bp away from their cognate box, and ~7.3% (114/1’562) of the new -10 boxes have the TGn motif upstream of them. However, only a mere five of the new -10 boxes and four of the new 35 boxes are significantly associated with increasing fluorescence by more than +0.5 a.u. (Fig 3C,D).”

In addition, we now study the role of UP elements. This analysis showed that the UP element plays a negligible role in promoter emergence within our dataset. It is discussed in a new subsection of the results (lines 591-608) and in the newly added Figure S13.

Collectively, these additional analyses suggest that the presence of TGn plus a -10 box is insufficient to create promoter activity, and that the UP element does not play a significant role in promoter emergence or evolution.

> *- The model in Figure 4C is highly unlikely. There is no evidence in the literature that RNAP can hang on with one "arm" in this way. In particular, structural work has shown that sequencespecific interactions with the -10 element can only occur after the DNA has been unwound (PMID: 22136875). Further, -10 elements alone, even if a perfect match to the consensus, are non-functional for transcription. This is because RNAP needs to be directed to the -10 by other promoter elements, or transcription factors. Only once correctly positioned, can RNAP stabilise DNA opening and make sequence-specific contacts with the -10 hexamer. This makes the notion that RNAP may interact with the -10 alone, using only domain 2 of sigma, extremely unlikely.*

This is a valid criticism, and we thank the reviewer for catching this problem. In response, we have removed the model and pertinent figures throughout the entire manuscript.

> *(2) Reinventing the language used to describe promoters and binding sites for regulators.*
>
> *- The authors needlessly complicate the narrative by using non-standard language. For example, On page 1 they define a motif as "a DNA sequence computationally predicted to be compatible with TF binding". They distinguish this from a binding site "because binding sites refer to a location where a TF binds the genome, rather than a DNA sequence". First, these definitions are needlessly complicated, why not just say "putative binding sites" and "known binding sites" respectively? Second, there is an obvious problem with the definitions; many "motifs" with also be "bindings sites". In fact, by the time the authors state their definitions, they have already fallen foul of this conflation; in the prior paragraph they stated: "controlled by DNA sequences that encode motifs for TFs to bind". The same issue reappears throughout the paper.*

We agree that this was needlessly complicated. We now just refer to every sequence we study as a motif. A -10 box is a motif, a -35 box is a motif, a putative H-NS binding site is an H-NS motif, etc. The word “binding site” no longer occurs in the manuscript.

> *- The authors also use the terms "regulatory" and non-regulatory" DNA. These terms are not defined by the authors and make little sense. For instance, I assume the authors would describe promoter islands lacking transcriptional activity (itself an incorrect assumption, see below)as non-regulatory. However, as horizontally acquired sections of AT-rich DNA these will all be bound by H-NS and subject to gene silencing, both*

> *promoters for mRNA synthesis and spurious promoters inside genes that create untranslated RNAs. Hence, regulation is occurring.*

Another fair point. We have thus changed the terminology throughout to “promoter” and “nonpromoter.”

> *- Line 63: "In prokaryotes, the primary regulatory sequences are called promoters". Promoters are not generally considered regulatory. Rather, it is adjacent or overlapping sites for TFs that are regulatory. There is a good discussion of the topic here (PMID: 32665585).*

We have rewritten this. The sentence now reads (lines 67-69):

“A canonical prokaryotic promoter recruits the RNA polymerase subunit σ70 to transcribe downstream sequences (Burgess et al., 1969; Huerta and Collado-Vides, 2003; Paget and Helmann, 2003; van Hijum et al., 2009).”

> *(3) The authors ignore the role of translation.*
>
> *- The authors' assay does not measure promoter activity alone, this can only be tested by measuring the amount of RNA produced. Rather, the assay used measures the combined outputs of transcription and translation. If the DNA fragments they have cloned contain promoters with no appropriately positioned Shine-Dalgarno sequence then the authors will not detect GFP or RFP production, even though the promoter could be making an RNA (likely to be prematurely terminated by Rho, due to a lack of translation). This is known for promoters in promoter islands (e.g. Figure 1 in PMID: 33958766).*

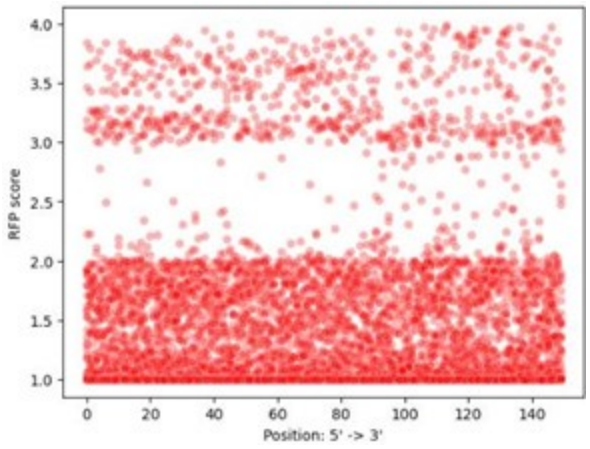
We agree that this is definitely a limitation of our study, which we had not discussed sufficiently. In response, we now discuss this limitation in a new section of the discussion (lines 680-686):

“Second, we measure protein expression through fluorescence as a readout for promoter activity. This readout combines transcription and translation. This means that we cannot differentiate between transcriptional and post-transcriptional regulation, including phenomena such as premature RNA termination (Song et al., 2022; Uptain and Chamberlin, 1997), post-transcriptional modifications (Mohanty and Kushner, 2006), and RNA-folding from riboswitch-like sequences (Mandal and Breaker, 2004).”

> *- In Figure S6 it appears that the is a strong bias for mutations resulting in RFP expression to be close to the 3' end of the fragment. Very likely, this occurs because this places the promoter closer to RFP and there are fewer opportunities for premature termination by Rho.*

The reviewer raises a very interesting possibility. To validate it, we have performed the following analysis. We took the RFP expression values from the 9’934 daughters with single mutations in all 25 parent sequences (P1-RFP, P2-RFP, … P25-RFP), and plotted the location of the single mutation (horizontal axis) against RFP expression (vertical axis) in Author response image 2.

Author response image 2.



The distribution is uniform across the sequences, showing that distance from the RBS is not likely the reason for this observation. Since this analysis was uninformative with respect to distance from the RBS, we chose not to include it in the manuscript.

> *(4) Ignoring or misrepresenting the literature.*
>
> *- As eluded to above, promoter islands are large sections of horizontally acquired, high ATcontent, DNA. It is well known that such sequences are i) packed with promoters driving the expression on RNAs that aren't translated ii) silenced, albeit incompletely, by H-NS and iii) targeted by Rho which terminates untranslated RNA synthesis (PMIDs: 24449106, 28067866, 18487194). None of this is taken into account anywhere in the paper and it is highly likely that most, if not all, of the DNA sequences the authors have used contain promoters generating untranslated RNAs.*

Thank you for pointing out that our original submission was incomplete in this regard. We address these concerns by new analyses, including some new experiments. First, Rhodependent termination is associated with the RUT motif, which is very rich in Cytosines (PMID: 30845912). Given that our sequences confer between 65%-78% of AT-content, canonical rhodependent termination is unlikely. However, we computationally searched for rho-dependent terminators using the available code from PMID: 30845912, but the algorithm did not identify any putative RUTs. Because this analysis was not informative, we did not include it in the paper.

