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

Timothy Fuqua[1,2], Yiqiao Sun[1,2], Andreas Wagner[1,2,3*]

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.

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.

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)***.

[1] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland.
[2] Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland.
[3] The Sante Fe Institute, Sante Fe, NM, USA.
* Corresponding email: andreas.wagner "at" ieu.uzh.ch

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 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, <u>P6-RFP</u>, P8-RFP, P9-RFP, P10-RFP, P11-GFP, P12-GFP, P17-GFP, P18-GFP, <u>P18-RFP</u>, <u>P19-RFP</u>, 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.

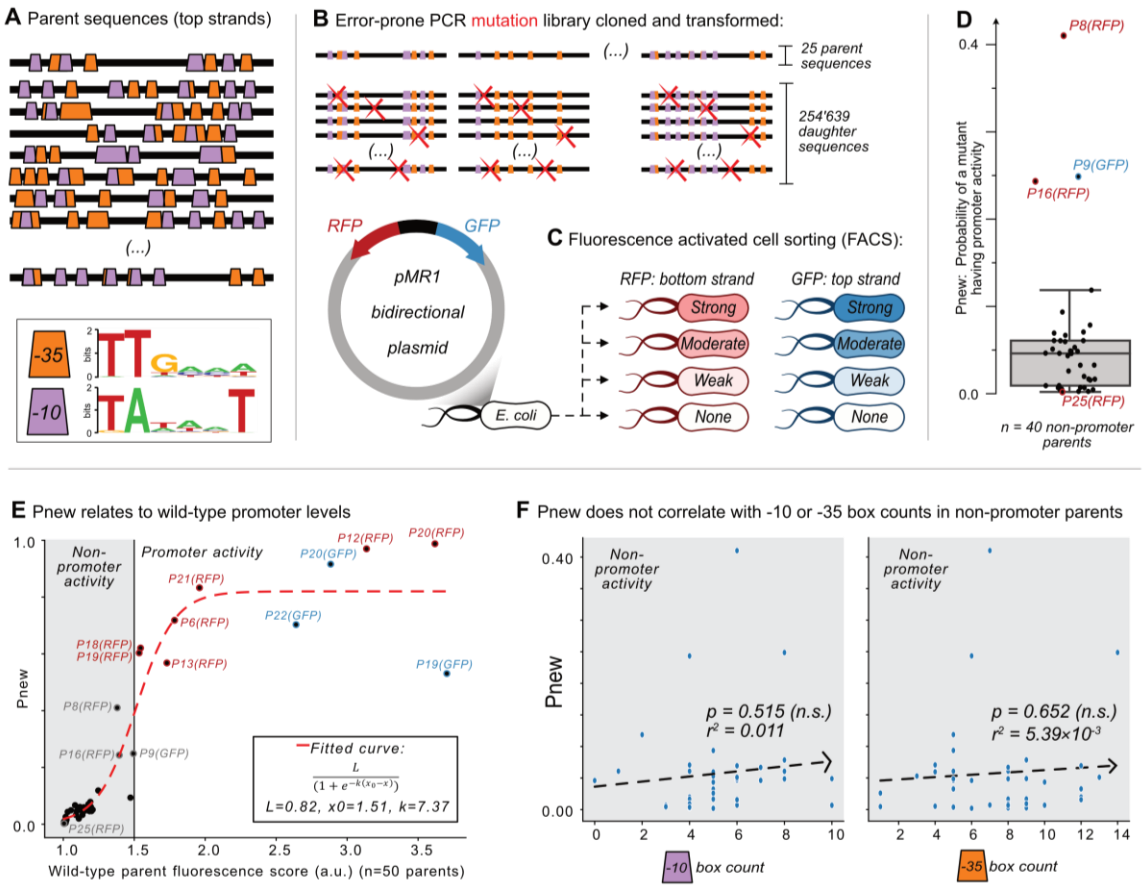

Figure 1. Mutagenesis reveals vastly different probabilities of promoter emergence. (A) The location of -10 and -35 boxes in a subset of the parent sequences. See **Fig S1** for the complete set (n=50). Orange trapezoids correspond to -35 boxes, and magenta trapezoids to -10 boxes, each identified using 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}).

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**).

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.

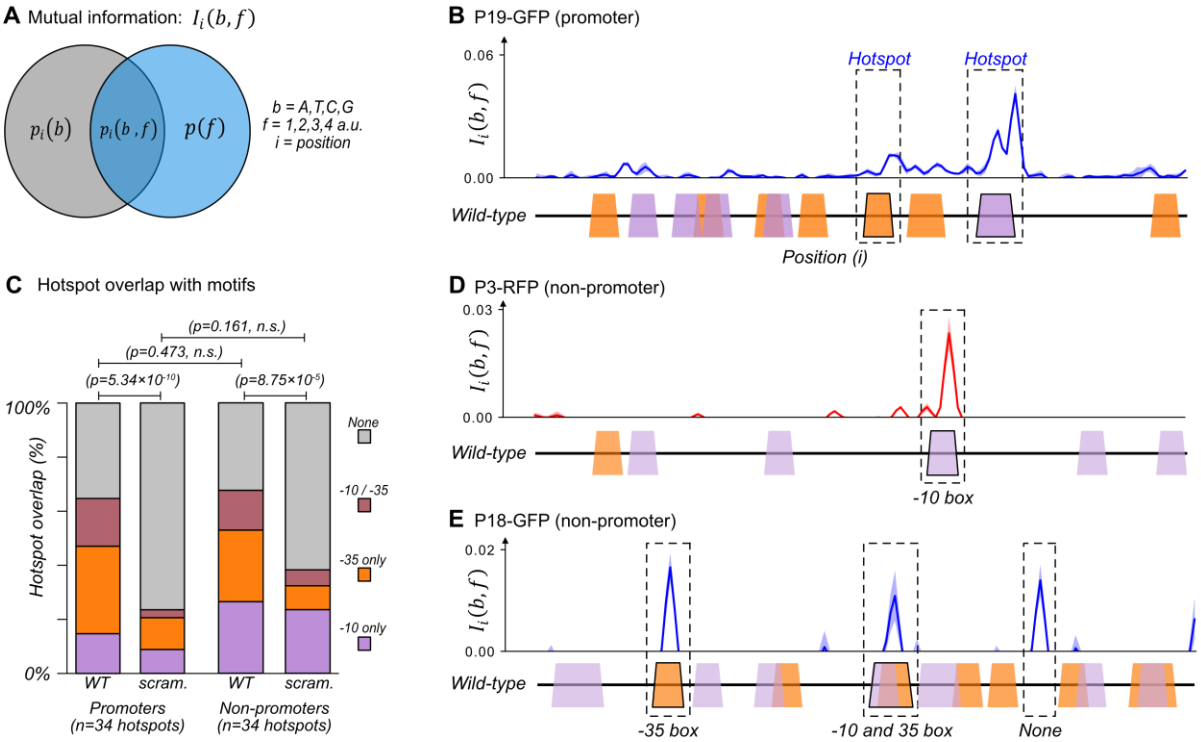

Figure 2. The majority of promoters emerge and evolve within a subset of preexisting promoter motifs. (A) We calculated the mutual information $I_i(b,f)$ between nucleotide identity (b = A,T,C,G) and fluorescence scores rounded to the nearest whole number (f= *1,2,3,4 a.u.*) for each position i in a 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.

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, $q=4.25\times10^{-42}$). 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.

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).

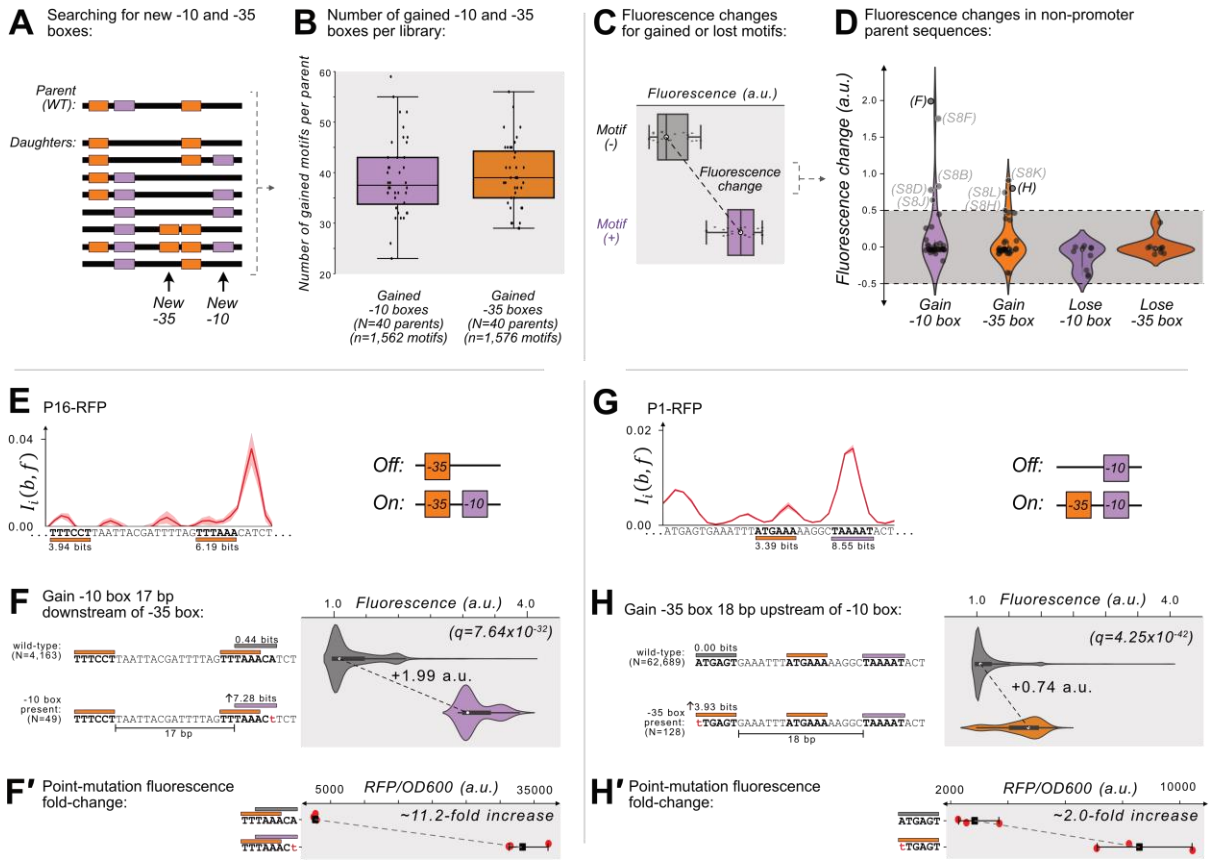

Figure 3. Gaining -10 and -35 boxes rarely creates de-novo promoters. (A) A cartoon for how we identify new -35 and -10 boxes (orange and magenta boxes, respectively) within each DNA sequence (thick, black, horizontal bars). Top: we first identify the -10 and -35 boxes in the wild-type parent sequence. 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.

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.