We analyzed the role of H-NS on promoter emergence and evolution within our dataset using both experimental and computational approaches. These additional analyses are now shown in the newly-added Figure 5 and the newly-added Figure S12. We found that H-NS represses P22-GFP and P12-RFP and affects the bidirectionality of P20. More specifically, to analyze the effects of H-NS, we first compared the fluorescence levels of parent sequences in a Δhns background vs the wild-type (dh5α) background in Figure 5A. We found 6 candidate H-NS targets, with P22-GFP and P12-RFP exhibiting the largest changes in fluorescence (lines 496506):

“We plot the fluorescence changes in Fig 5A as distributions for the 50 parents, where positive and negative values correspond to an increase or decrease in fluorescence in the Δhns background, respectively. Based on the null hypothesis that the parents are not regulated by H-NS, we classified outliers in these distributions (1.5 × the interquartile range) as H-NS-target candidates. We refer to these outliers as “candidates” because the fluorescence

changes could also result from indirect trans-effects from the knockout (Mattioli et al., 2020; Metzger et al., 2016). This approach identified 6 candidates for H-NS targets (P2-GFP, P19-GFP, P20-GFP, P22-GFP, P12-RFP, and P20-RFP). For GFP, the largest change occurs in P22-GFP, increasing fluorescence ~1.6-fold in the mutant background (two-tailed t-test, p=1.16×10-8) (Fig 5B). For RFP, the largest change occurs in P12-RFP, increasing fluorescence ~0.5-fold in the mutant background (two-tailed t-test, p=4.33×10-10) (Fig 5B)."

We also observed that the Δhns background affected the bidirectionality of P20 (lines 507-511):

"We note that for template P20, which is a bidirectional promoter, GFP expression increases ~2.6-fold in the Δhns background (two-tailed t-test, p=1.59×10-6). Simultaneously, RFP expression decreases ~0.42-fold in the Δhns background (two-tailed t-test, p=4.77×10-4) (Fig S12A). These findings suggest that H-NS also modulates the directionality of P20's bidirectional promoter through either cis- or trans-effects."

We then searched for regions where losing H-NS motifs in hotspots significantly changed fluorescence. We identified 3 motifs in P12-RFP and P22-GFP (lines 522-528):

"For P22-GFP, a H-NS motif lies 77 bp upstream of the mapped promoter. Mutations which destroy this motif significantly increase fluorescence by +0.52 a.u. (two-tailed MWU test, q=1.07×10-3) (Fig 5E). For P12-RFP, one H-NS motif lies upstream of the mapped promoter's -35 box, and the other upstream of the mapped promoter's -10 box. Mutations that destroy these H-NS motifs significantly increase fluorescence by +0.53 and +0.51 a.u., respectively (two-tailed MWU test, q=3.28×10-40 and q=4.42 ×10-50) (Fig 5F,G). Based on these findings, we conclude that these motifs are bound by H-NS."

We are grateful for the suggestion to look at the role of H-NS in our dataset. Our analysis revealed a more plausible explanation to what we formerly referred to as a "Tandem Motif" in the original submission. Previously, we had shown that in P12-RFP, when a -35 box is created next to the promoter's -35 box, or a -10 box next to the promoter's -10 box, that expression decreases. These new -10 and -35 boxes, however, also overlap with the two H-NS motifs in P12-RFP. We tested these exact point mutations in reporter plasmids and in the Δhns background, and found that the Δhns background rescues this loss in expression (see Figure S12). This analysis is in the newly added subsection: "The binding of H-NS changes when new 10 and -35 boxes are gained" and can be found at lines 529-563. We summarize the findings in a final paragraph of the section (lines 556-563):

"To summarize, we present evidence that H-NS represses both P22-GFP and P12-RFP in cis. H-NS also modulates the bidirectionality of P20-GFP/RFP in cis or trans. In P22-GFP, the strongest H-NS motif lies upstream of the promoter. In P12-RFP, the strongest H-NS motifs lie upstream of the -10 and -35 boxes of the promoter. We note that there are 16 additional H-NS motifs surrounding the promoter in P12-RFP that may also regulate P12-RFP (Fig S12G). Mutations in two of these two H-NS motifs can create additional -10 and -35 boxes that appear to lower expression. However, the effects of these mutations are insignificant in the absence of H-NS, suggesting that these mutations actually modulate H-NS binding."

We also agree that the majority of these sequences are likely driving the expression of many untranslated RNAs (see Purtov et al., 2014). We thus now define a promoter more carefully as follows (lines 113-119):

"In this study, we define a promoter as a DNA sequence that drives the expression of a (fluorescent) protein whose expression level, measured by its fluorescence, is greater than a defined threshold. We use a threshold of 1.5 arbitrary units (a.u.) of fluorescence. This definition does not distinguish between transcription and translation. We chose it because protein expression is usually more important than RNA expression whenever natural

selection acts on gene expression, because it is the primary phenotype visible to natural selection (Jiang et al., 2023)."

We also state this as a limitation of our study in the Discussion (lines 680-686):

"Second, we measure protein expression through fluorescence as a readout for promoter activity. This readout combines transcription and translation. This means that we cannot differentiate between transcriptional and post-transcriptional regulation, including phenomena such as premature RNA termination (Song et al., 2022; Uptain and Chamberlin, 1997), post-transcriptional modifications (Mohanty and Kushner, 2006), and RNA-folding from riboswitch-like sequences (Mandal and Breaker, 2004)."

> *- The authors state that GC content does not correlate with the emergence of new promoters. It is known that GC content does correlate to the emergence of new promoters because promoters are themselves AT-rich DNA sequences (e.g. see Figure 1 of PMID: 32297955). There are two reasons the authors see no correlation in this work. First, the DNA sequences they have used are already very AT-rich (between 65 % and 78 % AT-content). Second, they have only examined a small range of different AT-content DNA (i.e. between 65 % and 78 %). The effect of AT-content on promoter emerge is most clearly seen between AT-content of between around 40 % and 60 %. Above that level, the strong positive correlation plateaus.*

We respectfully disagree that the reviewer's point is pertinent because what the reviewer is referring to is the likelihood that the sequence is a promoter, which indeed increases with AT content, but we are focused on the likelihood that a sequence becomes a promoter through DNA mutation. We note that if a DNA sequence is more AT-rich, then it is more likely to have -10 and -35 boxes, because their consensus sequences are also AT-rich. However, H-NS and other transcriptional repressors also bind to AT-rich sequences. This could also explain the saturation observed above 60% AT-content in PMID 32297955. Perhaps we can address this trend in future works.