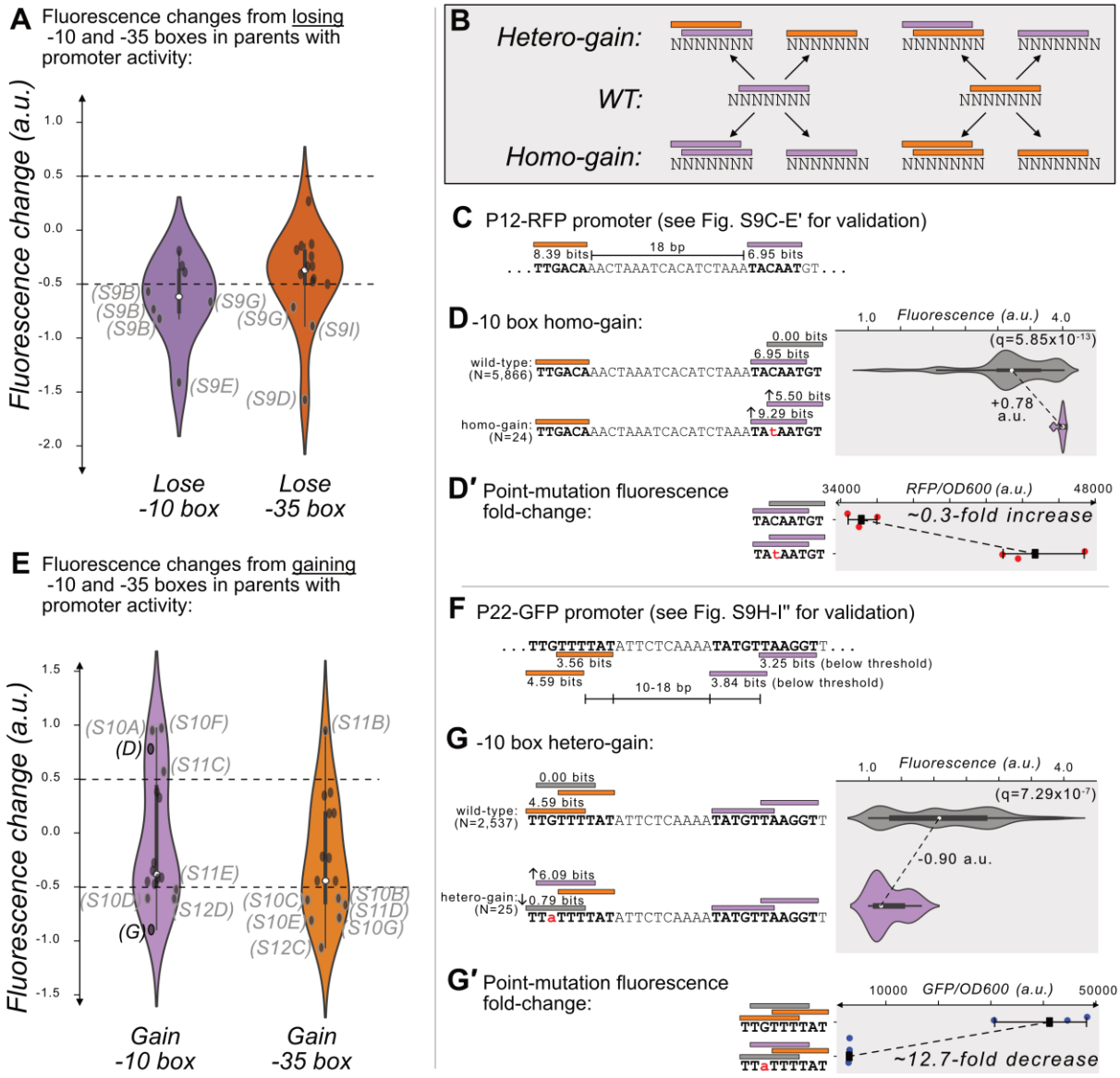

Figure 4. Gaining -10 and -35 boxes modulates promoter activity.
(**A**) The change in fluorescence (arbitrary units, a.u.) when losing -10 and -35 boxes in hotspots in the active parent sequences. Dashed lines indicate an effect size threshold of ± 0.5 a.u. Each black point corresponds to a gain or 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).

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.

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 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

553 background (two-tailed t-test, p=0.310). These data show that the *Δhns* background also
554 rescues the fluorescence loss associated with gaining a new -10 box. Thus, expression lowers
555 because mutations modulate H-NS binding, not by gaining the -10 box.

556 To summarize, we present evidence that H-NS represses both P22-GFP and P12-RFP in *cis.*
557 H-NS also modulates the bidirectionality of P20-GFP/RFP in cis or trans. In P22-GFP, the
558 strongest H-NS motif lies upstream of the promoter. In P12-RFP, the strongest H-NS motifs lie
559 upstream of the -10 and -35 boxes of the promoter. We note that there are 16 additional H-NS
560 motifs surrounding the promoter in P12-RFP that may also regulate P12-RFP (**Fig S12G**).
561 Mutations in two of these two H-NS motifs can create additional -10 and -35 boxes that appear
562 to lower expression. However, the effects of these mutations are insignificant in the absence
563 of H-NS, suggesting that these mutations actually modulate H-NS binding.
564
565 **The UP-element does not strongly influence promoter activity in our dataset.**
566 The UP element is an additional AT-rich promoter motif that can lie stream of a -35 box in a
567 promoter sequence ***(Estrem et al., 1998; Ross et al., 1993)***. We asked whether the creation
568 of UP-elements also creates or modulates promoter activity in our dataset. To this end, we first
569 identified a previously characterized position-weight matrix for the UP element
570 (`NNAAAWWTWTTTTNNWAAASYM`, PWM threshold score = 19.2 bits) ***(Estrem et al., 1998)*** (**Fig**
571 **S13A**)**.** We then computationally searched for UP-element-specific hotspots within the parent
572 sequences, i.e., locations in which mutations that gain or lose UP-elements lead to significant
573 fluorescence increases (Mann-Whitney U-test, **Fig S7** and methods. See **Data S8** for the
574 coordinates, fluorescence changes, and significance). The analysis did not identify any UP
575 elements whose mutation significantly changes fluorescence.

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

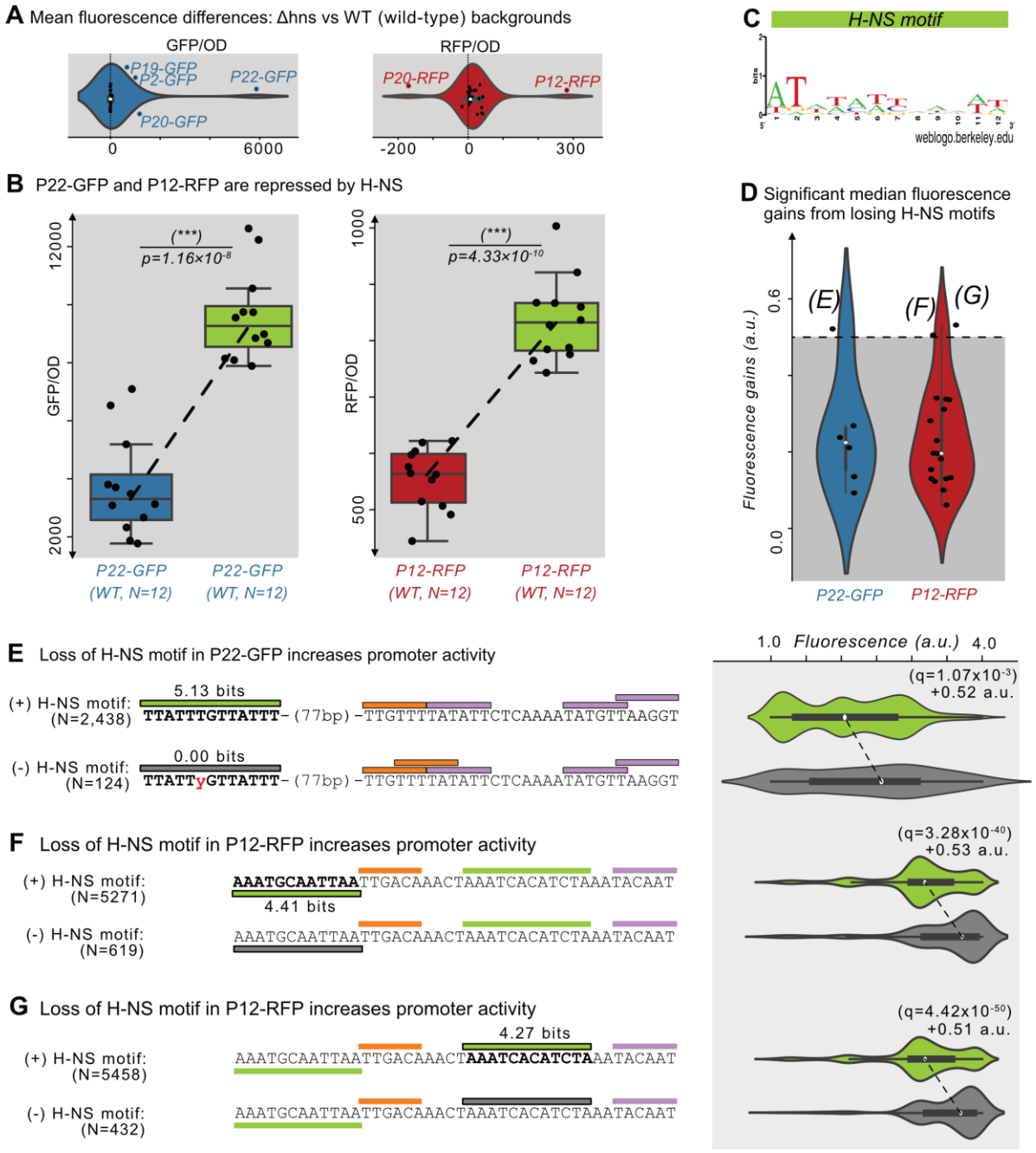

Figure 5. Histone-like nucleoid-structuring protein (H-NS) represses P12-RFP and P22-GFP. **(A)** The difference in average fluorescence levels for each parent sequence in a mutant background for the Histone-like nucleoid-structuring protein (H-NS) *Δhns* vs the wild-type background. Positive and negative 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.

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

Data S1 is a table of the DNA sequences for the primers and template sequences in this study. Data is stored as an Excel spreadsheet.

Data S2 is a dataframe with all daughter sequences and their respective fluorescence scores after quality filtering the data as described in the Methods subsection: Processing sequencing reads. Data is stored as a csv file.

Data S3 is a dataframe containing the location of each mutational information hotspot, and information on whether the hotspot overlaps with a -10 or -35 box. Data is stored as a csv file.

Data S4 is a dataframe with the results from our computational search for hotspots where -10 and -35 boxes are gained or lost, and the extent to which they associate with significant fluorescence changes. Data stored as a csv file.

Data S5 is an Excel spreadsheet that can help to rapidly reproduce the main figures (to the best of Excel's capabilities).

Data S6 is a dataframe with fluorescence readouts from our plate-reader experiments. Data is stored as a csv file.

Data S7 is a dataframe with the results from our computational search for hotspots where H-NS motifs are gained or lost, and the extent to which they associate with significant fluorescence changes. Data stored as a csv file.

Data S8 is a dataframe with the results from our computational search for hotspots where UP-element motifs are gained or lost, and the extent to which they associate with significant fluorescence changes. Data stored as a csv file.

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 55ºC for annealing and 72ºC 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.

785 **Molecular cloning**
786 We cloned the error-prone PCR library into pMR1 at the EcoRI and BamHI sites using
787 NEBuilder (New England Biolabs, USA, product #E2621) as follows: 100 ng of linear pMR1,
788 25 ng insert, 5 uL NEBuilder mastermix, and water to 10 uL total volume. To increase the
789 reaction efficiency, we added 1 uL of T4 DNA ligase buffer (New England Biolabs, USA,
790 product #B0202S). Using a thermal cycler (C1000 Touch Thermal Cycler, Bio-Rad, USA) we
791 incubated each assembly reaction for 1 hour at 50ºC.

792 Subsequently, we electroporated our product directly into *E. coli* DH5α electrocompetent cells
793 (Takara, Japan, product #9027) by adding 2 uL of (unpurified) product to 2× 50 uL aliquots of
794 *E. coli* (100 uL total) in a 2 mm electroporation cuvette (Cell Projects, England, product #EP-
795 202), and electroporated the cells using a Bio-Rad MicroPulser (Bio-Rad, USA). For recovery,
796 we then added 1 mL of Super Optimal Broth with Catabolite Repression Medium (SOC,
797 provided with the electrocompetent cells (Takara, Japan, product #9027), shaking the cells at
798 230 RPM at 37ºC for 1.5 hours (Infors HT, Switzerland, Multitron).

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

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

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

823 Because we used plasmids harboring the template sequence for PCR amplification, the
824 product of each PCR includes inserts with the mutations, and plasmids with the wild-type
825 inserts. To remove the template from the reaction mixture, we directly added 5 uL of rCutsmart
826 buffer (NEB, USA, product #B6004S) and 1 uL of DpnI (NEB, USA, product #R0176S) to each
827 reaction mix, and incubated the reactions at 36ºC for 1 hour followed by a 70ºC incubation for
828 15 minutes using a thermocycler (C1000 Touch Thermal Cycler, Bio-Rad, USA). We then
829 purified the products using a Qiagen QIAquick Gel Purification Kit (Qiagen, Germany, product
830 #28706), as described by the manufacturer.

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

837 **Measuring fluorescence with a plate reader**
838 We streaked bacteria from glycerol stocks onto LB-agar plates with the appropriate selection
839 antibiotics. For the plasmids that we transformed into the DH5α strain (wild-type), this included
840 100 ug/ml of chloramphenicol. For the plasmids that we transformed into the Δhns strain, this
841 included the 100 ug/ml of chloramphenicol with an additional 50ug/ml of kanamycin. The
842 plates were left to incubate overnight (~16 hours) at 37ºC. We then selected individual colonies
843 and inoculated them into 200 uL cultures of LB with 100 ug/ml of chloramphenicol at the
844 aforementioned concentrations. We grew the cultures overnight (~16 hours) at 37ºC, shaking
845 at 230 RPM with an Infors HT incubator (Multitron, Switzerland).

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

850 We pipetted 200 uL of the cultures into a 96-well microplate (Sarstedt AG, Germany, #3924)
851 and measured the fluorescence using a Tecan Spark plate reader (Tecan, Switzerland). We
852 measured GFP with an excitation of 485 nm (bandwidth 20 nm) and emission of 535 nm
853 (bandwidth 20 nm). We measured red fluorescence with an excitation of 560 nm (bandwidth
854 20 nm) and emission of 620 nm (bandwidth 20 nm). We additionally measured the optical
855 density at 600 nm (bandwidth 3.5 nm) (OD_{600}). We report all fluorescence readings from the
856 plate reader using the fluorescence counts normalized to the OD_{600}. See **Data S6** for the
857 readings.

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

868 **Sort-Seq**
869 To perform sort-seq, we first added 100 uL of the glycerol stock of our mutagenesis library to
870 1 mL of LB-chloramphenicol (100 ug/ml). We then allowed the resulting culture to shake at
871 230 RPM and 37ºC overnight (~16 hours) on an Infors HT (Multitron, Switzerland) shaking
872 incubator. Subsequently, we centrifuged the culture and washed the cell pellet twice with
873 Dulbeco's Phosphate Buffered Saline (PBS) (Sigma, USA, D8537). We then resuspended the
874 pellet in PBS.

875 To detect fluorescence caused by GFP and RFP expression, we measured green fluorescence
876 using a 488 nm laser (FITC-H) at 750 volts, and red fluorescence with a 633 nm laser (PE-H)

877 at 510 volts on an Aria III cell sorter (BD Biosciences, USA). We used the following
878 fluorescence gates for sort-seq (**Figure S2**).

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

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

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

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

895 We sorted the library using two consecutive rounds of sorting on consecutive days. On the
896 first day, we sorted 2'160'000, 71'545, 28'399, and 22'720 cells into GFP bins none, low,
897 moderate, and high, respectively; and 2'160'000, 540'000, 101'636, and 76'898 cells into RFP
898 bins none, low, moderate, and high, respectively. After sorting, we added 1 mL of SOC medium
899 (Sigma, USA, product #CMR002K) to the sorted cultures, incubating the cells at 37ºC and
900 shaking for two hours at 230 RPM (Infors HT, Switzerland, Multitron). Following this recovery
901 period, we added 9 mL of LB-Chloramphenicol (100 ug/ml chloramphenicol) to each culture
902 and incubated them overnight (~16 hours) under the same conditions.

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

911 After ~16 hours of overnight culture, we created glycerol stocks for each culture (8 cultures
912 total) as described in "Molecular Cloning." From the remaining culture volume, we isolated the
913 plasmids using a Qiagen QIAprep Spin Miniprep kit (Qiagen, Germany, product #27104) as
914 described by the manufacturer. We amplified the plasmid inserts using PCR as described in
915 "Molecular Cloning", using primers (**Data S1**) that add a unique multiplexing barcode onto the
916 5'-end of each PCR product from each bin. In addition, we isolated the inserts from the
917 unsorted library, and endowed them too with a unique barcode. This procedure yielded 9
918 (=8+1) sequence samples, each with a unique barcode. We then pooled equimolar amounts
919 of these samples, and sequenced them using Illumina paired-end sequencing (Eurofins
920 GmbH, Germany).

921 **Illumina sequencing**
922 The amplicon pool was sequenced by Eurofins Genomics (Eurofins GmbH, Germany) using
923 a NovaSeq 6000 (Illumina, USA) sequencer, with an S4 flow cell, and a PE150 (Paired-end
924 150 bp) run. In total, 282'843'000 reads and 84'852'900'000 bases were sequenced. Raw
925 sequencing reads can be found here: https://www.ncbi.nlm.nih.gov/bioproject/1071572

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

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

942 To map each daughter sequence to its respective template parent sequence, we calculated
943 the Hamming distance (i.e. the number of nucleotide differences between two sequences)
944 between each daughter sequence and the 78 parent sequences, and identified the parent as
945 the sequence with the smallest Hamming distance. To minimize our likelihood of mapping the
946 daughter sequences to the wrong parent sequence, we removed daughter sequences from
947 the dataset for which the closest Hamming distance was greater than 10. Because the error-
948 prone PCR created different numbers of daughter sequences per parent sequence, we also
949 removed all parent and their respective daughter sequences if there were less than 2'000
950 unique daughter sequences. After these calculations and filtration steps, our dataset contained
951 245'639 daughter sequences derived from 25 parent sequences.

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

954 $$F = \frac{\sum_{f=1}^{4}(f \times Reads_f)}{\sum_{f=1}^{4}(Reads_f)} \qquad (1)$$

955
956 Here, f is an integer index corresponding to each of the four fluorescence bins, i.e., f = 1,2,3,4
957 corresponds to no fluorescence, low, moderate, and high fluorescence, respectively. $Reads_f$
958 is the number of times the sequence was identified in each fluorescence bin f. We calculated
959 F for both green and red fluorescence. Each score is in arbitrary units (a.u.) of fluorescence.

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

963 **Mutual information**
964 To calculate mutual information between nucleotides and fluorescence scores, we first
965 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)***:

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

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 AND CONTRIBUTIONS

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.

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. doi:10.1073/pnas.1410569111

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, Zeitlinger J. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**:354–366. doi: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**:2006.0008. doi: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

1061 regulation in bacteria. *Proceedings of the National Academy of Sciences* **115**:E4796–E4805.
1062 doi:10.1073/pnas.1722055115
1063 Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002.
1064 Exploiting transcription factor binding site clustering to identify cis-regulatory modules
1065 involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* **99**:757–
1066 762. doi:10.1073/pnas.231608898
1067 Bracco L, Kotlarz D, Kolb A, Diekmann S, Buc H. 1989. Synthetic curved DNA sequences can act as
1068 transcriptional activators in Escherichia coli. *EMBO J* **8**:4289–4296.
1069 Burgess RR, Travers AA, Dunn JJ, Bautz EK. 1969. Factor stimulating transcription by RNA polymerase.
1070 *Nature* **221**:43–46. doi:10.1038/221043a0
1071 Burr T, Mitchell J, Kolb A, Minchin S, Busby S. 2000. DNA sequence elements located immediately
1072 upstream of the –10 hexamer in Escherichia coli promoters: a systematic study. *Nucleic Acids*
1073 *Res* **28**:1864–1870.
1074 Bykov A, Glazunova O, Alikina O, Sukharicheva N, Masulis I, Shavkunov K, Ozoline O. 2020. Excessive
1075 Promoters as Silencers of Genes Horizontally Acquired by Escherichia coli. *Front Mol Biosci*
1076 **7**:28. doi:10.3389/fmolb.2020.00028
1077 Cazemier AE, Verdoes JC, van Ooyen AJJ, Op den Camp HJM. 1999. Molecular and Biochemical
1078 Characterization of Two Xylanase-Encoding Genes from Cellulomonas pachnodae. *Appl*
1079 *Environ Microbiol* **65**:4099–4107.
1080 Chauhan V, Bahrudeen MNM, Palma CSD, Baptista ISC, Almeida BLB, Dash S, Kandavalli V, Ribeiro AS.
1081 2022. Analytical kinetic model of native tandem promoters in E. coli. *PLOS Computational*
1082 *Biology* **18**:e1009824. doi:10.1371/journal.pcbi.1009824
1083 Chiu T-P, Li J, Jiang Y, Rohs R. 2022. It is in the flanks: Conformational flexibility of transcription factor
1084 binding sites. *Biophysical Journal* **121**:3765–3767. doi:10.1016/j.bpj.2022.09.020
1085 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
1086 Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational
1087 molecular biology and bioinformatics. *Bioinformatics* **25**:1422–1423.
1088 doi:10.1093/bioinformatics/btp163
1089 Cooper C, Legood S, Wheat RL, Forrest D, Sharma P, Haycocks JRJ, Grainger DC. 2024. H-NS is a
1090 bacterial transposon capture protein. *Nat Commun* **15**:7137. doi:10.1038/s41467-024-
1091 51407-5
1092 Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH. 2006. Collision events between RNA
1093 polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids*
1094 *Research* **34**:5416–5425. doi:10.1093/nar/gkl668
1095 Dillon SC, Dorman CJ. 2010. Bacterial nucleoid-associated proteins, nucleoid structure and gene
1096 expression. *Nat Rev Microbiol* **8**:185–195. doi:10.1038/nrmicro2261
1097 Dornenburg JE, DeVita AM, Palumbo MJ, Wade JT. 2010. Widespread Antisense Transcription in
1098 Escherichia coli. *mBio* **1**:10.1128/mbio.00024-10. doi:10.1128/mbio.00024-10
1099 Estrem ST, Gaal T, Ross W, Gourse RL. 1998. Identification of an UP element consensus sequence for
1100 bacterial promoters. *Proc Natl Acad Sci U S A* **95**:9761–9766.
1101 Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of
1102 developmental enhancers. *Science* **350**:325–328. doi:10.1126/science.aac6948
1103 Forrest D, Warman EA, Erkelens AM, Dame RT, Grainger DC. 2022. Xenogeneic silencing strategies in
1104 bacteria are dictated by RNA polymerase promiscuity. *Nat Commun* **13**:1149.
1105 doi:10.1038/s41467-022-28747-1
1106 Fuqua T, Jordan J, van Breugel ME, Halavatyi A, Tischer C, Polidoro P, Abe N, Tsai A, Mann RS, Stern
1107 DL, Crocker J. 2020. Dense and pleiotropic regulatory information in a developmental
1108 enhancer. *Nature* **587**:235–239. doi:10.1038/s41586-020-2816-5
1109 Fuqua, Wagner. 2023. Mobile DNA is replete with hotspots for the de novo emergence of gene
1110 regulation. *bioRxiv* 2023.10.22.563463. doi:10.1101/2023.10.22.563463