> *- Once these authors better include and connect their results to the previous literature, they can also add some discussion of how previous papers in recent years may have also missed some of this important context.*

We apologize for this oversight. We have rewritten the Discussion section to include the following points below. Many of the newly added references come from the group of David Grainger, who works on H-NS repression, bidirectional promoters, promoter emergence, promoter motifs, and spurious transcription in *E. coli*. More specifically:

(1) The role of pervasive transcription and the likelihood of promoter emergence (lines 614-621):

"Instead, we present evidence that promoter emergence is best predicted by the level of background transcription each non-promoter parent produces, a phenomenon also referred to as "pervasive transcription" (Kapranov et al., 2007).

From an evolutionary perspective, this would suggest that sequences that produce such pervasive transcripts – including the promoter islands (Panyukov and Ozoline, 2013) and the antisense strand of existing promoters (Dornenburg et al., 2010; Warman et al., 2021), may have a proclivity for evolving de-novo promoters compared to other sequences (Kapranov et al., 2007; Wade and Grainger, 2014)."

(2) How our results contradict the findings from Bykov et al., 2020 (lines 622-640):

“A previous study randomly mutagenized the appY promoter island upstream of a GFP reporter, and isolated variants with increased and decreased GFP expression. The authors found that variants with higher GFP expression acquired mutations that 1) improve a -10 box to better match its consensus, and simultaneously 2) destroy other -10 and -35 boxes (Bykov et al., 2020). The authors concluded that additional -10 and -35 boxes repress expression driven by promoter islands. Our data challenge this conclusion in several ways.

First, we find that only ~13% of -10 and -35 boxes in promoter islands actually contribute to promoter activity. Extrapolating this percentage to the appY promoter island, ~87% (100% - 13%) of the motifs would not be contributing to its activity. Assuming the appY promoter island is not an outlier, this would insinuate that during random mutagenesis, these inert motifs might have accumulated mutations that do not change fluorescence. Indeed, Bykov et al. (Bykov et al., 2020) also found that a similar frequency of -10 and -35 boxes were destroyed in variants selected for lower GFP expression, which supports this argument. Second, we find no evidence that creating a -10 or -35 box lowers promoter activity in any of our 50 parent sequences. Third, we also find no evidence that destruction of a -10 or -35 box increases promoter activity without plausible alternative explanations, i.e. overlap of the destroyed box with a H-NS site, destruction of the promoter, or simultaneous creation of another motif as a result of the destruction. In sum, -10 and 35 boxes are not likely to repress promoter activity.”

(3) How other sequence features besides the -10 and -35 boxes may influence promoter emergence and activity (lines 661-671):

“These findings suggest that we are still underestimating the complexity of promoters. For instance, the -10 and -35 boxes, extended -10, and the UP-element may be one of many components underlying promoter architecture. Other components may include flanking sequences (Mitchell et al., 2003), which have been observed to play an important role in eukaryotic transcriptional regulation (Afek et al., 2014; Chiu et al., 2022; Farley et al., 2015; Gordân et al., 2013). Recent studies on E. coli promoters even characterize an AT-rich motif within the spacer sequence (Warman et al., 2020), and other studies use longer -10 and -35 box consensus sequences (Lagator et al., 2022). Another possibility is that there is much more transcriptional repression in the genome than anticipated (Singh et al., 2014). This would also coincide with the observed repression of H-NS in P22-GFP and P12-RFP, and accounts of H-NSrepression in the full promoter island sequences (Purtov et al., 2014).”

(4) The limits of our experimental methodology (lines 675-686):

“Additionally, future studies will be necessary to address the limitations of our own work. First, we use binary thresholding to determine i) the presence or absence of a motif, ii) whether a sequence has promoter activity or not, and iii) whether a part of a sequence is a hotspot or not. While chosen systematically, the thresholds we use for these decisions may cause us to miss subtle but important aspects of promoter evolution and emergence. Second, we measure protein expression through fluorescence as a readout for promoter activity. This readout combines transcription and translation. This means that we cannot differentiate between transcriptional and post-transcriptional regulation, including phenomena such as premature RNA termination (Song et al., 2022; Uptain and Chamberlin, 1997), posttranscriptional modifications (Mohanty and Kushner, 2006), and RNA-folding from riboswitch-like sequences (Mandal and Breaker, 2004) “

(5) An updated take-home message (lines 687-694):

“Overall, our study demonstrates that -10 and -35 boxes neither prevent existing promoters from driving expression, nor do they prevent new promoters from emerging by mutation. It shows how mutations can create new -10 and -35 boxes near or on top of preexisting ones to modulate expression. However, randomly creating a new -10 or -35 box will rarely create a new promoter, even if the new box is appropriately spaced upstream or downstream of a

cognate box. Ultimately our study demonstrates that promoter models need to be further scrutinized, and that using mutagenesis to create de-novo promoters can provide new insights into promoter regulatory logic."

> *(5) Lack of information about sequences used and mutations.*
>
> *- To properly assess the work any reader will need access to the sequences cloned at the start of the work, where known TSSs are within these sequences (ideally +/- H-NS, which will silence transcription in the chromosomal context but may not when the sequences are removed from their natural context and placed in a plasmid). Without this information, it is impossible to assess the validity of the authors' work.*

Thank you for raising this point. Please see Data S1 for the 25 template sequences (P1-P25) used in this study, and Data S2 for all of the daughter sequences.

For brevity, we have addressed the reviewer's request to look at the role of H-NS in their comment (4) "Ignoring or misrepresenting the literature."

We do not have information about the predicted transcription start sites (TSS) for the parent sequences because the program which identified them (Platprom) is no longer available. Regardless, having TSS coordinates would not validate or invalidate our findings, since we already know that the promoter islands produce short transcripts throughout their sequences, and we are primarily interested in promoters which can produce complete transcripts.

> *- The authors do not account for the possibility that DNA sequences in the plasmid, on either side of the cloned DNA fragment, could resemble promoter elements. If this is the case, then mutations in the cloned DNA will create promoters by "pairing up" with the plasmid sequences. There is insufficient information about the DNA sequences cloned, the mutations identified, or the plasmid, to determine if this is the case. It is possible that this also accounts for mutational hotspots described in the paper.*

We agree that these are important points. To address the criticism that we provided insufficient information, we now redesigned all our figures to provide this information. Specifically, the figures now include the DNA sequences, their PWM predictions, and the exact mutations that lead to promoter activity. The figures with these changes are Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12. We now also provide more details about pMR1 in a new section of the methods (lines 740-748):

"Plasmid MR1 (pMR1)

The plasmid MR1 (pMR1) is a variant of the plasmid RV2 (pRV2) in which the kan resistance gene has been swapped with the cm resistance gene (Guazzaroni and Silva-Rocha, 2014). Plasmid pMR1 encodes the BBa_J34801 ribosomal binding site (RBS, AAAGAGGAGAAA) 6 bp upstream of the start codon for GFP(LVA). The plasmid also encodes a putative RBS (AAGGGAGG) (Cazemier et al., 1999) 5 bp upstream of the start codon for mCherry on the opposite strand.

The plasmid additionally contains the low-to-medium copy number origin of replication p15A (Westmann et al., 2018).