1111 Galupa R, Alvarez-Canales G, Borst NO, Fuqua T, Gandara L, Misunou N, Richter K, Alves MRP,
1112 Karumbi E, Perkins ML, Kocijan T, Rushlow CA, Crocker J. 2023. Enhancer architecture and
1113 chromatin accessibility constrain phenotypic space during Drosophila development.
1114 *Developmental Cell* **58**:51-62.e4. doi:10.1016/j.devcel.2022.12.003
1115 Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS,
1116 Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A,
1117 Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L,
1118 López-Fuentes A, Hernández-Koutoucheva A, Moral-Chávez VD, Rinaldi F, Collado-Vides J.
1119 2016. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif
1120 clustering and beyond. *Nucleic Acids Res* **44**:D133–D143. doi:10.1093/nar/gkv1156
1121 Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. 2013. Genomic regions flanking E-box
1122 binding sites influence DNA binding specificity of bHLH transcription factors through DNA
1123 shape. *Cell Rep* **3**:1093–1104. doi:10.1016/j.celrep.2013.03.014
1124 Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters
1125 of transcription factor binding sites are a key component of human promoters and
1126 enhancers. *Genome Res* **20**:565–577. doi:10.1101/gr.104471.109
1127 Guazzaroni M-E, Silva-Rocha R. 2014. Expanding the Logic of Bacterial Promoters Using Engineered
1128 Overlapping Operators for Global Regulators. *ACS Synth Biol* **3**:666–675.
1129 doi:10.1021/sb500084f
1130 Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant
1131 alignments of multiple sequences. *Bioinformatics* **15**:563–577.
1132 doi:10.1093/bioinformatics/15.7.563
1133 Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ. 2012. RNA polymerase II collision interrupts
1134 convergent transcription. *Mol Cell* **48**:365–374. doi:10.1016/j.molcel.2012.08.027
1135 Hommais F, Krin E, Laurent-Winter C, Soutourina O, Malpertuy A, Le Caer JP, Danchin A, Bertin P.
1136 2001. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic
1137 nucleoid-associated protein, H-NS. *Mol Microbiol* **40**:20–36. doi:10.1046/j.1365-
1138 2958.2001.02358.x
1139 Huerta AM, Collado-Vides J. 2003. Sigma70 Promoters in Escherichia coli: Specific Transcription in
1140 Dense Regions of Overlapping Promoter-like Signals. *Journal of Molecular Biology* **333**:261–
1141 278. doi:10.1016/j.jmb.2003.07.017
1142 Inukai S, Kock KH, Bulyk ML. 2017. Transcription factor–DNA binding: beyond binding site motifs. *Curr*
1143 *Opin Genet Dev* **43**:110–119. doi:10.1016/j.gde.2017.02.007
1144 Ireland WT, Beeler SM, Flores-Bautista E, McCarty NS, Röschinger T, Belliveau NM, Sweredoski MJ,
1145 Moradian A, Kinney JB, Phillips R. 2020. Deciphering the regulatory genome of Escherichia
1146 coli, one hundred promoters at a time. *eLife* **9**:e55308. doi:10.7554/eLife.55308
1147 Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of*
1148 *Molecular Biology* **3**:318–356. doi:10.1016/S0022-2836(61)80072-7
1149 Jiang D, Cope AL, Zhang J, Pennell M. 2023. On the Decoupling of Evolutionary Changes in mRNA and
1150 Protein Levels. *Mol Biol Evol* **40**:msad169. doi:10.1093/molbev/msad169
1151 Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J,
1152 Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S,
1153 Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007. RNA Maps Reveal New RNA
1154 Classes and a Possible Function for Pervasive Transcription. *Science* **316**:1484–1488.
1155 doi:10.1126/science.1138341
1156 Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the
1157 biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National*
1158 *Academy of Sciences* **107**:9158–9163. doi:10.1073/pnas.1004290107
1159 Klein CA, Teufel M, Weile CJ, Sobetzko P. 2021. The bacterial promoter spacer modulates promoter
1160 strength and timing by length, TG-motifs and DNA supercoiling sensitivity. *Sci Rep* **11**:24399.
1161 doi:10.1038/s41598-021-03817-4

1162 Krieger G, Lupo O, Wittkopp P, Barkai N. 2022. Evolution of transcription factor binding through
1163 sequence variations and turnover of binding sites. *Genome Res* **32**:1099–1111.
1164 doi:10.1101/gr.276715.122
1165 LaFleur TL, Hossain A, Salis HM. 2022. Automated model-predictive design of synthetic promoters to
1166 control transcriptional profiles in bacteria. *Nat Commun* **13**:5159. doi:10.1038/s41467-022-
1167 32829-5
1168 Lagator M, Sarikas S, Steinrueck M, Toledo-Aparicio D, Bollback JP, Guet CC, Tkačik G. 2022.
1169 Predicting bacterial promoter function and evolution from random sequences. *eLife*
1170 **11**:e64543. doi:10.7554/eLife.64543
1171 Lässig M, Mustonen V, Walczak AM. 2017. Predicting evolution. *Nat Ecol Evol* **1**:1–9.
1172 doi:10.1038/s41559-017-0077
1173 Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic Regulatory Clusters in
1174 Drosophila. *Genome Res* **13**:579–588. doi:10.1101/gr.668403
1175 Loell KJ, Friedman RZ, Myers CA, Corbo JC, Cohen BA, White MA. 2024. Transcription factor
1176 interactions explain the context-dependent activity of CRX binding sites. *PLoS Comput Biol*
1177 **20**:e1011802. doi:10.1371/journal.pcbi.1011802
1178 Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F,
1179 Bauersachs S, Graf A, Young SL, Lieb JD, DeMayo FJ, Feschotte C, Wagner GP. 2015. Ancient
1180 Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome
1181 during the Evolution of Mammalian Pregnancy. *Cell Reports* **10**:551–561.
1182 doi:10.1016/j.celrep.2014.12.052
1183 Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome
1184 assemblies. *Bioinformatics* **27**:2957–2963. doi:10.1093/bioinformatics/btr507
1185 Mandal M, Breaker RR. 2004. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* **5**:451–463.
1186 doi:10.1038/nrm1403
1187 Martínez-Antonio A, Collado-Vides J. 2003. Identifying global regulators in transcriptional regulatory
1188 networks in bacteria. *Current Opinion in Microbiology* **6**:482–489.
1189 doi:10.1016/j.mib.2003.09.002
1190 Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, Melé M. 2020. Cis and
1191 trans effects differentially contribute to the evolution of promoters and enhancers. *Genome*
1192 *Biology* **21**:210. doi:10.1186/s13059-020-02110-3
1193 Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V,
1194 Salgado H, Juárez K, Contreras-Moreira B, Huerta AM, Collado-Vides J, Morett E. 2009.
1195 Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor
1196 Binding Sites in E. coli. *PLoS One* **4**:e7526. doi:10.1371/journal.pone.0007526
1197 Meng W, Belyaeva T, Savery NJ, Busby SJW, Ross WE, Gaal T, Gourse RL, Thomas MS. 2001. UP
1198 element-dependent transcription at the Escherichia coli rrnB P1 promoter: positional
1199 requirements and role of the RNA polymerase α subunit linker. *Nucleic Acids Res* **29**:4166–
1200 4178.
1201 Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ. 2016. Contrasting Frequencies and
1202 Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. *Molecular Biology*
1203 *and Evolution* **33**:1131–1146. doi:10.1093/molbev/msw011
1204 Mitchell JE, Zheng D, Busby SJW, Minchin SD. 2003. Identification and analysis of 'extended –10'
1205 promoters in Escherichia coli. *Nucleic Acids Res* **31**:4689–4695.
1206 Mohanty BK, Kushner SR. 2006. The majority of Escherichia coli mRNAs undergo post-transcriptional
1207 modification in exponentially growing cells. *Nucleic Acids Res* **34**:5695–5704.
1208 doi:10.1093/nar/gkl684
1209 Navarre WW, McClelland M, Libby SJ, Fang FC. 2007. Silencing of xenogeneic DNA by H-NS-facilitation
1210 of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes*
1211 *Dev* **21**:1456–1471. doi: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. doi:10.1093/dnares/dsl009

Paget MS, Helmann JD. 2003. The σ70family of sigma factors. *Genome Biology* **4**:203. doi: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**:e62601. doi: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. doi: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**:206. doi: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. doi: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**:1441006. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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**:1663. doi:10.1038/s41467-022-29321-5

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

Tierrafría VH, Rioualen C, Salgado H, Lara P, Gama-Castro S, Lally P, Gómez-Romero L, Peña-Loredo P, López-Almazo AG, Alarcón-Carranza G, Betancourt-Figueroa F, Alquicira-Hernández S, Polanco-Morelos JE, García-Sotelo J, Gaytan-Nuñez E, Méndez-Cruz C-F, Muñiz LJ, Bonavides-Martínez C, Moreno-Hagelsieb G, Galagan JE, Wade JT, Collado-Vides J 2022. 2022. RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microbial Genomics* **8**:000833. doi:10.1099/mgen.0.000833

1263 Uptain SM, Chamberlin MJ. 1997. Escherichia coli RNA polymerase terminates transcription
1264 efficiently at rho-independent terminators on single-stranded DNA templates. *Proceedings*
1265 *of the National Academy of Sciences* **94**:13548–13553. doi:10.1073/pnas.94.25.13548
1266 Urtecho G, Insigne KD, Tripp AD, Brinck MS, Lubock NB, Acree C, Kim H, Chan T, Kosuri S. 2023.
1267 Genome-wide Functional Characterization of Escherichia coli Promoters and Sequence
1268 Elements Encoding Their Regulation. *eLife* **12**. doi:10.7554/eLife.92558
1269 Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S. 2019. Systematic Dissection of Sequence Elements
1270 Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in
1271 Escherichia coli. *Biochemistry* **58**:1539–1551. doi:10.1021/acs.biochem.7b01069
1272 van Hijum SAFT, Medema MH, Kuipers OP. 2009. Mechanisms and Evolution of Control Logic in
1273 Prokaryotic Transcriptional Regulation. *Microbiology and Molecular Biology Reviews* **73**:481–
1274 509. doi:10.1128/MMBR.00037-08
1275 Villanueva-Cañas JL, Horvath V, Aguilera L, González J. 2019. Diverse families of transposable
1276 elements affect the transcriptional regulation of stress-response genes in Drosophila
1277 melanogaster. *Nucleic Acids Res* **47**:6842–6857. doi:10.1093/nar/gkz490
1278 Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT,
1279 Jasinska AJ, Turner JMA, Bertelsen MF, Murchison EP, Flicek P, Odom DT. 2015. Enhancer
1280 evolution across 20 mammalian species. *Cell* **160**:554–566. doi:10.1016/j.cell.2015.01.006
1281 Wade JT, Grainger DC. 2014. Pervasive transcription: illuminating the dark matter of bacterial
1282 transcriptomes. *Nat Rev Microbiol* **12**:647–653. doi:10.1038/nrmicro3316
1283 Wang L, Watters JW, Ju X, Lu G, Liu S. 2023. Head-on and co-directional RNA polymerase collisions
1284 orchestrate bidirectional transcription termination. *Molecular Cell* **83**:1153-1164.e4.
1285 doi:10.1016/j.molcel.2023.02.017
1286 Wang Y, Wang H, Wei L, Li S, Liu L, Wang X. 2020. Synthetic promoter design in Escherichia coli based
1287 on a deep generative network. *Nucleic Acids Research* **48**:6403–6412.
1288 doi:10.1093/nar/gkaa325
1289 Warman EA, Forrest D, Guest T, Haycocks JJRJ, Wade JT, Grainger DC. 2021. Widespread divergent
1290 transcription from bacterial and archaeal promoters is a consequence of DNA-sequence
1291 symmetry. *Nat Microbiol* **6**:746–756. doi:10.1038/s41564-021-00898-9
1292 Warman EA, Singh SS, Gubieda AG, Grainger DC. 2020. A non-canonical promoter element drives
1293 spurious transcription of horizontally acquired bacterial genes. *Nucleic Acids Res* **48**:4891–
1294 4901. doi:10.1093/nar/gkaa244
1295 Westmann CA, Alves L de F, Silva-Rocha R, Guazzaroni M-E. 2018. Mining Novel Constitutive
1296 Promoter Elements in Soil Metagenomic Libraries in Escherichia coli. *Frontiers in*
1297 *Microbiology* **9**.
1298 Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo promoters. *Nat*
1299 *Commun* **9**:1530. doi:10.1038/s41467-018-04026-w
1300 Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y,
1301 Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams
1302 BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See L-H, Fastuca M,
1303 Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender
1304 MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu Y-C,
1305 Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N,
1306 Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi
1307 VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa
1308 B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero
1309 J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ,
1310 Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S,
1311 Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler
1312 EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang K-H, Skoultchi A, Gosh S, Disteche C,
1313 Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams

LB, Zhou X-Q, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B, Mouse ENCODE Consortium. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:355–364. doi:10.1038/nature13992