A map of the plasmid is available on the Github repository: https://github.com/tfuqua95/promoter_islands "

The reviewer also makes a valid point about promoter elements of the plasmid itself. We addressed it with the following new analyses. First we re-examined each of the examples where new -10 and -35 boxes are gained or lost, to see if any of these hotspots occur on the

flanking ends of the parent sequences. We looked specifically at the ends because they could potentially interact with -10 and -35 box-like sequences on the plasmid to form a promoter.

Only one of these hotspots (out of 27) occurred at the end of the cloned sequences, and is thus a candidate for the phenomenon the reviewer hypothesized. This hotspot occurs in P9-GFP, where gaining a -10 box at the left flank increases expression (see Figure S8E-F'). There is indeed a -35 box 22-23 bp upstream of this -10 box on the plasmid, which could potentially affect promoter activity.

We tested the GFP expression of a construct harboring the point mutation which creates this -10 box on the left flank of P9-GFP. However, there was no significant difference in fluorescence between this construct and the wile-type P9-GFP (see Figure S8E-F'). Thus, this -35 box on pMR1 is not likely creating a new promoter.

> *(6) Overselling the conclusions.*
>
> *Line 420: The paper claims to have generated important new insights into promoters. At the same time, the main conclusion is that "Our study demonstrates that mutations to -10 and -35 boxes motifs are the primary paths to create new promoters and to modulate the activity of existing promoters". This isn't new or unexpected. People have been doing experiments showing this for decades. Of course, mutations that make or destroy promoter elements create and destroy promoters. How could it be any other way?*

In hindsight, we agree that the original conclusion was not very novel. Our new conclusion is that -10 and -35 boxes do not repress transcription, and that our current promoter models, even with the additional motifs like the UP-element and the extended -10, are insufficient to understand promoters (lines 687-694):

"Overall, our study demonstrates that -10 and -35 boxes neither prevent existing promoters from driving expression, nor do they prevent new promoters from emerging by mutation. It shows how mutations can create new -10 and -35 boxes near or on top of preexisting ones to modulate expression. However, randomly creating a new -10 or -35 box will rarely create a new promoter, even if the new box is appropriately spaced upstream or downstream of a cognate box. Ultimately our study demonstrates that promoter models need to be further scrutinized, and that using mutagenesis to create de-novo promoters can provide new insights into promoter regulatory logic."

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *I would like to start by thanking the authors for presenting an interesting and well-written article for review. This paper is a welcome addition to the field, addressing modern questions in the longstanding area of bacterial gene regulation. It is both enlightening and inspiring. While I do have suggestions, I hope these are not perceived as a lack of optimism for the work.*

Thank you for your kind words and suggestions, and for providing an astute and constructive review. We feel that manuscript has greatly improved with your suggested changes.

> *ABSTRACT:*
>
> *Line 11: The sentence, "It is possible that these motifs influence..." Could be rewritten to be clearer as it is the most important point of the manuscript. It is not obvious that*

> *you're talking about how the local landscape of motifs affects the probability of promoters evolving/devolving in this location.*

We have changed the sentence to read, “Here, we ask whether the presence of such motifs in different genetic sequences influences promoter evolution and emergence.”

> *INTRODUCTION:*
>
> *Line 68: Is the -35 consensus motif not TTGACA? Here it is listed as TTGAAA.*

Corrected from TTGAAA to TTGACA

> *RESULTS:*
>
> *Line 92-94. In finding that the. The main takeaway from this work is that different sequences have different likelihoods of mutations creating promoters and so I believe this claim could be explored deeper with more quantitative information. Could the authors supplement this claim by including? Could you look at whether there is a correlation between the baseline expression of a parent sequence and Pnew? I expect even the inactive sequences to have some variability in measured expression.*

Thank you for this great idea. We followed up on it by plotting the baseline parent sequence fluorescence scores against Pnew. You are indeed correct, i.e., Pnew increases with baseline expression following a sigmoid function, and is now shown in Figure 1D. To report our new observations, we have added the following section to the Results (lines 219-232):

“Although mutating each of the 40 non-promoter parent sequences could create promoter activity, the likelihood Pnew that a mutant has promoter activity, varies dramatically among parents. For each non-promoter parent, Fig 1D shows the percentage of active daughter sequences. The median Pnew is 0.046 (std. ± 0.078), meaning that ~4.6% of all mutants have promoter activity. The lowest Pnew is 0.002 (P25-GFP) and the highest 0.41 (P8-RFP), a 205-fold difference.

We hypothesized that these large differences in Pnew could be explained by minute differences in the fluorescence scores of each parent, particularly if its score was below 1.5 a.u. Plotting the fluorescence scores of each parent (N=50) and their respective Pnew values as a scatterplot (Fig 1E), we can fit these values to a sigmoid curve (see methods). This finding helps to explain why P8-RFP has a high Pnew (0.41) and P25-GFP a low Pnew (0.002), as their fluorescence scores are 1.380 and 1.009 a.u., respectively. The fact that the inflection point of the fitted curve is at 1.51 a.u. further justifies our use of 1.5 a.u. as a cutoff for promoter and non-promoter activity.”

> *Another potentially interesting analysis would be to see if k-mer content is correlated with Pnew. That is, determine the abundance of all hexamers in the sequence and see if Pnew is correlated with the number of hexamers present that is one nucleotide distance away from the consensus motifs (such as TcGACA or TAcAAT).*

We performed the suggested analysis by searching for k-mers that correlate with Pnew and found that no k-mer significantly correlates with Pnew (lines 240-248):

“We then asked whether any k-mers ranging from 1-6 bp correlated with the non-promoter Pnew values (5,460 possible k-mers). 718 of these 1-6 bp k-mers are present 3 or more times in at least one non-promoter parent. We calculated a linear regression between the frequency of these 718 k-mers and each Pnew value, and adjusted the p-values to respective q-values (Benjamini-Hochberg correction, FDR=0.05). This analysis revealed six k-mers: CTTC, GTTG,

ACTTC, GTTGA, AACTTC, TAACTT which correlate with Pnew. However, these correlations are heavily influenced by an outlying Pnew value of 0.41 (P8-RFP) (Fig S5C-H), and upon removing P8-RFP from the analysis, no k-mer significantly correlates with Pnew (data not shown)"

> *Line 152-157: How did you define the thresholds for 'active' or 'inactive'? It is not clear in the methods how this distinction was made.*

We have more clearly defined these thresholds in the text. A sequence with promoter activity has a fluorescence score greater than 1.5 a.u. (lines 168-172):

"We declared a daughter sequence to have promoter activity or to be a promoter if its score was greater than or equal to 1.5 a.u., as this score lies at the boundary between no fluorescence and weak fluorescence based on the sort-seq bins (methods). Otherwise, we refer to a daughter sequence as having no promoter activity or being a non-promoter."

> *Lines: 152-157: In trying to find the parent expression levels, no figure was available showing the distribution of parent expression levels. Furthermore, In looking at Data S2 & filtering out for sequences with distance 0 from the parent, I found the most active sequences did not match up with the sequences described as active in this section (e.g. p19 and p20 have a higher topstrand mean over P22, yet are not listed as active top strand sequences).*

We really appreciate you taking the time to examine the supplemental data. We previously listed the parents that had *only* GFP activity but no RFP activity (P22), and *only* RFP activity but no GFP activity (P6, P12, P13, P18, P21). We then said that P19 and P20 were bidirectional promoters, because they showed *both* GFP and RFP activity. In hindsight, we realize that our wording was confusing. We thus rewrote the affected paragraph, such that the bidirectional promoters are now in both lists of GFP/RFP active parents. We also now make the distinction between "templates" which comprise our 25 promoter island fragments, and "parents", where we treat both strands separately (50 parents total). The paragraph in question now reads (lines 173-187):

"Because some sequences in our library are unmutated parent sequences, we determined that 10/50 of the parent sequences already encode promoter activity before mutagenesis. Specifically, three parents drove expression on the top strand (P19-GFP, P20-GFP, P22-GFP), and five did on the bottom strand (P6-RFP, P12-RFP, P13-RFP, P18-RFP, P19-RFP, P20-RFP, P21-RFP). Two parents harbor bidirectional promoters (P19 and P20). The remaining 40 parent sequences are non-promoters, with an average fluorescence score of 1.39 a.u. We note that some of these parents have a fluorescence score higher than 1.39 a.u., but less than 1.50 a.u. such as P8-RFP (1.38 a.u.), P16-RFP (1.39 a.u.), P9-GFP (1.49 a.u.), and P1-GFP (1.47 a.u.). Whether these are truly "promoters" or not, is based solely on our threshold value of 1.5 a.u. We also note that 30% (15/50) of parents have the TGn motif upstream of a -10 box, but only 20% (3/15) of these parents have promoter activity (underlined with promoter activity: P4-RFP, P6-RFP, P8-RFP, P9RFP, P10-RFP, P11-GFP, P12-GFP, P17-GFP, P18-GFP, P18-RFP, P19-RFP, P22-RFP, P24-GFP, P25-GFP, P25RFP). See Fig S4 for fluorescence score distributions for each parent and its daughters, and Data S2 for all daughter sequence fluorescence scores."

> *Please include a supplementary figure showing the different parent expression levels (GFP mean +/- sd). Also, please explain the discrepancy in the 'active sequences' compared to Data S2 or correct my misunderstanding.*

We have added this plot to Figure S4B. The discrepancy arose because we listed the parents that had *only* GFP activity but no RFP activity (P22), and *only* RFP activity but no GFP activity

(P6, P12, P13, P18, P21). We then said that P19 and P20 were bidirectional promoters, because they showed *both* GFP and RFP activity. previous response regarding the ambiguity.

> *Line 182: I do not see 'Fuqua and Wagner 2023' in the references (though I am familiar with the preprint).*

We have added Fuqua and Wagner, BiorXiv 2023 to the references.

> *Lines 197 - 200: The distribution of hotspot locations should be compared to the distribution of mutations in the library. e.g. It is not notable that 17% of mutations are in -10 motifs if 17% of all mutations are in -10 motifs.*

Thank you for raising this point. To address it, we carried out a computational analysis where we randomly scrambled the nucleotides of each parent sequence while maintaining the coordinates for each mutual information “hotspot.” This scrambling results in significantly less overlap with hotspots and boxes. This analysis is now depicted in Figure 2C and written in lines 272-296.

> *Lines 253-264: Examples 3B, 3D, and 3F should indicate the spacing between the new and existing motifs. Are these close to the 15-19 bp spacer lengths preferred by sigma70?*

Point well taken. We now annotate the spacing of motifs in Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, and S11. We note that in many cases, high-scoring PWM hits for the same motif can overlap (i.e. two -10 motifs or two -35 motifs overlap). Additionally, the proximity of a 35 and -10 box does not guarantee that the two boxes are interacting. Together, these two facts can result in an ambiguity of the spacer size between two boxes. To avoid any reporting bias, we thus often report spacer sizes as a range (see Figure panels 4F, S8D, S8F-L, S9A, S9H, S10A, and S10E). The smallest spacer we annotate is in Figure 4F with 10 bp, and the largest is in Figure S8D with 26 bp. Any more “extreme” distances are not annotated, and for the reader to decide if an interaction is present or not.

> *Line 255: While fun, I am concerned about the 'Shiko' analogy. My understanding is the prevailing theory is that -35 recognition occurs before -10 recognition (https://doi.org/10.1073/pnas.94.17.9022, 10.1101/sqb.1998.63.141). Given this, the 'Shiko -35' concept in 3H is a bit awkward as it suggests that sigma70 stops at -10 motifs before planting down on the -35. Considering the cited paper is still in the preprint stages (and did not observe these Shiko -35 emergences), I am concerned about how this particular example will be received by the community. Perhaps more care could be done to verify that this example is consistent with generally accepted mechanisms of promoter recognition or a short clarification could be added to clarify the extent of the analogy.*

Thank you for raising this point. We decided to remove the Shiko analogy, because several readers assumed that it relates to the physical binding of RNA polymerase, rather than being an evolutionary mechanism of mutations forming complementary motifs in a stepwise manner.

> *Lines 323-326: It would be helpful to describe a more systematic approach to defining emergence events into different categories. A clear definition of each category in the methods or main text would help others consistently refer to these concepts in the future. This could be helped by showing the actual parent vs daughter sequences as a supplementary figure to figures 4B, 4D, & 4G.*

We agree this could have been more clearly communicated. We have addressed this by 1) simplifying the nomenclatures of these categories and 2) clearly defining these categories,

and 3) showing the actual parent vs daughter sequences in Figure 4, and Supplemental Figures S9, S10, S11, and S12. More specifically:

(1) Simplifying the nomenclature. We highlight events where gaining new -10 and -35 boxes can modify the promoter activity of parent sequences with promoter activity. This occurs when a new -10 or -35 box appears that partially overlaps with the -10 or -35 box of the actual promoter. Thus, we rename two terms: hetero-gain and homo-gain, shown in Figure 4B:

(2) We clearly define these categories (lines 430-435):

“We found that these mutations frequently create new boxes overlapping those we had identified as part of a promoter (Fig S9). This occurs when mutations create a -10 box overlapping a -10 box, a -35 box overlapping a 35 box, a -10 box overlapping a -35 box, or a -35 box overlapping a -10 box. We call the resulting event a “homogain” when the new box is of the same type as the one it overlaps, and otherwise a “hetero-gain”. In either case, the creation of the new box does not always destroy the original box.”

In the original manuscript, there was an additional third category, where gaining a -35 box upstream of the promoter’s -35 box, and gaining a -10 box upstream of the promoter’s -10 box decreased expression. We referred to this as a “tandem motif” and it can be found in Figure S12C,D. However, in response to comment “(4) Ignoring or misrepresenting the literature” from Reviewer #3, we carried out an analysis of the binding of H-NS (see Figure 5 and Figure S12). This analysis revealed that this “tandem motif” phenomenon was actually the result of changing the affinity of H-NS to these regions. Thus, the “tandem motif” is probably spurious.