SUPPLEMENTARY FIGURES

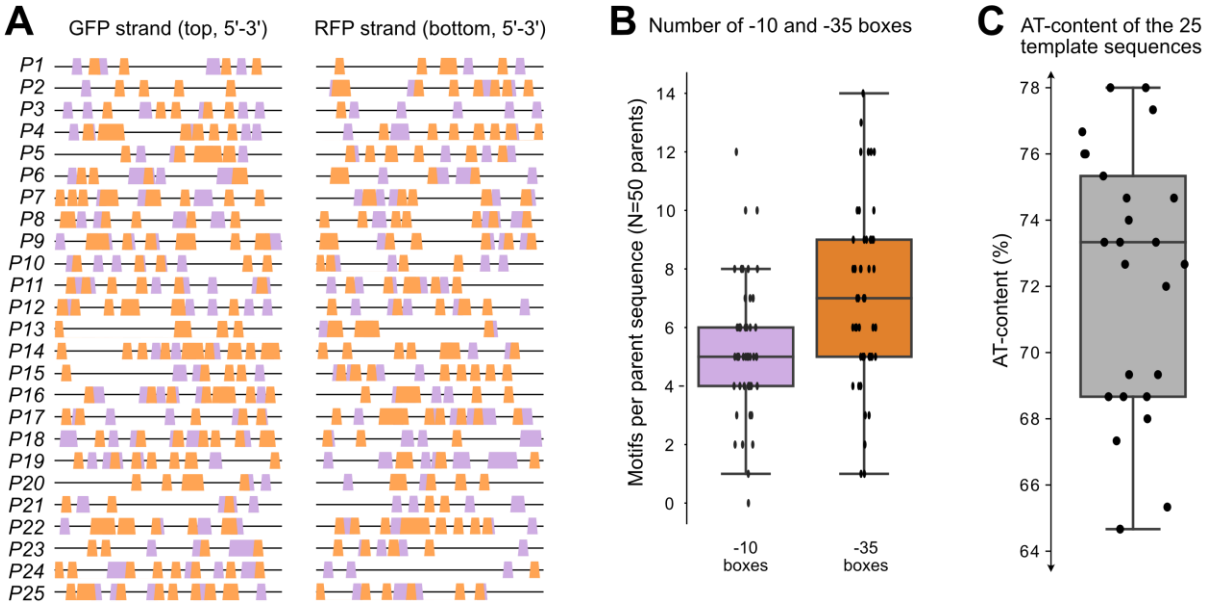

Figure S1. Promoter island sequences. (A) The locations and arrangements of -10 and -35 boxes located across both the top and bottom strands of the 150 base pair DNA template sequences (P1-P25). Orange trapezoids: -35 boxes. Magenta trapezoids: -10 boxes. Left: top (GFP) strand. Right: bottom (RFP) strand. 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

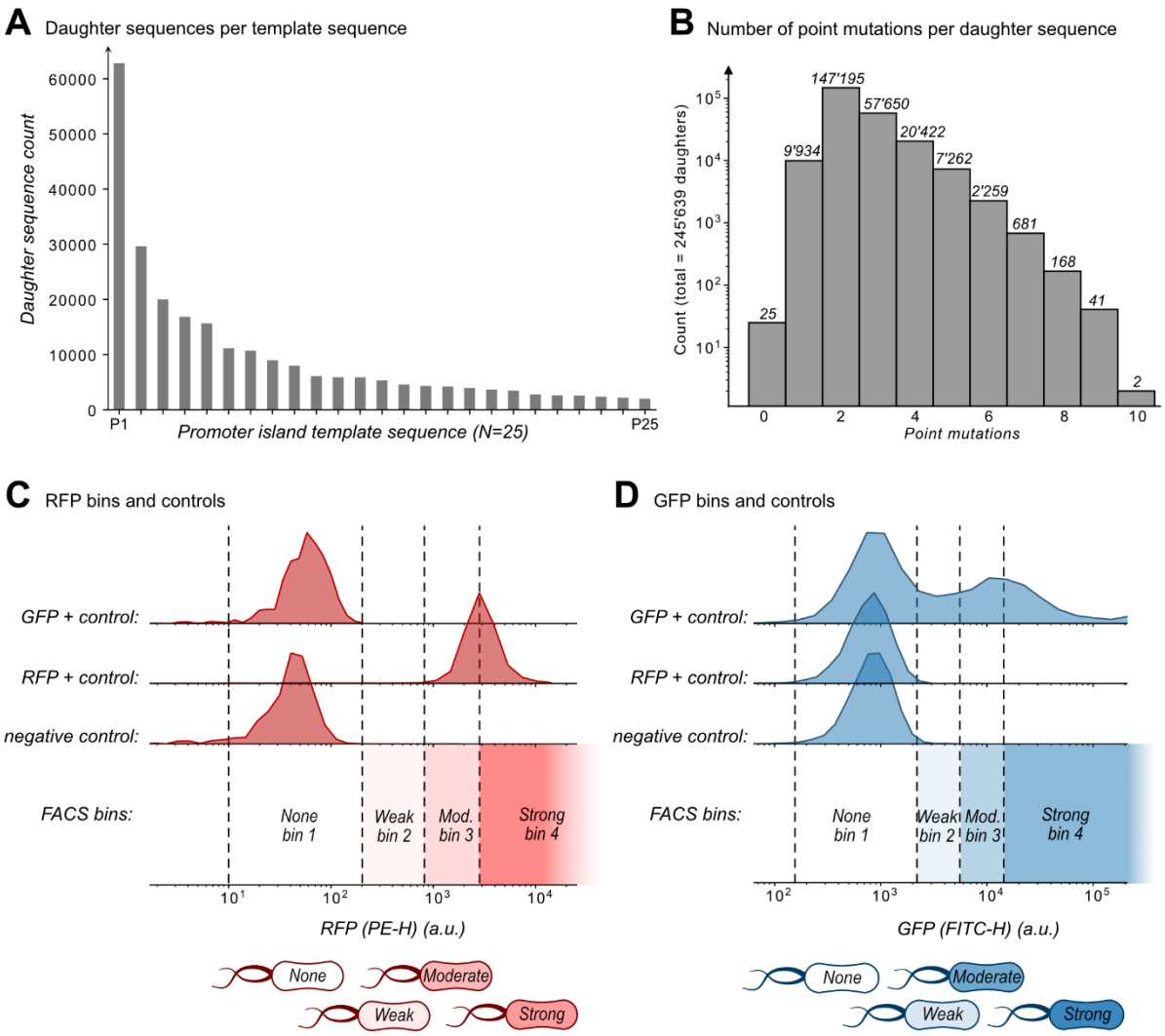

Figure S2. Mutagenesis library and sort-seq bins. (A) A histogram with the number of unique daughter sequences per template sequence (N=245'639 daughters and N=25 template sequences). **(B)** The number of point mutations per daughter sequence (N=245'639 daughter sequences). The number above each bin 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.

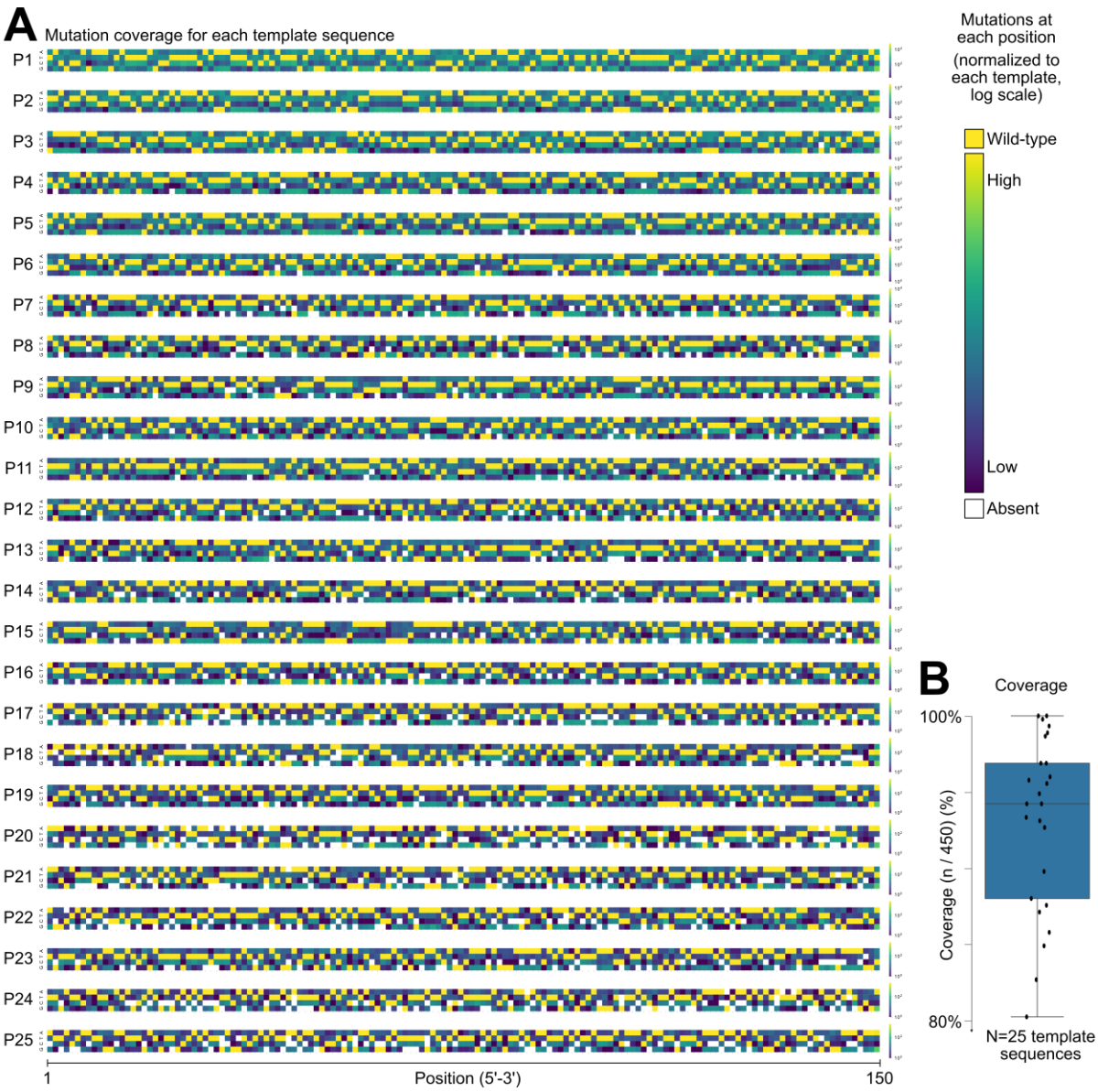

Figure S3. Mutational coverage for the template sequences. (A) Heatmaps depicting the frequency of each nucleotide (A,T,C,G, y-axis) at each position (5'-3', x-axis) for all of the daughter sequences of each template sequence (P1-P25, N=25 templates). Daughter sequences contain 1-10 point mutations (see **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.

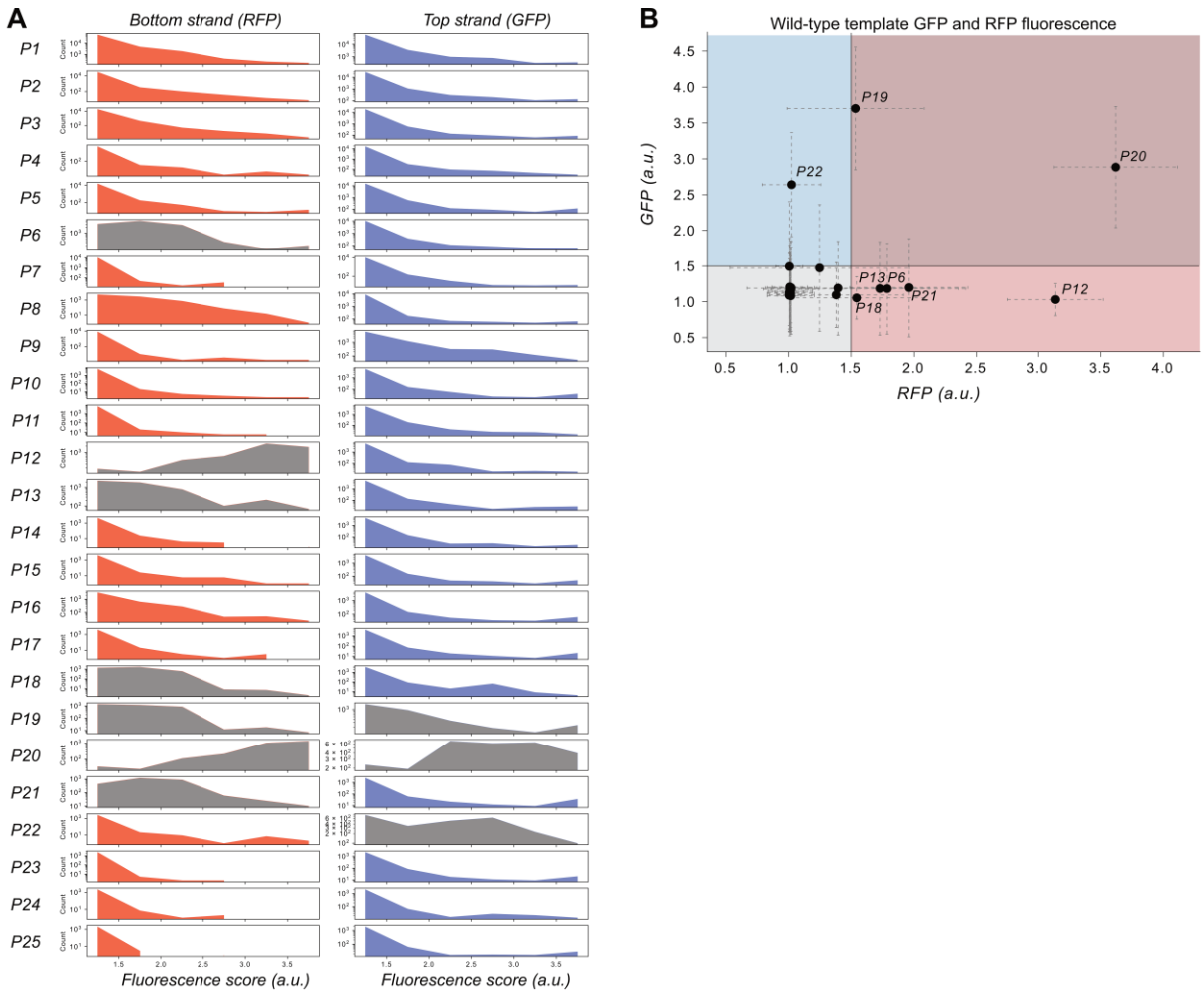

Figure S4. Histogram of the fluorescence score distributions for each parent and its respective daughter sequences. (A) Rows: template sequences (P1-P25). Left column: bottom strand (RFP fluorescence scores); Right column: top strand (GFP fluorescence scores.) Each panel is a histogram of the frequency 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.

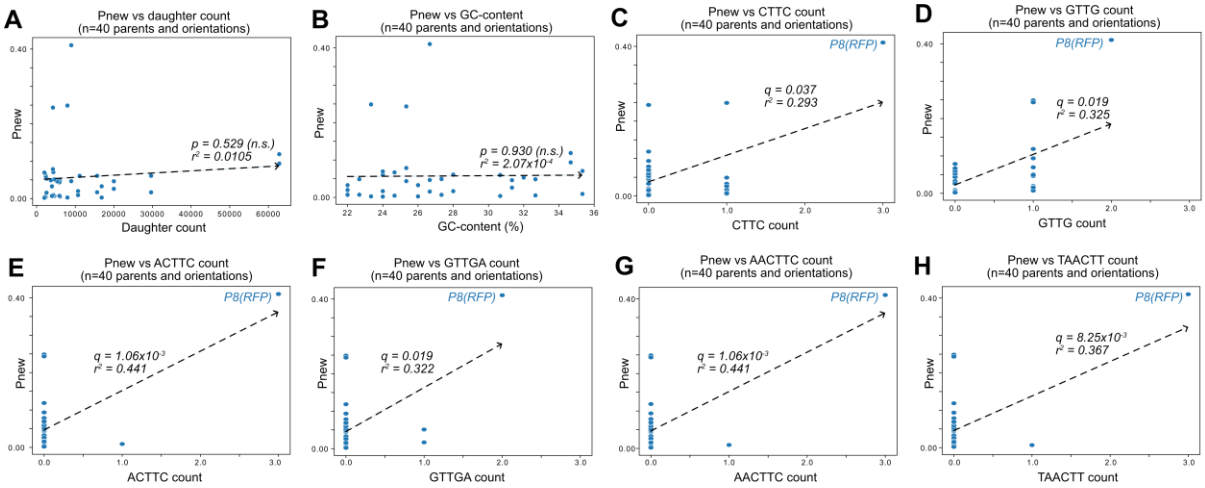

Figure S5. Correlation of P_{new} with sequence composition. (A) Scatterplots of P_{new} (see **Fig 1**) and the number of daughter sequences per parent. P_{new} is the ratio of daughter sequences with a fluorescence score greater than 1.5 a.u. to the total number of daughter sequences for each parent. 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).

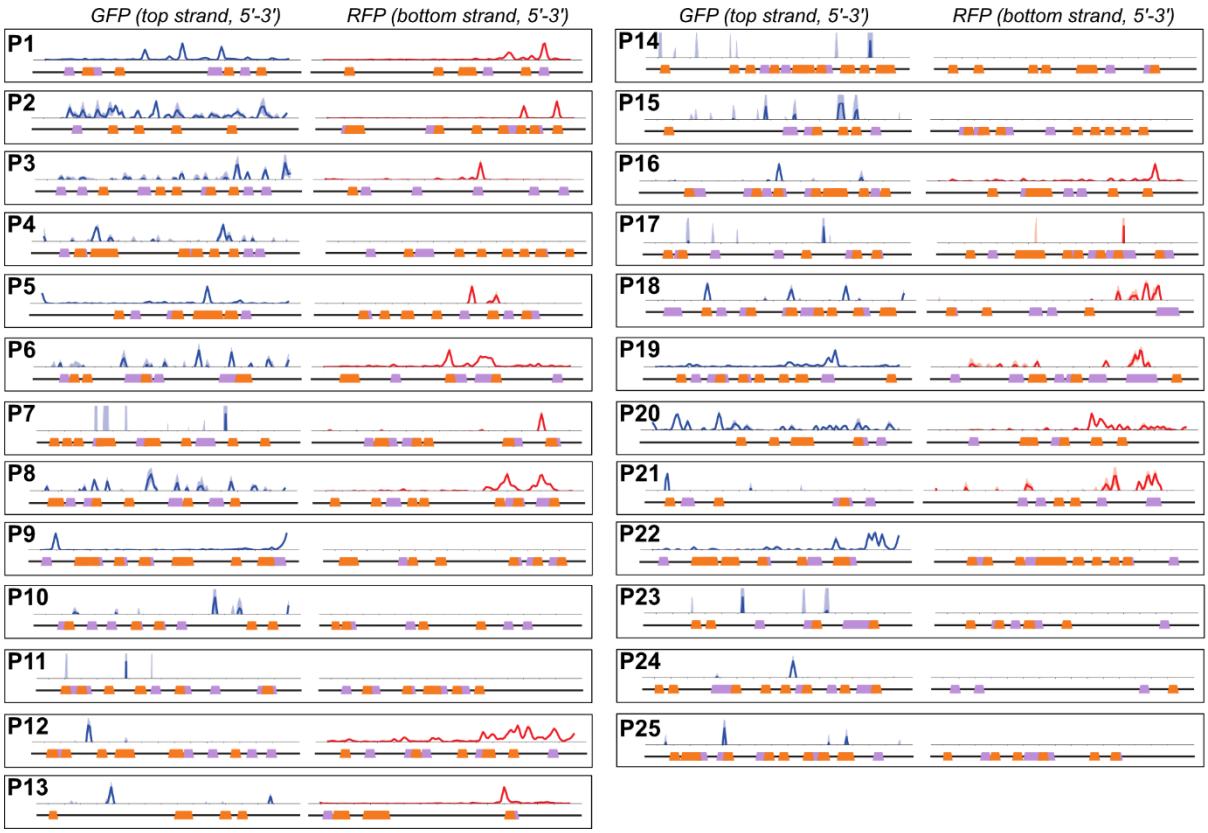

Figure S6. Mutual information and promoter motifs in the parent sequences. Each panel corresponds to a unique parent sequence, with its numerical identifier in bold. Within each panel there are two plots, corresponding to the top strand (left), and the bottom strand (right), both shown from 5' (left) 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.

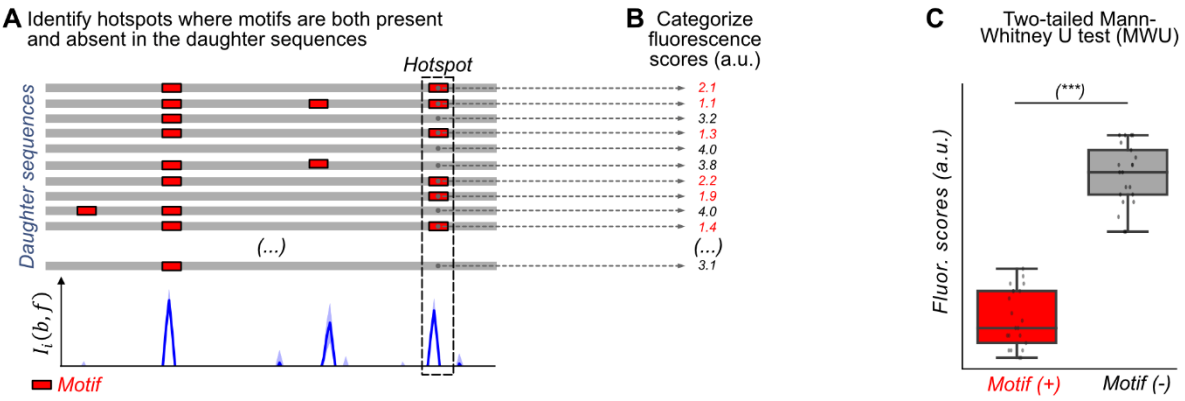

Figure S7. Identifying where motifs are gained and lost in hotspots that are associated with changes in fluorescence. These panels do not show data from any parent sequence or hotspot, but show hypothetical data to illustrate how we identified the associations of base identity and fluorescence in our 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.

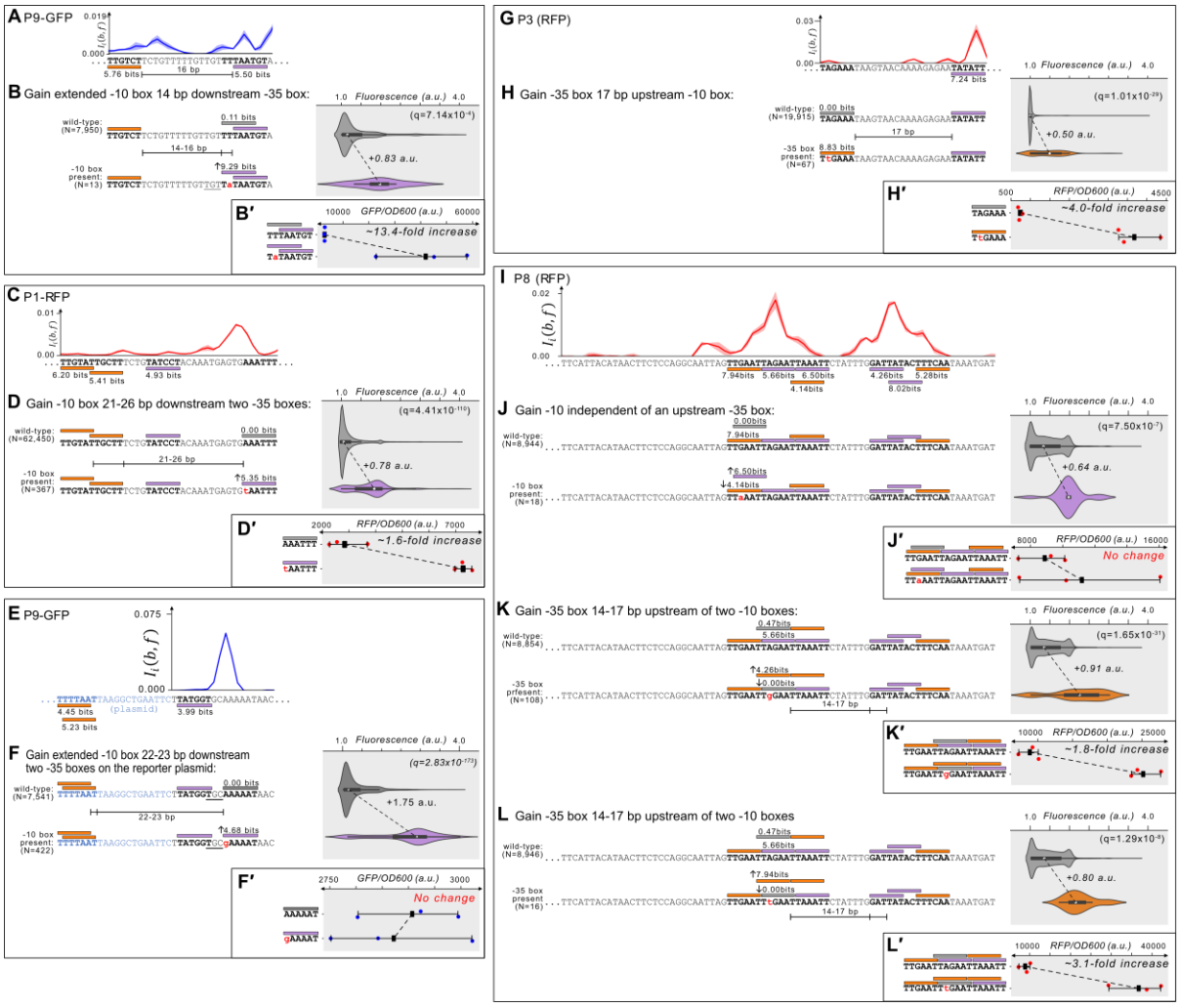

Figure S8. Additional examples of which gaining -10 and -35 boxes creates de-novo promoters. **(A)** Parent P9-GFP. Top: Mutual information $I_i(b,f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset 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).