> *DISCUSSION:*
>
> *Line 378-379: Since hotspots are essentially areas where promoters appear, wouldn't it be obvious that having more hotspots (i.e. areas where more promoters appear) would equate to a higher probability of new promoters? It would be helpful to clarify why this isn't obvious. This could be resolved by adding more complexity to the statement, such as showing that the level of mutual information found in a hotspot or across all hotspots in a sequence is correlated with Pnew.*

A fair criticism. In response, we have chosen to remove the analysis of this trend from the manuscript entirely. (Additionally, Pnew and mutual information calculations both relied on the fluorescence scores of daughter sequences, so the finding was circular in its logic.)

> *Line 394-396: This comparison of findings to Bykov et al should include a bit more justification for the proposed mechanism and how it specifically was observed in this paper. What did they observe and how do these findings relate?*

We gladly followed this suggestion, and added the following two paragraphs to the discussion (lines 622-640).

“A previous study randomly mutagenized the appY promoter island upstream of a GFP reporter, and isolated variants with increased and decreased GFP expression. The authors found that variants with higher GFP expression acquired mutations that 1) improve a -10 box to better match its consensus, and simultaneously 2) destroy other -10 and -35 boxes (Bykov et al., 2020). The authors concluded that additional -10 and -35 boxes repress expression driven by promoter islands. Our data challenge this conclusion in several ways.

First, we find that only ~13% of -10 and -35 boxes in promoter islands actually contribute to promoter activity. Extrapolating this percentage to the appY promoter island, ~87% (100% - 13%) of the motifs would not be contributing to its activity. Assuming the appY promoter island is not an outlier, this would insinuate that during random mutagenesis, these inert

motifs might have accumulated mutations that do not change fluorescence. Indeed, Bykov et al. (Bykov et al., 2020) also found that a similar frequency of -10 and -35 boxes were destroyed in variants selected for lower GFP expression, which supports this argument. Second, we find no evidence that creating a -10 or -35 box lowers promoter activity in any of our 50 parent sequences. Third, we also find no evidence that destruction of a -10 or -35 box increases promoter activity without plausible alternative explanations, i.e. overlap of the destroyed box with a H-NS site, destruction of the promoter, or simultaneous creation of another motif as a result of the destruction. In sum, -10 and 35 boxes are not likely to repress promoter activity. “

> *METHODS:*
>
> *Line 500: Could you provide more details on PMR1 (e.g. size, copy number, RBS strength) or a reference? I could not find this easily.*

Thank you for pointing out this oversight. In response, we have added the following subsection to the methods (lines 740-748):

“Plasmid MR1 (pMR1)

The plasmid MR1 (pMR1) is a variant of the plasmid RV2 (pRV2) in which the kan resistance gene has been swapped with the cm resistance gene (Guazzaroni and Silva-Rocha, 2014). Plasmid pMR1 encodes the BBa_J34801 ribosomal binding site (RBS, AAAGAGGAGAAA) 6 bp upstream of the start codon for GFP(LVA). The plasmid also encodes a putative RBS (AAGGGAGG) (Cazemier et al., 1999) 5 bp upstream of the start codon for mCherry on the opposite strand.

The plasmid additionally contains the low-to-medium copy number origin of replication p15A (Westmann et al., 2018).

A map of the plasmid is available on the Github repository: https://github.com/tfuqua95/promoter_islands.”

> *Line 581: What was the sequencing instrument &/or depth?*

We now report this information as follows (Methods, lines 918-922):

“Illumina sequencing

The amplicon pool was sequenced by Eurofins Genomics (Eurofins GmbH, Germany) using a NovaSeq 6000 (Illumina, USA) sequencer, with an S4 flow cell, and a PE150 (Paired-end 150 bp) run. In total, 282’843’000 reads and 84’852’900’000 bases were sequenced. Raw sequencing reads can be found here: https://www.ncbi.nlm.nih.gov/bioproject/1071572.”

> *SUPPLEMENT:*
>
> *Supplementary Figure 2: Why does the GFP control produce a bimodal distribution?*

The GFP+ culture was inoculated directly from a glycerol stock. The bimodal distribution probably results from a subset of the bacteria having lost the GFP-coding insert, because the left-most peak coincides with the negative control.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *This paper would benefit from a clear definition of what constitutes an active promoter as this is only mentioned as justification for the use of arbitrary values for fluorescence.*

Good point. To clarify, we now include this new paragraph in the introduction (lines 112-119):

"In this study, we define a promoter as a DNA sequence that drives the expression of a (fluorescent) protein whose expression level, measured by its fluorescence, is greater than a defined threshold. We use a threshold of 1.5 arbitrary units (a.u.) of fluorescence. This definition does not distinguish between transcription and translation. We chose it because protein expression is usually more important than RNA expression whenever natural selection acts on gene expression, because it is the primary phenotype visible to natural selection (Jiang et al., 2023)."

> *There needs to be a clear distinction in the use of the word sequences as often interchange sequences when meaning the 25 parent sequences and then the 50 possible sequences directions the promoter can act. It is confusing going from one to the other.*

We agree that this distinction is important. To make it clearer, we now introduce an additional term (lines 119-130). Our experiments start from 25 promoter island fragments (P1-P25), which we now call template sequences. Each template sequence comprises both DNA strands. The parent sequences are the top and bottom strands of each template sequence. Therefore, there are now 50 parent sequences (P1-GFP, P1-RFP, P2-GFP…, P25-RFP). By treating each strand as its own sequence, we no longer have to refer to the strand, avoiding the earlier confusion.

> *The description of the hotspots is often unclear and trying to determine if 3 out of 9 hotspots come from one parent sequence or multiple is not possible. A table denoting this information would be most helpful.*

We agree, and now provide this information in Data S3.

> *Finally, the description of the proposed mechanism of promoter activation via mutation of motifs should not be in the results but in the discussion, as it has insufficient evidence and would require further experimental validation.*

We remedied this problem by providing experimental validation of the proposed mechanisms. Specifically, we created the precise mutations that caused a loss or gain of a -10 or a -35 box, and measured the level of gene expression they drive with a plate reader. Because we chose to provide this experimental validation, we opted to leave the mechanisms of promoter activation in the results section.

> *The (Fuqua and Wagner 20023) paper is not in the references.*

We have added Fuqua and Wagner, BiorXiv 2023 to the references.

> *I enjoyed the paper and wish the authors the best for their future work.*

Thank you for taking the time to review our manuscript!

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *The paper has major flaws. For example:*
>
> *The data need to be analysed with correct promoter sequence element sequences (TTGACA for the -35 element).*

The discrepancy lies in the frequency of A's vs C's at position #5 of the PWM. Our PWM was built with more A's than C's at this position, but also includes C's in this position. However, we respectfully disagree that using a different -35 box PWM is going to change the outcomes of

our study. First, positions 4-6 of the PWM barely have any information content (bits) compared to positions 1-3 (see Fig 1A). This assertion is not just based on our own PWM, but based on ample precedent in the literature. In PMID 14529615, TTG is present in 38% of all -35 boxes, but ACA only 8%. In PMID 29388765, with the -10 instance TATAAT, the -35 instance TTGCAA yields stronger promoters compared to the -35 instance TTGACA (See their Figure 3B). In PMID 29745856 (Figure 2), the most information content lies in positions 1-3, with the A and C at position 5 both nearly equally represented, as in our PWM. In PMID 33958766 (Figure 1) an experimentally-derived -35 box is even reduced to a "partial" -35 box which only includes positions 1 and 2, with consensus: TTnnnn. Additionally, the -35 box PWM that we used significantly and strongly correlates with an experimentally derived -35 box (see Supporting Information from Figure S4 of Belliveau et al., PNAS 2017. Pearson correlation coefficient = 0.89). We now provide DNA sequences for each of the figures to improve accessibility and reproducibility. A reader can now use any PWM or method they wish to interpret the data.