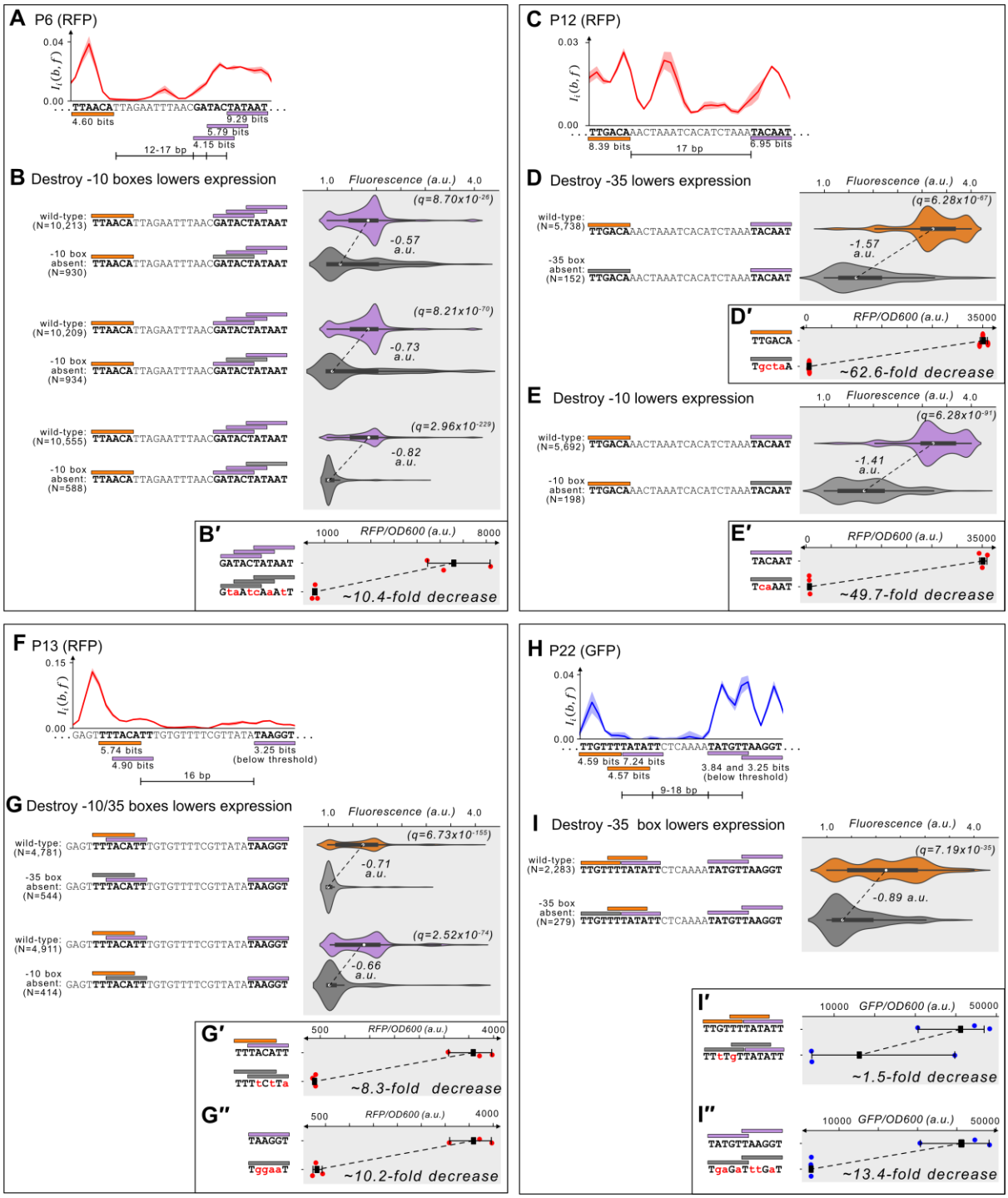

Figure S9. Mapping promoters in active parents.
(A) Parent P6-RFP. Top: Mutual information $I_i(b,f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 standard deviation when the dataset is randomly split into three equally sized subsets (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.

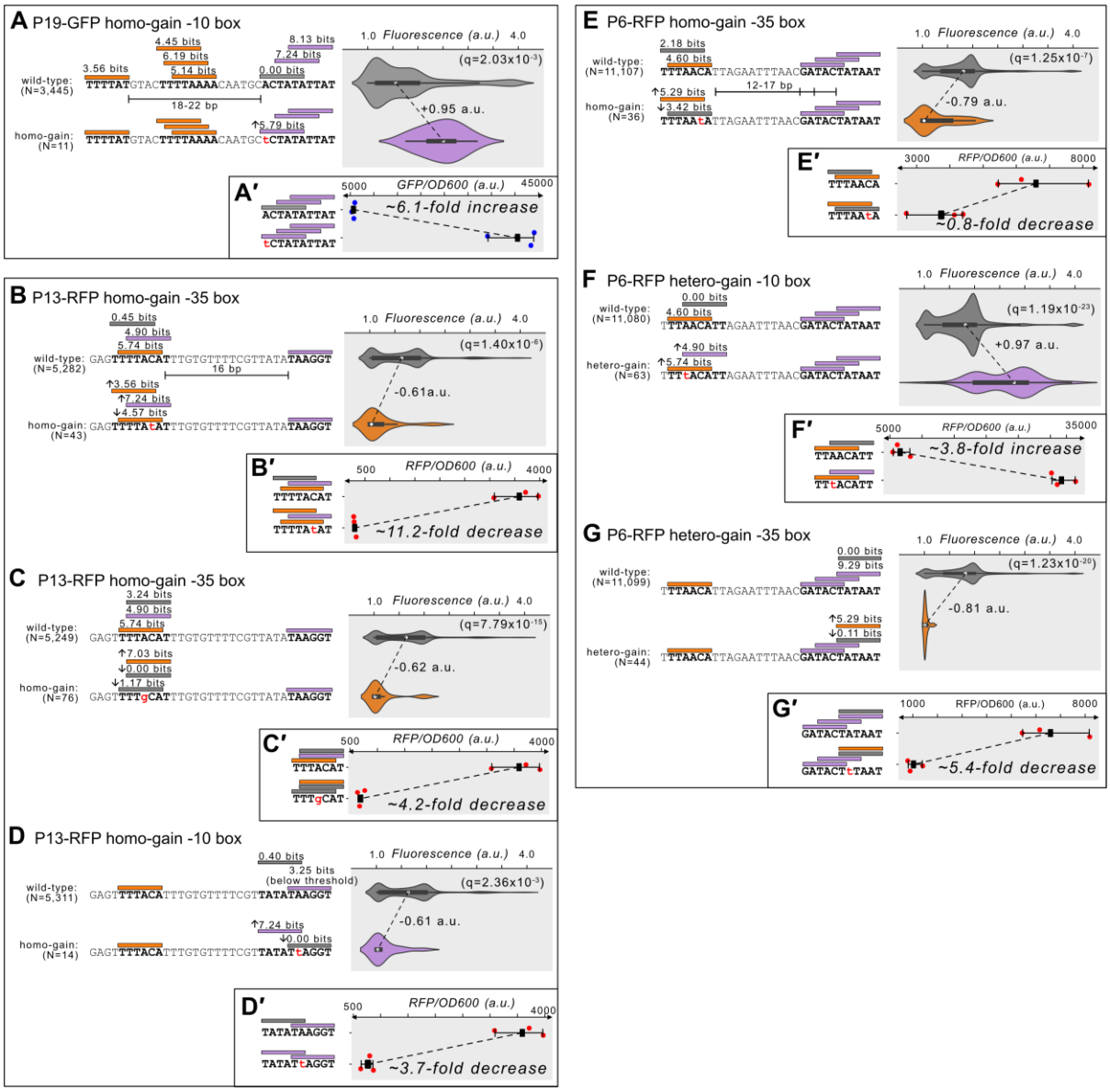

Figure S10. Additional examples of mutations modulating promoter activity.
(A) Top: Parent P19-GFP and its PWM predictions (orange rectangle = -35 box, magenta rectangle = -10 box, gray rectangle = region of interest). We plot the fluorescence scores of all daughters without a -10 box in the region of 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

1502 three colonies with and without the degenerate consensus sequence identified in (D). **(E)** Analogous to (A) but for
1503 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
1504 (A′) but the fluorescence levels from three colonies with and without the degenerate consensus sequence identified
1505 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,
1506 $q=1.19\times10^{-23}$). **(F′)** Analogous to (A′) but the fluorescence levels from three colonies with and without the
1507 degenerate consensus sequence identified in (F). **(G)** Analogous to (A) but for gaining a -35 box in the gray region
1508 of interest on P6-RFP (two-tailed MWU test, $q=1.23\times10^{-20}$). **(G′)** Analogous to (A′) but the fluorescence levels from
1509 three colonies with and without the degenerate consensus sequence identified in (G).

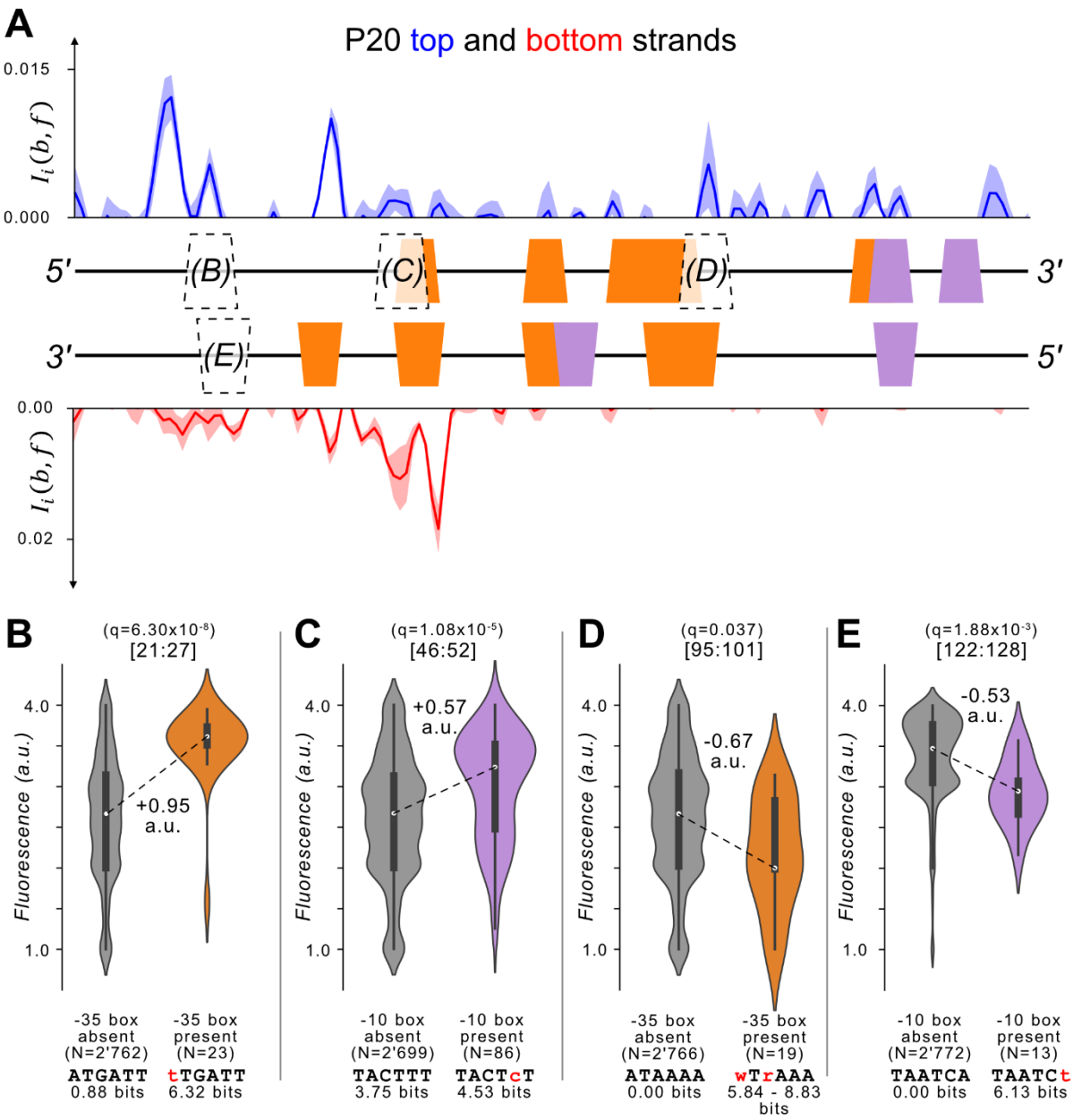

Figure S11. Template P20. **(A)** We plot the mutual information for the top (blue) and bottom (red) strands of template P20. Top: P20-GFP. We plot the mutual information $I_i(b,f)$ between nucleotide identity and fluorescence at each position. Solid line: mean mutual information, shaded region: ± 1 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, $q=1.88\times10^{-3}$).

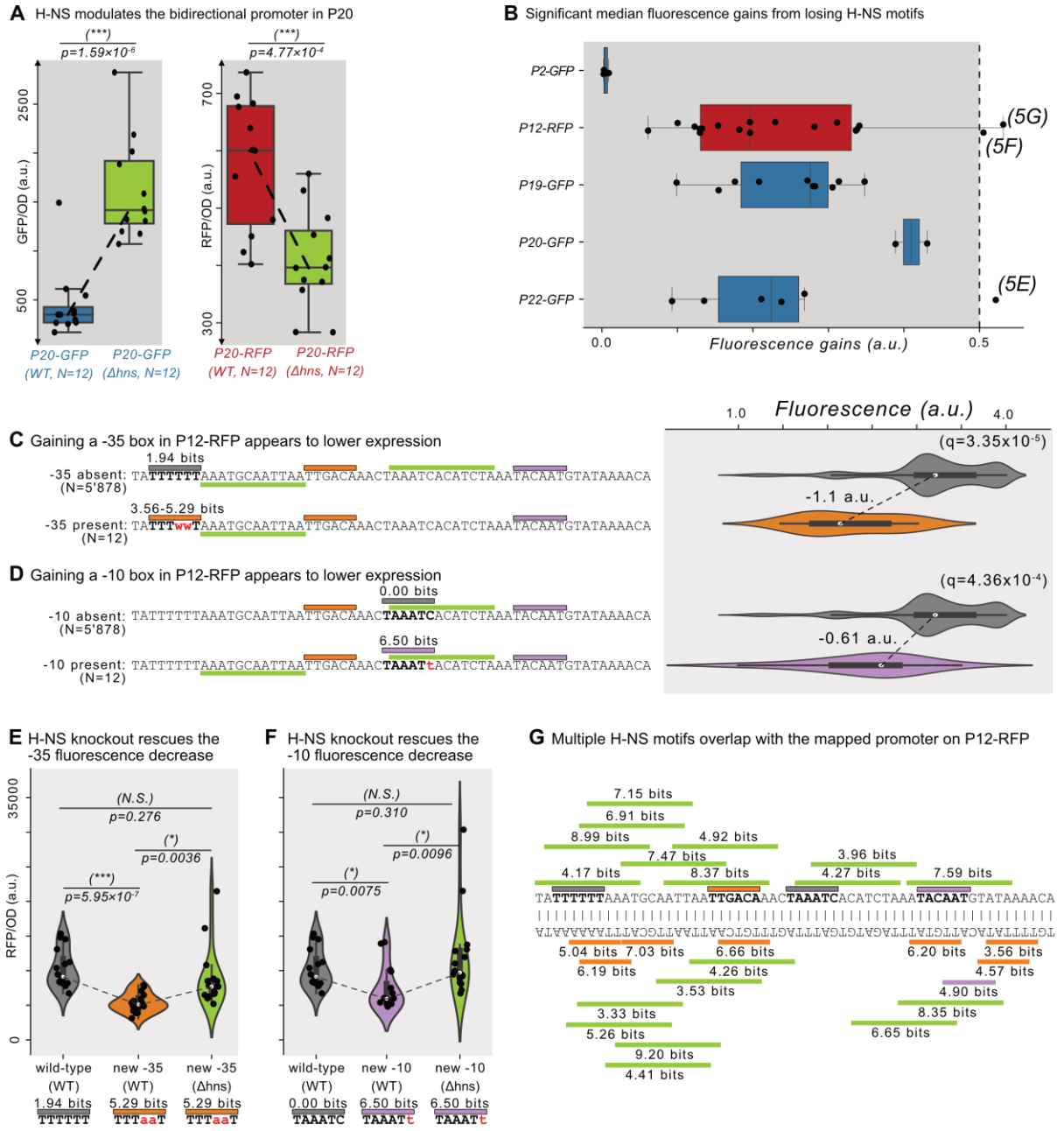

Fig S12. Additional H-NS sites. (A) The fluorescence levels of bacteria colonies (N=12 each) harboring a parent sequence in the wild-type (left) vs the *Δhns* background. Boxes represent the interquartile range (IQR) and the center line the median. Whiskers correspond to 1.5×IQR. We test the null hypothesis 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.

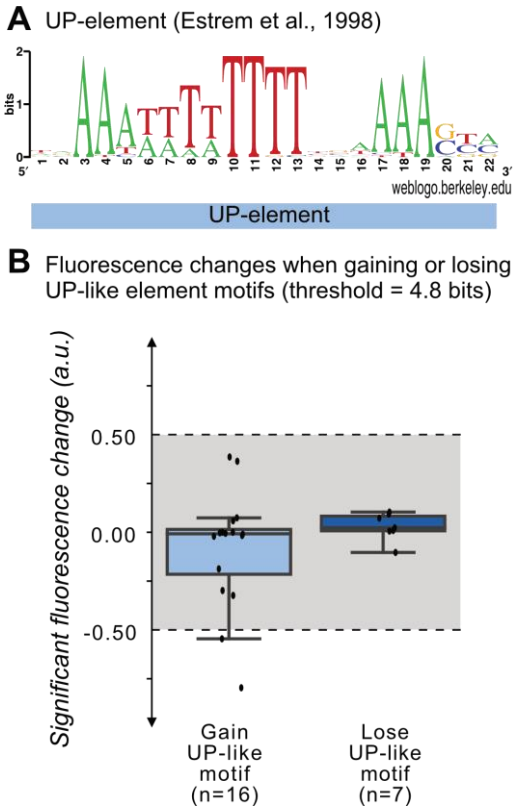

Figure S13. The change in fluorescence when gaining and losing UP-like motifs. (A) A sequence logo for the UP-element derived from a position weight matrix for this element ***(Estrem et al., 1998)***. **(B)** Change in fluorescence (arbitrary units, a.u.) when losing or gaining UP element motifs (threshold =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.

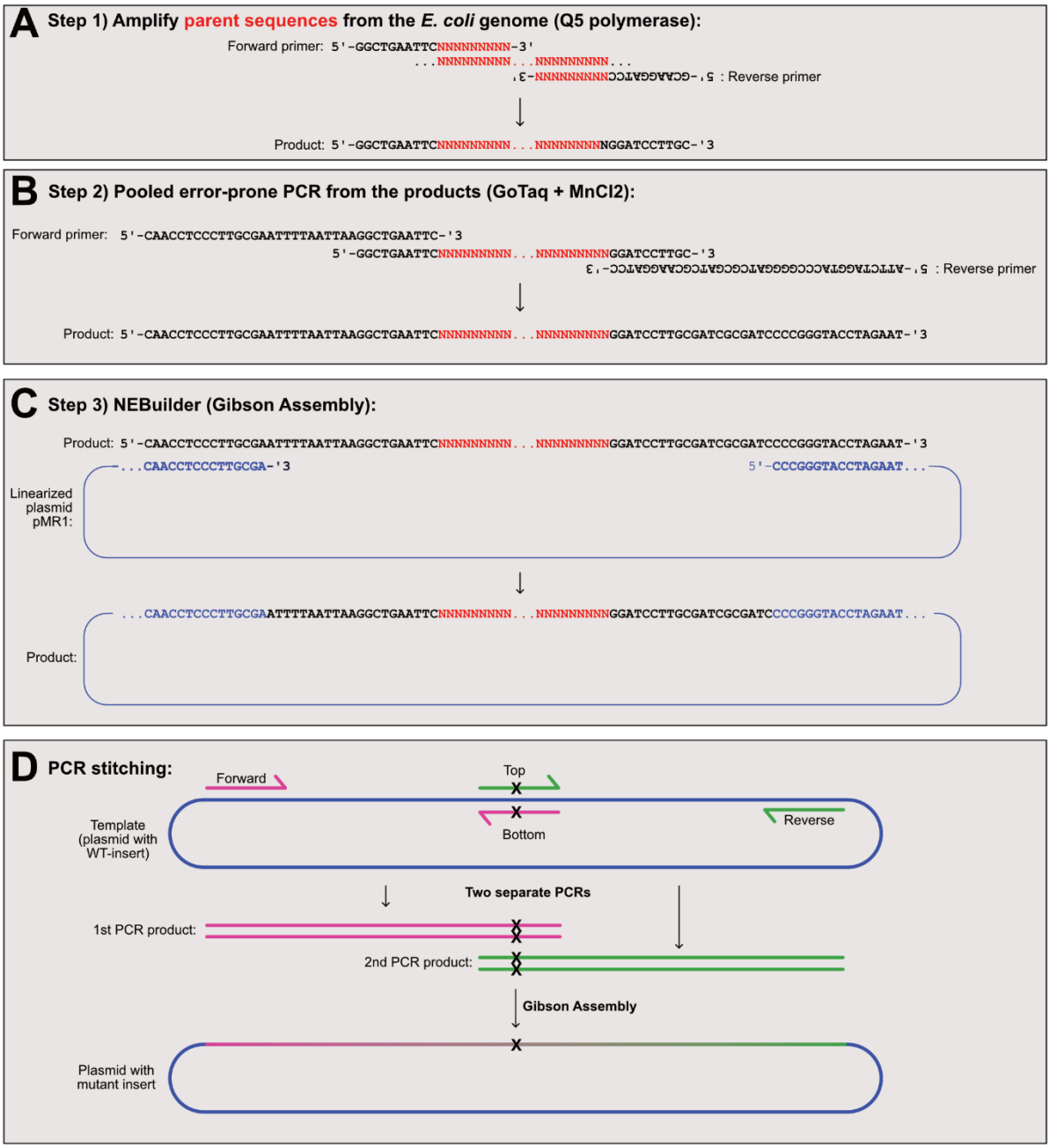

Fig S14. Molecular cloning of parent sequences. (A) We first amplify the parent sequences from the *E. coli* genome using Q5 polymerase. Top: the forward and reverse primers contain constant overhang regions, and unique sequences (red N's) unique to the parent sequence. Bottom: the product is a unique 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.