> *The data need to be analysed taking into account the role of other promoter elements and sequences for translation.*

Point well taken.

Thank you for bringing this oversight to our attention. We have performed two independent analyses to explore the role of TGn in promoter emergence in evolution. First, we computationally searched for -10 boxes with the bases TGn immediately upstream of them in the parent sequences, and found 18 of these "extended -10 boxes" in the parents (lines 143145):

"On average, each parent sequence contains ~5.32 -10 boxes and ~7.04 -35 boxes (Fig S1). 18 of these -10 boxes also include the TGn motif upstream of the hexamer."

However, only 20% of these boxes were found in parents with promoter activity (lines 182-185):

"We also note that 30% (15/50) of parents have the TGn motif upstream of a -10 box, but only 20% (3/15) of these parents have promoter activity (underlined with promoter activity: P4-RFP, P6-RFP, P8-RFP, P9-RFP, P10-RFP, P11GFP, P12-GFP, P17-GFP, P18-GFP, P18-RFP, P19-RFP, P22-RFP, P24-GFP, P25-GFP, P25-RFP)."

Second, we computationally searched through all of the daughter sequences to identify new -10 boxes with TGn immediately upstream. We found 114 -10 boxes with the bases TGn upstream. However, only 5 new -10 boxes (2 with TGn) were associated with increasing fluorescence (lines 338-345):

"Mutations indeed created many new -10 and -35 boxes in our daughter sequences. On average, 39.5 and 39.4 new 10 and -35 boxes emerged at unique positions within the daughter sequences of each mutagenized parent (Fig 3A,B), with 1'562 and 1'576 new locations for -10 boxes and -35 boxes, respectively. ~22% (684/3'138) of these new boxes are spaced 15-20 bp away from their cognate box, and ~7.3% (114/1'562) of the new -10 boxes have the TGn motif upstream of them. However, only a mere five of the new -10 boxes and four of the new -35 boxes are significantly associated with increasing fluorescence by more than +0.5 a.u. (Fig 3C,D)."

In addition, we now study the role of UP elements. This analysis showed that the UP element plays a negligible role in promoter emergence within our dataset. It is discussed in a new subsection of the results (lines 591-608).

"The UP-element does not strongly influence promoter activity in our dataset.

The UP element is an additional AT-rich promoter motif that can lie stream of a -35 box in a promoter sequence (Estrem et al., 1998; Ross et al., 1993). We asked whether the creation of UP-elements also creates or modulates promoter activity in our dataset. To this end, we first identified a previously characterized position-weight matrix for the UP element (NNAAAWWTWTTTTNNWAAASYM, PWM threshold score = 19.2 bits) (Estrem et al., 1998) (Fig S13A). We then computationally searched for UP-element-specific hotspots within the parent sequences, i.e., locations in which mutations that gain or lose UP-elements lead to significant fluorescence increases (Mann-Whitney U-test, Fig S7 and methods. See Data S8 for the coordinates, fluorescence changes, and significance). The analysis did not identify any UP elements whose mutation significantly changes fluorescence.

We then repeated the analysis with a less stringent PWM threshold of 4.8 bits (1/4th of the PWM threshold score). This time, we identified 74 "UP-like" elements that are created or destroyed at unique positions within the parents. 23 of these motifs significantly change fluorescence when created or destroyed. However, even with this liberal threshold, none of these UP-like elements increase fluorescence by more than 0.5 a.u. when gained, or decrease fluorescence by more than 0.5 a.u. when lost (Fig S13B). This finding ultimately suggests that the UP element plays a negligible role in promoter emergence within our dataset."

Collectively, these additional analyses suggest that the presence of TGn plus a -10 box is insufficient to create promoter activity, and that the UP element does not play a significant role in promoter emergence or evolution.

> *The full sequences used need to be provided and mutations resulting in new promoters need to be shown.*

To Figures 3, 4, 5, and Supplemental Figures S8, S9, S10, S11, and S12, we have added the sequences which created or the destroyed the promoters, and their PWM scores.

> *The paper needs to be rewritten to take into account the relevant literature on i) promoter islands (i.e. sections of horizontally acquired AT-rich DNA) ii) generation and loss of promoters by mutation.*

We have rewritten the introduction. The majority of these points are now addressed in the following two new paragraphs (lines 92-112):

"Recent work shows that mutations can help new promoters to emerge from promoter motifs or from sequences adjacent to such motifs (Bykov et al., 2020; Fuqua and Wagner, 2023; Yona et al., 2018). However, encoding -10 and -35 boxes is insufficient to drive complete transcription of a gene coding sequence. For instance, the E. coli genome contains clusters of -10 and -35 boxes that are bound by RNA polymerase and produce short oligonucleotide fragments, but rarely create complete transcripts. Such clusters are called promoter islands, and are strongly associated with horizontally-transferred DNA (Bykov et al., 2020; Panyukov and Ozoline, 2013; Purtov et al., 2014; Shavkunov et al., 2009).

There are two proposed explanations for why promoter islands do not create full transcripts. First, the TF H-NS may repress promoter activity in promoter islands. This is because in a Δhns background, transcript levels from the promoter islands increases (Purtov et al., 2014). However, mutagenizing a specific promoter island (appY) until it transcribes a GFP reporter, reveals that in-vitro H-NS binding does not significantly change when GFP levels increase (Bykov et al., 2020). Thus, it is not clear whether H-NS actually represses the complete transcription of these sequences. The second proposed explanation is that excessive promoter motifs silence transcription. The aforementioned study found that promoter activity increases when mutations improve a -10 box to better match its consensus (TAAAAAT→TATACT), while simultaneously destroying surrounding -10 and -35 boxes (Bykov

et al., 2020). However, we note that if these surrounding motifs never contributed to GFP fluorescence to begin with, then mutations could also simply have accumulated in them during random mutagenesis without affecting promoter activity."

In closing, we would like to thank all three reviewers again for your time to engage with this manuscript.

Summary of specific changes that we have made to each section of the manuscript

• Abstract

- We updated the abstract to include the finding that more than 1'500 new -10s and 35s are created in our dataset, but only ~0.3% of them actually create de-novo promoter activity.

- We no longer highlight the conclusion that the majority of promoters emerge and evolve from -10 and -35 boxes.

• Introduction

- We have added more background information about the UP-element and the TGn motif.

- We better describe the promoter islands and the results identified by Bykov et al., 2020.

• Results: Promoter island sequences are enriched with motifs for -10 and -35 boxes.

- We clarify how the -10 and -35 PWMs we use were derived.

- We refer to the 25 promoter island fragments as "Template sequences" (P1-P25). The "parent sequences" now correspond to the top and bottom strands of each template (N=50, P1-GFP, P1-RFP, P2-GFP, …, P25-RFP).

- We elaborate that ~7% of the -10 boxes in the template sequences have the TGn motif.

- In the previous version of the manuscript, if there were overlapping -10 boxes or overlapping -35 box, we counted these to be a single -10 box or a single -35 box, respectively. In the new version of the manuscript, we now treat each motif as an independent box. Because of this, the number of -10 and -35 boxes per parent have slightly increased.

•Results: Non-promoters vary widely in their potential to become promoters.

- We make a clear distinction between promoters and non-promoters, and define the parent sequences.

- We note that only 20% of parents with an "extended -10 box" have promoter activity.

• Results: Promoter emergence correlates with minute differences in background promoter levels.

- We added an analysis where we compare Pnew to the parent fluorescence levels, even if they are below 1.5 a.u. We find that the distribution of Pnew matches a sigmoid function.

• Results: Promoter emergence does not correlate with simple sequence features

- We added an analysis comparing k-mer counts to Pnew.

- We updated the way we count -10 and -35 boxes, and recalculated the correlation with Pnew. The P and R2 values have changed, but Pnew still does not significantly correlate with -10 or -35 box counts.

• Results: Promoters emerge and evolve only from specific subsets of -10 and -35 boxes

- We have added an analysis where we computationally scramble the wild-type parent sequences while maintaining the coordinates of the mutual information hotspots. This reveals that the overlap with -10 and -35 motifs is not a coincidence of dense promoter motif encoding.

We found a computational error in our analysis and updated the percent overlap between -10 boxes and -35 boxes with mutual information hotspots. The results are similar. o 14% of -10 boxes overlap with hotspots with our new way of defining -10 and -35 boxes.

• Results: New -10 and -35 boxes readily emerge, but rarely lead to de-novo promoter activity

- We quantify how often a new -10 and -35 box is created at a unique position within our collection of promoter fragments, and how often this results in a -10 and -35 box being appropriately spaced, and how often this actually leads to de-novo promoter activity. o We quantify how often a TGn sequence lies upstream of a new -10 box.

• Results: Promoters can emerge when mutations create motifs but not by destroying them.

- For each example, we added the DNA sequences of the wild-type region of interest and the mutant region of interest that results in the gain of promoter activity, and their respective PWM scores.

- We created constructs to validate each example by testing their fluorescence on a plate reader.

- We removed the P1-GFP example from the main figure, as it was a false-positive in the dataset. It is now in Fig S8.

- We removed the Shiko Emergence metaphor because it could be confused with a binding mechanism for RNA polymerase.

• Results – Gaining new motifs over existing motifs increases and decreases promoter activity.

- We removed the “Tandem motif” because it is more likely caused by H-NS binding.

- We renamed the mechanisms to be “hetero-gain” and “homo-gain” for simplicity, and clearly define how we classified each sequence into each category.

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from constructs harboring the predicted point mutations.

• Results – Histone-like nucleoid-structuring protein (H-NS) represses P12-RFP and P22-GFP.

- This is a new analysis, which explores the role of the TF H-NS in repressing the parent sequences.

- We identified putative H-NS motifs in P12-RFP and P22-GFP.

- We show experimentally that in a H-NS null background, a bidirectional promoter (P20) becomes unidirectional, even though P20 does not contain an obvious H-NS motif.

- In the original version of the manuscript, we describe a phenomenon where gaining a -35 box upstream of a promoter’s -35 box, or a -10 box upstream of a promoter’s -10 box significantly decreases expression. We called this phenomenon a “tandem motif.” However, in the newest version of the manuscript, we find that these fluorescence decreases are rescued in a H-NS null background, suggesting the finding was actually due to H-NS binding modulation and not -10 and -35 boxes.

• Results – The UP-element does not strongly influence promoter activity in our dataset.

We used a PWM for the UP element to see if gaining or losing UP motifs was significantly correlated with increasing or decreasing expression. Even with a liberal PWM threshold, the analysis did not find any UP elements.

• Discussion

- We rewrote the discussion to account for the new analyses and the results on H-NS, the UP-element, and the extended -10.

- We better explain how our results clash with the results from the Bykov paper.

- We fit our results into the context of David Grainger's papers.

• Methods

- Added an explanation about pMR1.

- Added methods describing how we created the point mutation constructs.

- Added the methods for the plate reader.

- Added the methods for Illumina sequencing.

- Added the methods for the sigmoid curve-fitting.

• Figure 1

- Panel E compares how Pnew (the probability of a daughter sequence having a fluorescence score greater than 1.5 a.u.) associates with the fluorescence scores of each parent sequence.

- Panel F was originally in Figure S5. In the originally submitted version of the manuscript, if there were overlapping -10s or overlapping -35s, we counted these to be a single -10 or a single -35, respectively. In the new version of the manuscript, we now treat each motif as an independent box. Because of this, the r2 and p values have changed, but the conclusions have not (Pnew still does not significantly correlate with -10 or -35 box counts).

• Figure 2

- Panel C now includes a stacked barplot showing the percentage of -10 and -35 boxes that overlap with mutual information hotspots when the parent sequences are randomly scrambled computationally.

• Figure 3

- Panels A-C were added to explain how we define a new -10/-35 box, how many such new boxes each parent has. These panels also illustrate how we associate the presence or absence of a motif with significant changes in fluorescence scores of the daughter sequences.

- We moved the example of P1-GFP to Figure S8 because when we tested the specific mutation which leads to gaining the -10 box, fluorescence did not change.

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from reporter constructs harboring the point mutations predicted by our computational analyses.

- Cartoons of RNA polymerase have been removed.

• Figure 4

- The tandem-motif has been removed from the figure.

- Cartoons of RNA polymerase have been removed.

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from constructs harboring the point mutations predicted by our computational analyses.

• Figure 5

- This is a new figure analyzing the role of H-NS in promoter evolution and emergence.

• Figure S4

- Panel B now shows the wild-type parent scores and their standard deviations from the sort-seq experiment.

• Figure S5

- Panels with -10 and -35 box counts moved to Figure 1.

- The panel comparing Pnew to hotspot counts was removed.

- Correlations between different k-mers and Pnew are added to panels C-H.

• Figure S8

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from constructs harboring the point mutations predicted by our computational analyses.

• Figure S9

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from constructs harboring the point mutations predicted by our computational analyses.

• Figure S10

- We now include the DNA sequences, the PWM scores, the spacer lengths, and the fluorescence values from constructs harboring the point mutations predicted by our computational analyses.

• Figure S11

- Added DNA sequences and PWM scores.

• Figure S12

- A new figure with further insights about H-NS.

• Figure S13

- A new figure regarding the UP-element analysis.

• Figure S14

- Added Panel D to show how we created mutant reporter constructs for validation.

https://doi.org/10.7554/eLife.98654.2.sa0